Escuela Politécnica Superior de Ávila

# Robust Statistical and Artificially Intelligent Approaches for the Analysis of 2D and 3D Morphological Data

**Lloyd Austin David Courtenay**

VNiVERSiDAD
D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

Doctorado Internacional
Geotecnologías Aplicadas a Construcción, Energía e Industria

Supervisors: Diego González-Aguilera & José Yravedra

December 13, 2022

# Robust Statistical and Artificially Intelligent Approaches for the Analysis of 2D and 3D Morphological Data

**Lloyd Austin David Courtenay**

Doctorado Internacional
Geotecnologías Aplicadas a Construcción, Energía e Industria

December 13, 2022

VNiVERSiDAD
D SALAMANCA

**DEPARTAMENTO DE INGENIERÍA
CARTOGRÁFICA Y DEL TERRENO**

Hornos Caleros, nº 50 - Ávila
Tel . (34) 920 35 35 00 Fax . (34) 920 35 35 05 ict@usal.es

# INFORME DE LA TESIS DOCTORAL

"*Robust Statistical and Artificially Intelligent Approaches for the Analysis of 2D and 3D Morphological Data*"

presentada por D. Lloyd Austin David Courtenay

La Tesis Doctoral presentada por el doctorando D. Lloyd Austin David Courtenay reúne todos los requisitos que se le pueden exigir a una Tesis Doctoral con mención internacional, y que a continuación paso a detallar:

- El tema presentado por el doctorando en su Tesis Doctoral se enmarca dentro de una de las líneas de investigación más intensas y novedosas dentro de las disciplinas de la Fotogrametría de Rango Cercano y la Visión Computacional. Más concretamente, **se abordan desde una perspectiva transdisciplinar enfoques estadísticos robustos y acercamientos de inteligencia artificial para el análisis de datos morfológicos en 2D y 3D, aplicados a casos de Arqueología, Paleoantropología, Tafonomía y Dermatología**.

- El estado del arte ha sido rigurosamente documentado y analizado desde un punto de vista científico y tecnológico como así consta a lo largo de la gran cantidad de referencias incorporadas a la Tesis Doctoral.

- La metodología y algoritmos desarrollados en la Tesis Doctoral presentan aspectos novedosos que permiten abordar con éxito los objetivos propuestos. Más concretamente, entre los progresos realizados hay que destacar muy especialmente:

    - El uso de recursos informáticos avanzados y **técnicas estadísticas multivariantes** que permiten mejorar en gran medida la precisión de los estudios que suelen realizarse con criterios visuales y cualitativos.
    - El **empleo de herramientas matemáticas**, que suelen utilizarse en contextos biológicos y antropológicos, para la cuantificación de diferentes morfologías.
    - La **utilización exitosa de algoritmos de inteligencia artificial** como herramienta muy poderosa en la clasificación de elementos en función de su morfología.
    - La creación de **nuevos enfoques metodológicos** aplicados a la tafonomía que nunca habían sido empleados antes.

- Por otro lado, hay que reseñar la **posible transferencia de tecnología** derivada en forma del **desarrollo de software** que será susceptible de ser registrada como propiedad intelectual, y que ha servido como soporte para la experimentación de la metodología y algoritmos desarrollados:

    - *RLibraries*, para el tratamiento estadístico multivariante de los datos.
    - *TPS Measurement* software, para la realización de medidas en análisis morfológicos.
    - *Trampling algorithm*, para aplicar una red neuronal en procesos de análisis de marcas.

- Finalmente, hay que destacar muy especialmente el altísimo nivel de producción científica derivado del propio desarrollo de la Tesis Doctoral por parte del Doctorando, el cual permite avalar la calidad y relevancia de la misma. Hay que reseñar la publicación de **9 artículos indexados JCR**, así como sus aportaciones realizadas en otra elevada cantidad de contribuciones científicas desarrolladas a partir de las innovaciones aplicadas en esta tesis doctoral.

Por todo lo anteriormente reseñado, emito un informe con todos mis pronunciamientos favorables, y autorizo su presentación como Tesis Doctoral en el Departamento de Ingeniería Cartográfica y del Terreno de la Universidad de Salamanca.

Ávila, 30 de noviembre de 2022

LOS DIRECTORES DE LA TESIS DOCTORAL

Fdo. Diego González Aguilera

Fdo. José Yravedra Sainz de los Terreros

Smithsonian
*National Museum of Natural History*
Department of Anthropology
Human Origins Program

December 7, 2022

Briana Pobiner: Reviewer Report for Lloyd Austin David Courtenay PhD dissertation

It was a real pleasure to have the opportunity to serve as an external reviewer for Courtenay's dissertation "*Robust Statistical and Artificially Intelligent Approaches for the Analysis of 2D and 3D Morphological Data*".

I am a taphonomist and zooarchaeologist, and I do not have strong training in the methods used and developed by Courtenay; therefore, my focus in reviewing this dissertation was mainly on the conclusions and implications for my own area of research. But before I get to that, I must note that I have been an external reviewer for several PhD dissertations, and this one is by far the most well organized and well written, with beautiful, clear figures. It is really impressive in its breadth and significance. Courtenay should be commended for the huge amount of high-level work that went into producing this dissertation. Chapters 1, 2, and 4, outside of the nine publications included which are helpfully organized into five themes, show his clear understanding and synthesis of the topics he studied and the innovative approaches he has not only applied, but also developed.

As far as my own areas of expertise, I can comment on the five publications on bone surface modifications: *Applied Sciences* 2019, *PLoS One* 2020, *Applied Sciences* 2020, *Scientific Reports* 2021, *Animals* 2021, *Quaternary Science Reviews* submitted. In taphonomy and zooarchaeology, Courtenay has carved out a strong niche for himself in the field with a focus on new data science-based techniques and approaches to help us measure and classify bone surface modifications more accurately. These methods have implications for our understanding of the abiotic (trampling marks) and biotic (cut marks, carnivore tooth marks) processes that affected modern and fossil bone assemblages.

In sum, this dissertation is an extremely cutting-edge (no pun intended) impressive body of research, especially for a dissertation. I look forward to seeing Courtenay continue to shape the fields of zooarchaeology and taphonomy in the future.

With sincere regards,
Briana Pobiner
Briana Pobiner

School of Life and Health Sciences,
University of Roehampton.
Holybourne Ave SW15 4JD
London (United Kingdom)
n.tamayo@roehampton.ac.uk

London, 9 December 2022

By request of Dr. González-Aguilera, here I present my report on Mr. Lloyd Courtenay's PhD thesis entitled "*Robust Statistical and Artificially Intelligent Approaches for the Analysis of 2D and 3D Morphological Data*" and supervised by Dr. Diego González-Aguilera and Dr. José Yravedra in the Escuela Politécnica Superior de Ávila (Universidad de Salamanca).

This PhD thesis presents an innovative investigation that combines Geometric Morphometrics, Fourier Shape Descriptors and Artificially Intelligent Algorithms for generation, handling and quantification of 2D and 3D morphological data in different contexts. The outcomes of this PhD work demonstrate that the use of these advanced computational methodologies reduces the subjectivity in the analyses of complex and large morphological datasets. This PhD thesis successfully addresses the issue of small sample sizes in paleoanthropology and archaeology by augmentation and simulation of data. Also, a relevant finding of this PhD thesis is the use of Artificially Intelligent Algorithms such as Neural Networks and Support Vector Machines for the (identification of new specimens). These are very important outcomes in the context of accuracy, reproducibility and reliability of the analysis of morphological data.

This PhD thesis is organized in a series of logical and coherent steps that facilitates its comprehension and the applicability of its results. The theoretical framework of this PhD research is very well explained and the use of these methodologies for the aims is well justified. The discussion is very well developed, and I would like to emphasize the exercise in honesty carried out by the candidate by discussing the advantages and disadvantages of the different techniques. On top of that, this PhD reflects a rigorous revision of the scientific literature, as the candidate not only discusses classic works about the topic, but also very recently published investigations, showing an excellent handling of the existing literature.

The candidate shows the applicability of the methodologies developed in this PhD thesis not only to different fields often related to each other, such as archaeology, taphonomy and paleoanthropology, but also to ecology and medicine, with important applications to oncology. This shows the great potential and utility of the outcomes of this PhD thesis in different research fields, which is, in my opinion, one of the most important achievements in any PhD thesis.

The transdisciplinarity of this PhD thesis is reflected in nine research works that have been already published or are in the process of publication in international journals: six articles have been published in high impact international journals including Applied Sciences, PloSONE, Scientific Reports, Animals and Journal of Clinical Medicine and three are under review in American Journal of Biological Anthropology, Quaternary Sciences Reviews and Systematic Biology by the time of this PhD thesis submission. In all these publications, the candidate is the first author or first co-author, reflecting leadership skills in conceptualization, analyses and writing of the manuscripts and associated PhD chapters in an international context.

In view of the dates of publication of the articles included in this PhD thesis, I assume that a great part of the PhD work has been carried out during the COVID-19 pandemic. This pandemic has highly impacted the scientific performance of many scientists whose research directly depended on the access to the infrastructure and material necessary for their research, bringing the importance of the use of open resources to the fore. In this regard, I would like to highlight the effort made by the candidate in establishing a workflow based on open-source software in his PhD research, making the code and data available in open online repositories. This is a great contribution of this PhD thesis, not only because it promotes transparency and reproducibility in science, but also because it lays the groundwork for future researchers to continue building their research on the outcomes reported therein. As a researcher working on virtual morphology, I am very looking forward to putting into practice the methodologies detailed in this PhD research.

I can confirm that this PhD thesis meets the requirements to be submitted to the Escuela de Doctorado of the Universidad de Salamanca as part of the Programa de Doctorado interuniversitario en Geotecnologías aplicadas a la Construcción, Energía e Industria, within the framework of an international PhD and to be defended by the candidate and evaluated by a PhD committee of international experts. I consider that the PhD thesis, together with the scientific articles that are part of it, show that the candidate is ready to become an independent researcher in a post-doctoral stage.

I remain at your disposal to answer any question you may have.

Yours faithfully,

Dr. Nicole Torres-Tamayo
School of Life and Health Sciences
University of Roehampton (London, UK)
E-mail: n.tamayo@roehampton.ac.uk

# List of Scientific Publications

The present doctoral thesis consists of a compendium of seven scientific articles published, and two in the process of being published, in international journals of high scientific impact. These include (in chronological order);

### A Hybrid Geometric Morphometric Deep Learning Approach for Cut and Trampling Mark Classificaiton

Lloyd A. Courtenay[1,2,3], Rosa Huguet[2,3,4], Diego González-Aguilera[1], José Yravedra[5,6]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Institut Català de Paleoecologia Humana I Evolució Social (IPHES), Zona Educacional 4, Campus Sescelades URV (Edifici W3) E3, 43700, Tarragona, Spain. [3]Universitat de Rovira I Virgili (URV), Department d'Historia i Hostoria de l'Art, Avignuda de Catalunya 35, 43002, Tarragona, Spain. [4]Unit Associated to CSIC, Departamento de Paleobiologia, Museo de Ciencias Naturales, calle José Gutiérrez Abascal, s/n, 28006 Madrid, Spain. [5]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [6]C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren s/n, 28040 Madrid, Spain.

**Journal**: Applied Sciences
**Publication Date**: 23 December, 2019
**DOI**: https://doi.org/10.3390/app10010150

### Obtaining New Resolutions in Carnivore Tooth Pit Morphological Analyses: A Methodological Update for Digital Taphonomy

Lloyd A. Courtenay[1], Darío Herranz-Rodrigo[2,3], Rosa Huguet[4,5,6], Miguel Ángel Maté-González[1,7,8], Diego González-Aguilera[1,7], José Yravedra[2,3].

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [3]C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren s/n, 28040 Madrid, Spain. [4]Institut Català de Paleoecologia Humana I Evolució Social (IPHES), Zona Educacional 4, Campus Sescelades URV (Edifici W3) E3, 43700, Tarragona, Spain. [5]Universitat de Rovira I Virgili (URV), Department d'Historia i Hostoria de l'Art, Avignuda de Catalunya 35, 43002, Tarragona, Spain. [6]Unit Associated to CSIC, Departamento de Paleobiologia, Museo de Ciencias Naturales, calle José Gutiérrez Abascal, s/n, 28006 Madrid, Spain. [7]Gran Duque de Alba Institution, Diputación Provincial de Ávila, Paseo Dos de Mayo, 8, 05002, Ávila, Spain. [8]Department of Topographic and Cartography Engineering, Higher Technical School of Engineers in Topography, Geodesy and Cartography, Technical University of Madrid, Madrid, Spain.

**Journal**: PLoS ONE
**Publication Date**: October 8, 2020
**DOI**: https://doi.org/10.1371/journal.pone.0240328

## Geometric Morphometric Data Augmentation Using Generative Computational Learning Algorithms

Lloyd A. Courtenay[1], Diego González-Aguilera[1]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain.

**Journal**: Applied Sciences
**Publication Date**: December 21, 2020
**DOI**: https://doi.org/10.3390/app10249133

## Developments in Data Science Solutions for Carnivore Tooth Pit Classification

Lloyd A. Courtenay[1], Darío Herranz-Rodrigo[2,3], Diego González-Aguilera[1], José Yravedra[2,3]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [3]C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren s/n, 28040 Madrid, Spain.

**Journal**: Scientific Reports
**Publication Date**: 13 May, 2021
**DOI**: https://doi.org/10.1038/s41598-021-89518-4

## 3D Insights into the Effect of Captivity on Wolf Mastication and Their Tooth Marks; Implications in Ecological Studies of Both Past and Present

Lloyd A. Courtenay[1], Darío Herranz-Rodrigo[2,3], José Yravedra[2,3], José Mª Vázquez-Rodríguez[4], Rosa Huguet[5,6,7], Isabel Barja[8,9], Miguel Ángel Maté-González[1,10], Maximiliano Fernández-Fernández[11,12], Ángel-Luis Muñoz-Nieto[1], Diego González-Aguilera[1,11]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [3]C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren s/n, 28040 Madrid, Spain. [4]Department of Prehistory and Archaeology, Humanities Faculty, UNED University, C/Senda del Rey, 7, 28040, Madrid, Spain. [5]Institut Català de Paleoecologia Humana I Evolució Social (IPHES), Zona Educacional 4, Campus Sescelades URV (Edifici W3) E3, 43700, Tarragona, Spain. [6]Universitat de Rovira I Virgili (URV), Department d'Historia i Hostoria de l'Art, Avignuda de Catalunya 35, 43002, Tarragona, Spain. [7]Unit Associated to CSIC, Departamento de Paleobiologia, Museo de Ciencias Naturales, calle José Gutiérrez Abascal, s/n, 28006 Madrid, Spain. [8]Zoology Unit, Department of Biology, Autónoma University of Madrid, C/Darwin 2, Campus Universitario de Cantoblanco, 28049 Madrid, Spain. [9]Center of Investigation in Biodiversity and Global Change (CIBC-UAM), Universidad Autónoma de Madrid, 28049 Madrid, Spain [10]Department of Topographic and Cartography Engineering, Higher Technical School of Engineers in Topography, Geodesy and Cartography, Technical University of Madrid, Madrid, Spain. [11]Gran Duque de Alba Institution, Diputación Provincial de Ávila, Paseo Dos de Mayo, 8, 05002, Ávila, Spain. [12]Department of Sciences of Communication and Sociology, Faculty of Communication Sciences, University Rey Juan Carlos, Camino del Molino, s/n, 28943 Madrid, Spain.

**Journal**: Animals
**Publication Date**: August 6, 2021
**DOI**: https://doi.org/10.3390/ani11082323

## A Novel Approach for the Shape Characterisation of Non-Melanoma Skin Lesions using Elliptic Fourier Analyses and Clinical Images

Lloyd A. Courtenay[1], Inés Barbero-García[1], Julia Aramendi[1,2], Diego González-Aguilera[1], Manuel Rodríguez-Martín[3], Pablo Rodríguez-Gonzalvez[4], Javier Cañueto[5,6,7], Concepción Román-Curto[5,6]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Geology, Facultad de Ciencia y Tecnología, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU), Barrio Sarriena s/n, 48940 Leoia, Spain. [3]Department of Mechanical Engineering, Universidad de Salamanca, Zamora, 49029, Spain. [4]Department of Mining Technology, Topography and Structures, University of León, Ponferrada, León, Spain. [5]Department of Dermatology, University Hospital of Spain, Paseo de San Vicente 58-182, 37007, Salamanca, Spain [6]Instituto de Investigación Biomédica de Salamanca (IBSAL), Paseo de San Vicente, 58-182, 37007, Salamanca, Spain [6]Instituto de Biología Molecular y Celular del Cáncer (IBMCC)/Centro de Investigación del Cáncer (lab 7). Campus Miguel de Unamuno s/n. 37007 Salamanca.

**Journal**: Journal of Clinical Medicine
**Publication Date**: July 28, 2022
**DOI**: `https://doi.org/10.3390/jcm11154392`

## Unraveling carnivore competition for animal resources at the 1.46 Ma Early Pleistocene Site of Barranco León (Orce, Granada, Spain)

Lloyd A. Courtenay[1,2], José Yravedra[2,3,4,5], Darío Herranz-Rodrigo[2,3], Juan José Rodríguez-Alba[2], Alexia Serrano-Ramos[6], Verónica Estaca-Gómez[2], Diego González-Aguilera[1], José Antonio Solano[6,7], Juan Manuel Jiménez-Arenas[6,7,8]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [3]C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren s/n, 28040 Madrid, Spain. [4]Grupo de Investigación Ecosistemas Cuaternarios. Complutense University, Prof. Aranguren s/n, 28040, Madrid, Spain. [5]Grupo de Investigación Arqueología Prehistórica. Complutense University, Prof. Aranguren s/n, 28040, Madrid, Spain. [6]Department of Prehistory and Archaeology, University of Granada, Campus Universitario de Cartuja, 18071, Granada, Spain. [7]Museum Primeros Pobladores de Europa "Josep Gibert", Cam. San Simon, 18858, Orce, Granada, Spain [8]Institute of Peace and Conflict Research, University of Granada, C/ Rector López Argüeta s/n, 18001, Granada, Spain

**Journal**: Quaternary Science Reviews
**Publication Date**: December 10, 2022
**DOI**: `https://doi.org/10.1016/j.quascirev.2022.107912`

## Recruiting a Skeleton Crew – Methods for Simulating and Augmenting Palaeoanthropological Data using Monte Carlo based Algorithms.

Lloyd A. Courtenay[1,2], Julia Aramendi[3], Diego González-Aguilera[1]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [3]Department of Geology, Facultad de Ciencia y Tecnología, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa, Spain.

**Under Review in**: American Journal of Biological Anthropology

## A Graph Based Geometric Morphometric approach to the analysis of primate radii: A new mathematical model for the processing of landmark data.

Lloyd A. Courtenay[1,2], Julia Aramendi[3], Diego González-Aguilera[1]

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain. [2]Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain. [3]Department of Geology, Facultad de Ciencia y Tecnología, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa, Spain.

**Under Review in**: Journal of Anatomy

# Abstract

Notions of geometry and morphology are a fundamental component of the way we perceive, describe, and essentially interact with objects. The shape and size of an element can be highly informative, and will thus condition the way we carry out basic functions in our daily lives. Morphology can be useful for; the detection of anomalies and patterns, the characterisation of an object or organism, as well as the identification of casuality (cause and effect). Nevertheless, finding an efficient and objective means of characterising morphology in both micro and macroscopic elements is often a great challenge in many fields of science. While many approaches to these types of tasks have long relied on a visual or qualitative description of shape and form, unfortunately most of these methods are influenced by notable degrees of subjectivity, typically product of human-based error and dependent on experience and perspective.

In the present Doctoral Thesis, a wide array of different techniques for the extraction and analysis of morphological data is explored and discussed. Specifically, the main goals of this study are to define a general workflow that can be used for the quantification of different elements, with the hope of developing a transparent and transdisciplinary approach that can be applied to many fields of science.

For this purpose, multiple techniques for the digitisation of both 2D and 3D data have been used, including (in order of prevalence); structured light surface scanning, micro-computed tomography, 3D-digital microscopy, and traditional photography for clinical imaging. Investigation into the best means of extracting morphological information was then explored, primarily including geometric morphometric analyses, while also testing combinations of traditional metric and elliptic Fourier analyses as well. Following this, the present Doctoral Thesis dedicates a significant portion of research into finding the best means of statistically analysing this type of information. Here the use of multiple robust statistical approaches is employed, as well as both parametric and non-parametric testing, in order to obtain the highest possible accuracy in the conclusions withdrawn.

So as to leverage this information for tasks such as classification or diagnostics, this Doctoral Thesis also focuses great attention on the use of Computational Learning for the development of Artificially Intelligent algorithms in decision-making tasks. Prior to the development of classification algorithms, however, research delved into the possible limitations present in data science. Namely, problems due to sample size and the "curse of dimensionality". So as to overcome these limitations, different research lines were developed; first exploring the multitude of available techniques for data augmentation and simulation, followed by experimentation with different types of algorithms for classification tasks.

The results obtained from this study reveal many different techniques to be useful for the modelling, extraction, and study of morphological information. Here it has been shown in a variety of different scenarios how, especially when combined with robust statistical approaches, both geometric morphometrics and elliptic Fourier analyses are powerful tools for the description of shape and form. Throughout this research, data simulation has also proven to be a fundamental step in the workflow, providing Artificially Intelligent Algorithms such as Neural Networks and Support Vector Machines sufficient information for the identification of new specimens.

So as to promote transparency and improve reproducibility, the present doctoral thesis is also accompanied by a large collection of open-source code, datasets, and different software.

The applicability of these methodological approaches, and thus their transdisciplinary nature, has been demonstrated across multiple case studies. These include applications in archaeological and palaeontological taphonomy, palaeoanthropology, animal wellfare, and dermatology. Through this compendium of research articles, the presented methods have been able to discover a number of different features from each of these fields. These range from the ability to identify extinct carnivore taxa in archaeological sites based on their tooth marks, to the first empirical quantification of skin lesion asymmetry as a diagnostic tool in dermatological oncology. In addition, through the presentation of a new mathematical model for the description of morphology, this study has been able to provide a new, more efficient, means of extracting biomechanical information from great primate limb long bones. Each of these discoveries present promising advantages for the study of other types of morphological data as well.

The present Doctoral Thesis thus hopes to provide a new perspective on the means in which the morphology of different elements can be studied, promoting a more robust, transdisciplinary approach. Through this, future research will focus on applying these techniques to other fields of science, while working on fine tuning this methodological workflow to obtain higher precision and accuracy.

# Resumen

Las nociones de geometría y morfología son un componente fundamental del modo en que percibimos, describimos y esencialmente interactuamos con los objetos. La forma y el tamaño de un elemento pueden ser altamente informativos, y condicionan el modo en que realizamos funciones básicas en nuestra vida cotidiana. La morfología puede ser útil para; la detección de anomalías y patrones, la caracterización de un objeto u organismo, así como la identificación de la casualidad (causa y efecto). Sin embargo, encontrar un acercamiento eficaz y objetivo de caracterizar la morfología de los elementos micro y macroscópicos suele ser un gran reto en muchos campos de la ciencia. Aunque muchos enfoques de este tipo de tareas se han basado durante mucho tiempo en una descripción visual o cualitativa de la forma, lamentablemente la mayoría de estos métodos están influenciados por notables grados de subjetividad, típicamente producto del error humano y dependiente de la experiencia y el conocimiento.

En la presente Tesis Doctoral, se explora una amplia gama de diferentes técnicas para la extracción y el análisis de datos morfológicos. En concreto, los principales objetivos de este estudio son definir un flujo de trabajo general que pueda ser utilizado para la cuantificación de diferentes elementos, con la esperanza de desarrollar un enfoque transparente e transdisciplinar que pueda aplicarse a muchos campos de la ciencia.

Para ello, se han utilizado múltiples técnicas de digitalización de datos tanto en 2D como en 3D, entre ellas (por orden de prevalencia): escaneo de superficies con luz estructurada, tomografía microcomputada, microscopía digital 3D y la fotografía para la obtención de imágenes clínicas. A continuación, se investigó cuál era el mejor medio para extraer información morfológica, principalmente mediante análisis de Morfometría Geométrica, aunque también se probaron combinaciones de análisis métricos tradicionales y avanzados basados en análisis de Fourier elípticos. A raíz de esto, la presente Tesis Doctoral dedica una parte importante de la investigación a la búsqueda de los mejores medios de análisis estadístico de este tipo de información. Más concretamente, se emplea el uso de múltiples enfoques estadísticos robustos, así como pruebas paramétricas y no paramétricas, con el fin de obtener la mayor precisión posible en las conclusiones extraídas.

Con el fin de aprovechar esta información para tareas como la clasificación o el diagnóstico, esta Tesis Doctoral también presta gran atención al uso del Aprendizaje Computacional para el desarrollo de algoritmos de la Inteligencia Artificial en tareas de toma de decisiones. Sin embargo, antes del desarrollo de algoritmos de clasificación, la investigación profundizó en las posibles limitaciones presentes en la ciencia de datos. A saber, los problemas debidos al tamaño de la muestra y la "maldición de la dimensionalidad". Para superar estas limitaciones, se desarrollaron diferentes líneas de investigación; primero explorando la multitud de técnicas disponibles para el aumento de datos y la simulación; seguido por la experimentación con diferentes tipos de algoritmos para tareas de clasificación.

Los resultados obtenidos en este estudio revelan que muchas técnicas diferentes son útiles para el modelado, extracción y estudio de la información morfológica. Aquí se ha demostrado en una variedad de diferentes escenarios diferentes cómo, especialmente cuando se combinan con enfoques estadísticos robustos, tanto la morfometría geométrica y los análisis de Fourier elípticos son herramientas poderosas para la descripción de la forma. A lo largo de esta investigación, la simulación de datos también ha demostrado ser un paso fundamental en el flujo de trabajo, proporcionando mediante técnicas de inteligencia artificial (ej. redes neuronales, máquina de soporte de vectores) información suficiente para la identificación de nuevos especímenes.

Con el fin de promover la transparencia y mejorar la reproducibilidad, la presente Tesis Doctoral también va acompañada de una amplia colección de código abierto, conjuntos de datos y diferentes programas informáticos.

La aplicabilidad de estos enfoques metodológicos, y por tanto, su carácter transdisciplinar, ha quedado demostrado a través de múltiples casos de estudio validados con éxito. Entre ellos se incluyen aplicaciones en arqueología y paleontología, la paleoantropología, el bienestar animal, y la dermatología. A través de este compendio de artículos de investigación, los métodos presentados han sido capaces de contribuir con una serie de características en cada uno de estos campos. Estas van desde la capacidad de identificar taxones de carnívoros extintos en yacimientos arqueológicos basándose en sus marcas dentales, hasta la primera cuantificación empírica de la asimetría de las lesiones cutáneas como herramienta de diagnóstico en oncología dermatológica. Además, mediante la presentación de un nuevo modelo matemático para la descripción de la morfología, este estudio ha sido capaz de proporcionar un nuevo medio, más eficiente, de extraer información biomecánica de los huesos largos de las extremidades de los grandes primates. En definitiva, cada uno de estos descubrimientos presenta prometedoras ventajas para el estudio de otros tipos de datos morfológicos.

La presente Tesis Doctoral pretende, por tanto, aportar una nueva perspectiva sobre los medios con los que se puede estudiar la morfología de diferentes elementos, promoviendo un enfoque más robusto y transdisciplinar. Por ello, las investigaciones futuras se centrarán en aplicar estas técnicas a otros campos de la ciencia, al tiempo que permitan trabajar en el ajuste de este flujo de trabajo metodológico para obtener una mayor precisión y exactitud.

# Funding

This doctoral thesis was primarily funded by the Spanish Ministry of Science and Innovation, the European Regional Development Fund, and the Spanish State Investigation Agency, in the public initiative titled *Programa Estatal de Promoción de Talento y su Empleabilided en I+D+i del Plan Estatal de Investigación Científica y Técnica y de Innovación 2017-2020*, in the 2019 call for funding. This FPI Predoctoral Grant, with reference number PRE2019-089411, is associated to the National project RTI2018-099850-B-I00, led by Prof. Diegto González-Aguilera and Dr. Manuel Rodríguez-Martín of the University of Salamanca, and in collaboration with the University of Cantabria. The associated research project is titled *FaTIMA: Fusiıon de tecnologías no destructivas y métodos de simulación numérica para la inspección y monitorización en uniones de nuevos materiales y en procesos de fabriación aditiva*.

Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the present Doctoral Thesis.

# Aknowledgements

The time that has passed since beginning my Doctoral Thesis in October, 2019, has felt long, and has been far from easy. None of this would have been possible, however, without the support of so many different people I have met along the way.

First of all, I would like to reflect on all of the individuals who have formed a fundamental part of my formative years as an academic. I would like to thank all of the friends and colleagues I made over the course of my degree at the Complutense, and express a great deal of gratitude as well to all the staff and members of the Institut Catalá de Paleoecologia Humana I Evolució Social. While there are too many to acknowledge from both institutions, I would especially like to thank Andreu Ollé for many of the comments, debates and discussions we have had over the years, and I would like to express my utmost appreciation towards Rosa Huguet, who remains to be the main person I turn to when I need some sound advice and motivation! It is hard to express how much I appreciate the contributions she has made to my work and my growth as a professional.

Since beginning my PhD in Ávila, I have met a great deal of people who have also helped me professionally along the way, both in Spain and many as far away as Africa. First of all, I would like to thank all of the colleagues I have worked with in the TIDOP Research Group here in Ávila. I don't think I could have come as far as I have without the memorable experience of working with such a tight group of engineers. The many conversations with colleagues (and friends, over the "occasional" pint), such as Esteban Ruíz de Oña and Alberto Morcillo Sanz, have always been enlightening, and have definitely inspired me to look beyond my training as an archaeologist and delve further into the world of programming. Likewise, the opportunity to hear about all the different lines of research and the ways in which they can be applied, has probably been my biggest inspiration in searching for new ways to apply a wide array of different techniques to archaeology.

My first steps into the world of robust statistics, for example, were thanks to the guidance of one of my supervisors, Diego González-Aguilera. I would also like to thank, however, other people such as Susana del Pozo, who was there with me when we began trying to find the best way of using this type of statistics at the beginning of the cancer research projects. I also remember hearing about the research of many other TIDOP members, who have helped me begin to look more into other lines of research beyond my own, which I would have never thought about otherwise, and which have eventually become a fundamental part of the research I carry out today.

Along these lines, I must also thank the individuals who have helped me brush up on my mathematics over the years. In particular I would like to thank Josué Casado Rabasco.

Outside of Ávila, I have had the fantastic opportunity to travel, meet, and work with some incredible researchers from other institutions. My first stay at another research center was in the wonderful city of Granada, a city I fell in love with due to its amazing atmosphere and cultural peculiarities. Thinking of my stay in Granada, I would like to thank Juan Manuel Jiménez Arenas, the Orce research team, and the Archaeological Museum of Granada, for the invitation and the opportunity to work on such an amazing group of sites. Along these lines, I would like to extend my thanks to all those who have participated in the excavation and investigation of the Orce sites.

Following Granada, the invitation to visit the University of Bordeaux and work with the incredible finds from the Lower Pleistocene sites in Omo (Ethiopia), was one of the best opportunities and experiences I have been given over the last couple of years. In light of this, I would like to thank Jean-Philip Brugal for funding my first trip, and for his work leading the TAPHEN research initiative that has brought together so many international taphonomists. More importantly, I would like to express my gratitude to Antoine Souron for his enthusiasm and interest in my work. Not only is Antoine a colleague who I enjoy working with, and look forward to continue working with in the foreseeable future, but he is also someone I am happy to call a friend. A trip to Bordeaux would not be complete without spending some time with him and his very welcoming family, enjoying Ethiopian food, gossiping about experiences working in Africa, and, as always, having the "occasional" pint.

From a slightly different perspective, I would most certainly like to thank Fernando Díez-Martín. Although my academic trajectory has changed quite a lot since the beginning of my PhD, of which Fernando formed an important part at the very beginning, I am incredibly grateful to him for his support and encouragement, despite some unfortunately complicated circumstances. I hope to soon return to the amazing sites at Olduvai Gorge, Tanzania, so as to continue my work with him, as well as all the other incredible people who I have met in Tanzania since my first visits to the Gorge in early 2019.

A lot of this would not be possible without my directors, for whom I am very grateful. First of all, I would like to express my never-ending appreciation to Diego González-Aguilera, without whom this Doctoral Thesis would not have happened, or at least would have ended up being completely different! I truly believe that a significant part of this studies' quality is thanks to his support, inspiration, and very useful suggestions. Diego played a fundamental role in helping me get funding, and, most of the time, has had to put up with the very bizarre ideas I usually concoct for new research topics. I'm not sure many other engineers would have been as open-minded. Likewise, I would like to thank José Yravedra, who has been guiding and teaching me since the very beginning of my career as a researcher in 2016. We have been through quite a lot of difficult times, but I am indebted to him for his support and advice. For a long time, José has been one of my greatest inspirations and mentor. Therefore I will be forever grateful to him for providing me with so many opportunities over the years.

Needless to say, beyond the more professional aspects of my life, I would not be here without the incredible friends I have made, many of whom I now consider part of my extended family and could not imagine a life without. My friends are some of the people who have had to put up with me the most, remain unfalteringly loyal, and who always seem to find a way to pick me up and make life more enjoyable, no matter how difficult things can get.

Beginning with my life here in Ávila, I would like to thank the rather odd assortment of individuals who we have now come to label "DOPIF". Without DOPIF, life in Ávila would have been ridiculously tedious, especially without our little social club for lunches. I would like to particularly thank my close friend Innes, who has definitely lightened things up since the moment she arrived. Just as importantly, I would like to thank Pedro and Laurah (*mis hermanos brasileños*); no weekend has been complete without seeing the sun come up from outside of their apartment (*who can it be now?*).

Thinking back, I have very fond memories of many, many people. From Madrid, I am particularly grateful to Laura, Ro, and Fran, who have kept in touch since we finished our degree. I am also thankful to the friendship I have with Abel, who has been a person I have often confided in over the years. Likewise, living with Noé, Miguel, and Andrea, in Tarragona, forms part of some of the greatest memories I have. More memorable experiences include going to festivals with Sara and Juan, numerous trips to Galicia to see Noé and his friends, and the COVID-19 Skype meetings we used to have with "Máster", Zaira, and many other friends. As can be expected, the "occasional" pint was involved.

Among the many journeys I have made, each of the trips would not have been complete without the large social circles I always seem to have found myself in. I am always excited to see Blade and Aline on my trips to Bordeaux, while Elena and Maria made my stays in the Archaeological Museum of Granada more enjoyable. David and Noelia, whom I both met on my trips *por debajo de despeñaperros*, are not only some of the kindest people I know, but are are some of my best friends. They frequently remind me that my Spanish isn't *that* good because I often feel like I need subtitles.

I cannot thank you all enough for helping me keep my sanity.

My highest level of gratitude goes towards my closest friends, some old, most new. I am thankful to Juan and his family for lightening up my weekends in the Rastro, for the many, many concerts we went to together, and for the great laughs we shared (mostly at our own expense). Julia, who has been a colleague for many years, but has since (in her own words) "forced" us to be friends (not that I needed persuading), is not only a person whose professional opinion I value the most, but whose friendship and support has been a major pillar for me; I am very grateful for our friendship. Finally, I thank Noé and Enrique... for pretty much everything. There are far too many things to write down, but I'm sure they are aware of how much I appreciate them. They are *always* there for me.

Last but not least, I want to express my admiration to my family.

First of all, I thank the entire Courtenay family whose name I bear so proudly. Auntie Sheila and Uncle Barrie, you mean the world to me, and although I don't visit the UK much, I think about you all the time. John, thank you for being there for us, and most of all, for putting up with the crazy household you seem to want to join! Jordan, you are one of the few people who seems to really understand me. I am not just grateful to you for being there for me; I also just want you to know how proud I am of you for who you are as a person.

Finally, and most importantly, I would like to express my greatest appreciation to the two most important people in my life, for whom I would do anything to make proud. To quote from *The Book Thief*, this work is dedicated to "those who gave me eyes to see";

*The person who is always there for me*

**Liz Courtenay**

*The person who I miss more than anything in the world*

**Ginger Courtenay**

*"Good luck exploring the infinite abyss."*

Zach Braff (a.k.a. Andrew Largeman)
- *Garden State* -

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| ADASYN | Adaptive Synthetic Sampling |
| AI | Artificial Intelligence |
| AIA | Artificially Intelligen Algorithms |
| AMH | Anatomically Modern Human |
| ANN | Artificial Neural Network |
| ANOVA | Analysis of Variance |
| ARB | Abnormal Repetitive Behaviours |
| ASA | American Statistical Association |
| AUC | Area Under Curve |
| BCC | Basal Cell Carcinoma |
| BEN | Benign skin lession |
| BF | Bayes Factors |
| BFB | Bayes Factor Bounds |
| bpPCA | Between-Group Principal Component Analyses |
| BL | Barranco León |
| BOA | Bayesian Optimization Algorithm |
| BSM | Bone Surface Modification |
| BWMV | Biweight Midvariance |
| CI | Confidence Intervals |
| CL | Computational Learning |
| CGAN | Conditional Generative Adversarial Network |
| CNN | Convolutional Neural Network |
| CS | Centroid Size |
| CT | Computed Tomographic |
| CTREE | Conditional Inference Tree |
| CV | Computer Vision |
| CVA | Canonical Variance Analysis |
| D | Depth or Distance (depending on context) |
| DBSCAN | Density-Based Spatial Clustering Algorithm with Noise |
| DL | Deep Learning |
| DR | Dimensionality Reduction |
| EDMA | Euclidean Distance Matrix Analysis |
| EFA | Elliptic Fourier Analysis |
| EI | Expected Improvement |
| FA | Fourier Analysis |
| FAMD | Factor Analysis of Mixed Data |
| FFNN | Feed Forward Neural Network |
| FN | False Negative |
| FOV | Field Of View |
| FP | False Positive |
| FPR | False Positive Risk |
| FRV | Fourier Radius Variation |
| FSD | Fourier Shape Descriptors |
| FTA | Fourier Tangent Angle |

| | |
|---|---|
| GAN | Generative Adversarial Network |
| GB | Gradient Boosting |
| GCN | Graph Convolutional Network |
| GPA | Generalised Procrustes Analysis |
| GPUCB | Gaussian Process Upper Confidence Bound |
| GMM | Geometric Morphometrics (or GM) |
| GRF | Generalised Resistant Fit |
| GRoF | Generalised Robust Fit |
| GUI | Graphical User Interface |
| HPD | Highest Posterior Density |
| IEC | Intraepithetal Carcinoma |
| IF | Isolation Forest |
| IR | Interquantile Range |
| JCI | Journal Citation Indicator |
| JIF | Journal Impact Factor |
| KNN | *k*-Nearest Neighbour |
| LDA | Linear Discriminant Analysis |
| LDC | Left Depth Convergent |
| LSGAN | Least Squares Loss Generative Adversarial Network |
| LM | Landmark |
| LOOCV | Leave-One-Out Cross-Validation |
| MAD | Median Absolute Deviation |
| MANOVA | Multivariate Analysis of Variance |
| MC | Monte Carlo |
| MCMC | Markov Chain Monte Carlo |
| $\mu$-CT | micro-Computed Tomography |
| ML | Machine Learning |
| MS | Mean Shift |
| MSE | Mean Squared Error |
| MT | Machine Teaching |
| NB | Naïve Bayes |
| NISP | Number of Identifiable Specimens |
| NJ | Neighbour Joining |
| NMAD | Normalised Median Absolute Deviation |
| NMSC | Non-Melanoma Skin Cancer |
| NN | Neural Network |
| NSVM | Neural Support Vector Machine |
| OA | Opening Angle |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| PDF | Probability Density Function |
| PLSDA | Partial Least Square Discriminant Analysis |
| RDC | Right Depth Convergent |
| ReLU | Rectified Linear Unit |
| RF | Random Forest |
| RFF | Random Fourier Function |
| RM | Repeatability Measure |
| RMSE | Root Mean Squared Error |
| ROC | Receiver Operating Characteristic |
| rTOST | robust Two One-Sided equivalency Test |
| RWA | Relative Warp Analysis |
| SCC | Squamous Cell Carcinoma |
| SMOTE | Synthetic Minority Oversampling Technique |
| SSLS | Structured Light Surface Scanning |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machine |

| | |
|---|---|
| TN | True Negative |
| TOST | Two One-Sided equivalency Test |
| TP | True Positive |
| TPS | Thin Plate Spline |
| t-SNE | t-Distributed Stochastic Neighbour Embedding |
| UMAP | Uniform Manifold Approximation and Projection |
| UPGMA | Unweighted Pair Group Method with Arithmetic Mean |
| VM3 | Venta Micena 3 |
| WIB | Width of Incistion in proximity with the Base |
| WIM | Width of Incision Midway |
| WIS | Width of Incision at Surface |
| WGAN | Wasserstein Loss Generative Adversarial Network |
| WGAN-GP | Gradient Penalty Wasserstein Generative Adversarial Network |

# Chapter 1

# Presentation of Doctoral Thesis

## 1.1 Presentation

This Doctoral Thesis, titled *Robust Statistical and Artificially Intelligent Approaches for the Analysis of 2D and 3D Morphological Data*, presents approximately 4 years of research with a background that stretches back to the mid-2010's. While the focus of this line of research has shifted over the years, providing a greater protagonism to engineering and data science, as opposed to archaeological science, the objectives of the present study are to develop tools that are widely applicable to different case studies, thus hoping to benefit the scientific community as a whole.

Finding efficiency and objectivity in the inspection and analysis of micro and macroscopic elements is a great challenge in many fields of science. For many years, multiple lines of research have relied on the visual (*in-visu*) description and characterisation of an element's morphology. Unfortunately, most of these methods are subject to significant degrees of human-based error that could be a product of experience, perspective, and the general subjectivity induced by the analyst.

The most innovative techniques that provide a means of overcoming these issues have been made available through the use of advanced Computer Vision applications, including the generation and handling of three-dimensional data. Both 2D and 3D information are a valuable source of data that can overcome numerous limitations through their generally non-destructive nature. Furthermore, the wide array of different techniques for the creation and obtaining of this data make many types of analyses easy to compute, fast, and geometrically quantifiable. Through this, researchers are required to adapt according to; (1) the methods available for obtaining this information, (2) the means in which data is extracted from them, and (3) the techniques used to process the final dataset.

Geomatics engineering and remote sensing are valuable fields that have greatly contributed to research of this nature. Remote sensing provides multiple means of modelling real world elements with increased resolution, regardless of the object's shape or size. These disciplines work best when combining different imaging techniques and sensors. From a different perspective, Geometric Morphometrics (GMM) is a powerful mathematical tool for the quantification of different morphologies, having had great success in multiple fields, such as systematic biology, physical anthropology, palaeoanthropology and archaeology, among others. GMMs allows for the quantification of both 2D and 3D elements using sets of homologous Cartesian coordinates, known as *landmarks*, that describe the general morphology of an element under study. In a similar light, Fourier Shape Descriptors (FSD) are also useful mathematical tools for the decomposition of shape into a set of manageable variables, especially in cases where landmarks are either hard or impossible to locate. Finally, Artificially Intelligent Algorithms (AIAs) and the field of computational learning provide an efficient means of processing highly complex and noisy datasets with greater accuracy and objectivity.

The combination of techniques from different fields of research is a very common practice in science. It should even be argued that collaboration is fundamental for research to grow and evolve. Nevertheless, the degree of interaction across disciplines is often limited, and can be referred to as *multidisciplinary* research. *Intedisciplinary* research, on the other hand, is used to refer to collaborations where integration between the disciplines is much greater, and finally *Transdisciplinarity* refers to research when two or more disciplinary perspectives transcend to form a new holistic approach. As will be shown throughout this body of research, the methodological approaches described here are applicable to a wide range of different applications, with case-studies in fields of archaeology, taphonomy, palaeoanthropology, modern day ecology, and even medicine. This Doctoral Thesis can thus be considered a development and valuable contribution to modern day research in many fields, promoting a more *transdisciplinary* approach to scientific analyses.

## 1.2 Objectives and Hypotheses

The objectives of the present Doctoral Thesis are to exploit methods for the extraction and analysis of morphological data, as well as the use of computational learning for the processing of this information, with the intention of objectively quantifying different morphologies. Here, numerous different mathematical and statistical models will be explored in a number of transdisciplinary applications. This body of work can thus be summarised in the investigation of the following primary objectives;

- The development of new techniques for the detection and modelling of micro- and macroscopic elements.
- The reduction of analyst induced subjectivity, employing methods from applied mathematics, robust statistics, and computational sciences, for the processing of highly complex and large datasets with greater efficiency and objectivity.
- Through points 1 and 2, this Doctoral Thesis will provide innovative approaches for the quantification of morphology, essentially presenting a methodology that can be incorporated into numerous disciplines of applied sciences, proposing a more transdisciplinary approach to scientific research.

As a bi-product, this body of research also aims to fulfill the following secondary objectives;

- The development of open-source software for the study of morphological data.
- The promotion of transparency and reproducibility in each of the methodological approaches.

In fulfilling these objectives, this body of research aims to answer the following parting hypotheses;

1. The use of advanced computational resources and robust statistical techniques can greatly improve the accuracy of studies that are typically performed using visual and qualitative criteria.

2. Mathematical tools, that are typically used in biological and anthropological contexts for the quantification of different morphologies, can be useful for the study of non-biological elements as well.

3. Artificially Intelligent Algorithms, especially when fuelled by advanced numeric and categorical simulation techniques, can prove a very powerful tool in the classification of elements based on their morphology.

## 1.3 Structure

In this first chapter, the objectives of the Doctoral Thesis are explained, followed by their theoretical justification in Chapter 2, specifying the advantages these types of approaches may have in transdisciplinary applications. The theory behind the obtaining and handling of morphological data will be discussed, followed by a description of the current statistical tools available for processing this data. An introduction to the key concepts of AI will also be presented, including a debate on the advantages and disadvantages of these techniques.

The third chapter will be dedicated to the scientific articles published as a product of this Doctoral Thesis. This chapter is split into 5 sections;

**1. Obtaining Data and Robust Statistics**

In this section, both the extraction of data from 3D models and the processing of this data using robust statistics are explored.

In Courtenay et al. (2020a), the reproducibility and reliability of different landmark models is evaluated, in relation to the extraction of morphological information from 3D models in the field of taphonomy. For this purpose, the authors explore different techniques for the analysis of inter and intra-observer error, using robust statistics to refine the calculation of error, and thus assess the method's reproducibility. A development is then proposed for the use of computational techniques (sliding semilandmarks) for the collection of 3D coordinate data, so as to improve GMM resolution and results. This study uses structured light surface scanning to produce 3D models, and is focused on the analysis of carnivore chewing modifications produced on the surfaces of bones, namely their *tooth pits*.

Building on this, Courtenay et al. (2021a) use the landmark model presented in Courtenay et al. (2020a) for a case study in ecological research, both applicable to ecological contexts of the past and present. Developments are made comparing and evaluating the processing of 3D information, as well as 2D information derived from 3D models, while assessing the best statistical tools in order to evaluate this type of data. In addition, Courtenay et al. (2021a) debate the reliability of experimental samples obtained from captive and wild carnivores, discussing how animal captivity influences animal stress and the tooth marks they produce. In order to support their results, the authors employ robust statistical measures, and confront an important topic regarding the evaluation of statistical $p$-Values, so as to ensure the highest precision in all statistical inferences proposed. This study uses structured light surface scanning to produce 3D models, and focuses primarily on the intraspecies variability of different wolf populations based on their tooth pits and scores.

**2. Data Simulation and Augmentation**

This section is dedicated to the simulation and augmentation of datasets, so as to overcome limitations imposed by small sample sizes; a fundamental step that will consequently condition the quality of results obtained in the third section of scientific publications included in this Doctoral Thesis. In both Courtenay and González-Aguilera (2020) and Courtenay et al. (Under Review-a), the authors discuss the possibility of *Machine Teaching*, as a more powerful means of constructing classification algorithms, while also exploring the different robust statistical techniques available for the evaluation of simulated data.

The first study, by Courtenay and González-Aguilera (2020), designs and implements a new application of unsupervised Neural Networks, known as Generative Adversarial Networks (GANs), for the simulation and generation of synthetic, yet realistic, morphological variables. Here the authors discuss methods for the evaluation of the quality of synthetic numeric data by means of robust statistical approaches and equivalency testing. This study is mostly focused on the simulation of GMM data of any origin. Finally, Courtenay and González-Aguilera (2020) present a valuable critique and reflection on the use of bootstrapping techniques when training classification algorithms.

Following from this, Courtenay et al. (Under Review-a) present an alternative approach to that of Courtenay and González-Aguilera (2020), proposing the implementation of Monte Carlo (MC) based algorithms for the simulation of new data, including Markov Chain Monte Carlo algorithms (MCMC). The main contribution of this work is the additional inclusion of algorithms for the simulation of both categorical and numeric variables in the field of palaeoanthropology. This study additionally presents a novel methodological approach for the augmentation of 3D models, simulating 3D data by employing a combination of MCMCs and GMM tools. Data for this study was obtained using a mixture of micro-Computed Tomography ($\mu$-CT), and traditional measurements.

In order to support open science, Courtenay et al. (Under Review-a) is also associated with an R library; *AugmentationMC*.

### 3. Artificially Intelligent Classification Algorithms

This section is dedicated to the creation of classification models, using primarily Neural Networks (NNs), and Support Vector Machines (SVMs), for the processing of morphological data.

Courtenay et al. (2020b) present a NN architecture for the processing of GMM data describing the difference between cut and trampling marks; two types of structurally similar Bone Surface Modification (BSM) that are of great importance in the study of human evolution. This study uses digital microscopy to produce the 3D models from which data is obtained from. This study additionally serves as a basic introduction to NNs and how they work.

Courtenay et al. (2021b) present two additional powerful models for the classification of morphological data obtained using the methodological approach of Courtenay et al. (2020a). Here, 3D data for the description of carnivore tooth pits were obtained using Structured Light Surface Scanning. In this study, the authors show how (1) the augmentation of data improves classification results (Courtenay and González-Aguilera, 2020; Courtenay et al., Under Review-a), especially when using *Machine Teaching*, and (2) how a large set of carnivores can be differentiated when working with morphological data extracted from their tooth pits. Courtenay et al. (2021b) also present a direct comparison between GANs and MCMCs for the augmentation of data, building on the notions discussed in Courtenay and González-Aguilera (2020) and Courtenay et al. (Under Review-a). From this study, the authors were able to conclude that MCMCs are much less computationally expensive. This study also reveals how both are equally powerful in producing statistically equivalent data to the original training data, however MCMCs in general have been found to produce more realistic simulations.

To perform classification tasks, Courtenay et al. (2021b) describe a fairly new NN architecture, using NNs as a kernel function, before passing the extracted features into a final SVM output layer. This algorithm is called a Neural Support Vector Machine (NSVM). Here, the authors additionally use Bayesian Optimization algorithms to fine tune algorithm hyperparameters. Based on this work, the authors obtain above human-level accuracy when differentiating between tooth pits produced by different carnivores, including; brown bears, spotted hyenas, wolves, African wild dogs, foxes, jaguars, leopards and lions.

### 4. Transdisciplinary Applications in Science

This section is dedicated to real-world applications of the different methodological approaches detailed through Sections 1 through to 3.

Courtenay et al. (2023) present a real world application of the hybrid GMM & AI methodological approach for the classification of tooth marks in the 1.4 Ma Lower Pleistocene archaeological site of Barranco León (BL), Orce, Granada, Spain. This builds on the work of Courtenay et al. (2020a), Courtenay et al. (2021b), and Courtenay et al. (Under Review-a). Using structured light surface scanning as a means of obtaining 3D models, the authors find the sample of archaeological tooth pits to present morphological affinities with 5 different extinct carnivore species, with the interesting and novel conclusion that the Mosbach wolf (*Canis mosbachensis*) played an important role in the formation of the BL assemblage. Asides from the Mosbach wolf, the authors were also able to detect the activity of some saber-tooth (machairodontine) Felids, *Homotherium latidens*, the giant Hyena, *Pachycrocuta brevirostris*, as well as other species such as the lesser known canid *Lycaon lycaonoides*, and the Etruscan bear, *Ursus etruscus*. In this study, the authors debate the interpretation of the BL site, and highlight the usefulness of their methods in the detection of extinct carnivore species based solely on the chewing damage they produce on bone.

In Courtenay et al. (2022a), the authors describe a novel transdisciplinary approach to the study of morphology in the field of dermatological oncology. From this perspective, Courtenay et al. (2022a) propose the use of morphological descriptors to analyse skin lesions, more specifically the case of Non-Melanoma type skin cancers, as well as benign cutaneous lesions. This study analyses coloured (Red-Green-Blue, RGB) images of cutaneous Squamous Cell Carcinoma, Basal Cell Carcinoma, Intraepithetal Carcinoma, as well as Benign Lesions. Due to the difficulty describing the morphology of dermatological elements without the presence of clear homologous points that could be used as landmarks, the authors propose the use of a landmark-free approach, known as Elliptic Fourier Analyses (EFA). In order to extract information from these images, the authors first use a manual segmentation of the images, provided by a dermatologist, and then refine this segmentation using Computer Vision techniques. Once the outline of lesions had been defined, EFA coefficients extracted from this data were shown to be a powerful tool for the analysis of skin lesion morphology, revealing malignant tumours to be mostly described by border irregularities. Alongside measurements of lesion asymmetry, and combined with machine learning algorithms such as SVMs, the authors show how benign and malignant lesions can be separated when using EFA coefficients and asymmetry metrics. This is one of the first applications of such an approach in dermatology, and presents a novel insight into cutaneous lesion diagnostics.

### 5. A New Mathematical Model for Morphological Analyses

In the final section of scientific publications, Courtenay et al. (Under Review-b) present a new mathematical model for the analysis of landmark data, developing techniques from traditional GMMs, while integrating new elements from Graph Theory and Geometric Learning. From this perspective, the authors debate the quality of Dimensionality Reduction (DR) techniques typically used in GMMs, proposing the hypothesis that traditional DR calculations, such as Principal Component Analyses (PCA), may lose sight of the relationship landmarks have with the rest of the configuration, and thus produce a drop in quality of the final constructed morphological variable feature space. To overcome this issue, the authors propose the use of message passing and neighborhood aggregation formulae to first embed superimposed landmark coordinates, before further statistical applications were performed. As a proof of concept, the authors present an application on data describing the morphology and biomechanics of primate radii.

The proposal of Graph-based GMMs was found to be particularly useful for detecting biomechanical patterns among different great ape radii. While traditional GMMs are informative on the morphological traits related with primate locomotion and long bone biomechanics, a Graph-based approach was found to construct more stable feature spaces than traditional PCA approaches, regardless of the data used as input. From this perspective, Graph-based approaches were able to detect morphological patterns in areas of the bone that are usually obscured by less powerful DR tehcniques.

The fourth and final chapter briefly explains the final remarks obtained throughout the course of this research, alongside proposals for future research and perspectives.

In support of the main contents of this body of work, Appendices A through to D present a detailed theoretical explanation of the different elements presented in the scientific publications. These Appendices serve as support or supplementary material, that may prove useful in understanding the different topics touched on in each of the research papers. Appendix A, therefore, details the mathematical theory behind GMMs, as well as a note in Appendix A.5 on proposals for a more robust approach to landmark superimposition. Appendix B describes the formulae and basis behind Fourier transforms as descriptors for Morphology. Appendix C describes the theory behind $p$-Value evaluations, as a means of proposing a more robust approach to hypothesis testing in statistics. Lastly, Appendix D briefly describes the internal functionality of a NN, as well as an explanation on the best practices for evaluating Computational Learning performance.

Finally, Appendices E and F describe the technical side of this Doctoral Thesis, including a detailed description of all software developed for each of the studies (Appendix E), as well as the scientific impact that this body of work has obtained (Appendix F).

# Chapter 2

# Definitions and Motivation

Geometry is the branch of mathematics that confronts the properties and relations of points, lines, surfaces, solids, and other high dimensional analogues. From this perspective, geometry defines components of space with relation to the distance, shape, size and relative position of elements. Coming from the Greek terms *"geo"* (earth) and *"metron"* (measurement), geometry is currently one of the oldest fields in mathematics.

From geometry, many sub-fields have emerged with multiple theoretical as well as real-world applications. Throughout this Doctoral Thesis, we will be dealing with the concept of *morphology*, and the fields related to the study of morphology commonly known as *morphometrics*. From here, we define morphometrics as the measurement (*"metron"*) of shape (*"morphe"*), while morphology is concerned with the analysis of structure and form, most typically used in biological contexts.

Geometrical data can be collected from almost any source, from more traditional methods, using caliper or scribe-based measurements (Corner et al., 1992a; Mortenson and Steinbok, 2006; Mendonca et al., 2013; Robinson and Terhune, 2017, *inter alia*), to advanced digitisation techniques such as medical scans (Mendonca et al., 2013; Robinson and Terhune, 2017; Barbero-García et al., 2019; Bello and Galway-Witham, 2019, *inter alia*), high-resolution microscopy (Bello and Soligo, 2008; Boschin and Crezzini, 2012; Ball et al., 2017; Pante et al., 2017; Bello and Galway-Witham, 2019; Souron et al., 2019; Courtenay et al., 2019a, 2020b,c, *inter alia*), photographs taken using a camera (Pollefeys, 2004; Luhmann, 2010; Rodríguez-Martín et al., 2015; Ruiz de Oña et al., 2022, *inter alia*), or even a mobile phone (Salazar-Gamarra et al., 2016; Kottner et al., 2017; Barbero-García et al., 2017, 2018, 2019, 2020; Lerma et al., 2018, *inter alia*).

In real world applications, notions of geometry and morphology are fundamental in the way we perceive and describe objects. The shape of an element can be highly informative from multiple different perspectives, such as the detection or characterization of anomalies, patterns, and other more structural characteristics of an object or organism. While shapes are geometric elements, however, they are mostly studied qualitatively; to the human eye a shape can be a triangle, but we seldom describe the triangle according to the trigonometric properties that lie underneath and define it.

In many cases, decisions about objects are often made using their shape and size (Dryden and Mardia, 1998). Here we describe just a few;

*Medicine*

In dermatology, for example, the shape, size and symmetry of a skin lesion are often used to diagnose possible cases of skin cancer (Friedman et al., 1985; MacKie, 1986; Tsao et al., 2015). While the use of specific tools, such as dermatoscopes, facilitate the visual inspection of these lesions, observer subjectivity and inter-observer variability are notable factors in diagnostic accuracy (Kittler et al., 2002; Piccolo et al., 2002; Friedman et al., 2008; Vestergarrd et al., 2008; Merlino et al., 2016). Similarly, both physical and mental pathologies or disabilities can also be described and assessed using geometric variables (e.g. Astley and Clarren, 1995; Drew and Sachs, 1997; Siegenthaler, 2015). In this latter example, however, methods for the quantification and detection of these biological anomalies have been explored in greater detail (Mutsvangwa et al., 2010; Barbero-García et al., 2017, 2018, 2019, 2020; Roussos et al., 2021).

*Photogrammetry and Computer Vision*

Many algorithms and techniques for photogrammetry and Computer Vision (CV) also exploit the use of geometry and morphology. Photogrammetry, whose main focus is the precise and accurate extraction of geometric and morphological elements, utilises a multitude of algorithms for the emphasis and detection of features from images (e.g. Ramer, 1972; Douglas and Peucker, 1973; Canny, 1986; Harris and Stephens, 1988; Lowe, 1999). Based on these methods, photogrammetric applications can be used to facilitate more complex applications such as object recognition and 3D modelling (Hartley and Zisserman, 2004; González-Aguilera, 2005; Forbsyth and Ponce, 2011; Demaagd et al., 2012; Föstner and Wrobel, 2016; Ruiz de Oña et al., 2022). CV, on the other hand, is focused on the fast and efficient extraction of geometric information from image data, presenting a wide array of different applications such as facial recognition (Bronstein et al., 2005), and character recognition (Fritzche, 1961; Raudseps, 1965; Ferson et al., 1985). In many of these cases, these types of applications can be useful for the automation of processes that could otherwise be costly, complex, or tiring. Moreover, 3D modelling applications have a wide array of different uses and purposes that are useful in almost any field of science.

*Material Sciences*

In the industrial sector, techniques for the detection of deformations and faults are also fundamental. While the majority of this work has been based on the visual inspection of elements, the use of advanced computer vision techniques such as digital image correlation is proving popular to improve anomaly detection (Herbert, 2010; Luhmann, 2010; Chen et al., 2011; Hua et al., 2012; Rodríguez-Martín et al., 2015; Zaczek-Peplinska et al., 2015). The basic principal in this line of research is to rapidly identify defects in different materials, which in turn ensures the safety of structures and machinery. To provide an example, the geometric properties of welds are highly informative in the structural integrity of industrial elements. In this sense, the analysis of plane misalignment (Rodríguez-Martín et al., 2016a), thickness and excess weld materials (Ruiz de Oña et al., 2022), weld bead geometry and surface imperfections (Gong et al., 2018; Rodríguez-Gonzálvez and Rodríguez-Martín, 2018), as well as problematic fissures and cracks (Henkel et al., 2016; Rodríguez-Martín et al., 2016b, 2020), are highly influenced by morphological and geometric variables. From this perspective, morphological descriptors may be of use in the fast and empirical identification of possible anomalies.

*Archaeology*

Morphological trends and patterns are also highly informative in the study of archaeological materials. For example, the emergence of symmetric stone tools in the Lower Palaeolithic, known as handaxes, is of great interest for the development of early *Homo* (Lepre et al., 2011; Diez-Martín et al., 2019; Sano et al., 2020, *inter alia*). Morphological studies of handaxes are thus useful to characterise and understand degrees of symmetry, as well as the general emergence of the shape of these tools (Costa, 2010; Serwatka, 2015; Iovita et al., 2017; García-Medrano et al., 2019), which can essentailly have both functional (Key et al., 2016; Key and Lycett, 2017a,b,c), and cognitive (Stout et al., 2000, 2006, 2008, 2015; Stout and Chaminade, 2012; Geribàs et al., 2010; Lombao et al., 2017; Muller et al., 2017), implications in early human evolution. Similar notions, however, can be equally important in the characterisation of other archaeological materials, such as other types of stone tools, rock and mobile art, ceramics, metal objects such as swords, broaches, belt buckles, and many others.

*Taphonomy*

Taphonomists dedicated to the identification of microscopic marks on archaeological and palaeontological bones, known as Bone Surface Modifications (BSMs), also rely heavily on the *in-visu* characterisation of different traces. In most early research, the marks left by stone tools when cutting meat (*cut marks*), those left by carnivores when feeding and chewing (*tooth marks*), as well as naturally produced marks typically produced by sedimentary abrasion (*trampling marks*), were originally described using a combination of qualitative features, as well as general description of their cross-section profile morphology. Throughout literature, references can be found to "V" shaped cut marks (Olsen and Shipman, 1988; Shipman, 1988; Domínguez-Rodrigo et al., 2009, *inter alia*), "U" shaped tooth marks (Binford, 1981; Haynes, 1983; Blumenschine, 1995, *inter alia*), and "_/" shaped trampling marks (Brain, 1967; Behrensmeyer et al., 1986; Andrews and Cook, 1985, *inter alia*). Nevertheless, the majority of these types of identification are heavily influenced by analyst experience and overall subjectivity (Domínguez-Rodrigo et al., 2017; Courtenay, 2019). In light of this, more detailed research into quantitatively describing and quantifying these morphological differences are relatively new, and have proven to be highly useful in the identification of problematic traces (Bello and Soligo, 2008; Boschin and Crezzini, 2012; Aramendi et al., 2017; Pante et al., 2017; Courtenay et al., 2017, 2019a, 2020c; Yravedra et al., 2017; Courtenay, 2019; Souron et al., 2019, *inter alia*).

*Systematic Biology and Palaeoanthropology*

Finally, for systematic biologists and palaeoanthropologists, morphology can be used as a diagnostic feature to identify species (Henning, 1966; Hawks, 2004), or describe evolutionary changes. The first descriptions of the hominin species *Zinjanthropus boisei* (Leakey, 1959), for example, originally described the OH5 skull (Olduvai Gorge, Tanzania) using 20 features, of which the terms “size”, “shape” and “form”, as well as other descriptors related to morphology, are used in all of the features. From this perspective, criteria such as the “less elongated” foramen magnum, or the “more steeply” rise of the occipital posteriar wall, were used to establish diagnostic criteria that could identify this species in other skulls (Leakey, 1959; Tobias, 1967). Beyond these qualitative variables, however, the definition of *Zinjanthropus boisei* was challenged, considering many of the “features” to not be entirely unique to the OH5 specimen, with the only difference with other genus, such as *Paranthropus*, being a change in size (Robinson, 1960). Since the incorporation of much more sophisticated techniques has become standard practice, however, our knowledge and ability to describe this species, *Paranthropus boisei*, has greatly improved (O’Higgins et al., 2011; Benazzi et al., 2011; Domínguez-Rodrigo et al., 2013; Daver et al., 2018; Lague et al., 2019; Richmond et al., 2020; Aramendi, 2021, *inter alia*).

While this is only a brief overview of the multiple number of applications morphological research can have, it can be seen that a more detailed look into quantifying morphology can be fundamental, as opposed to simple *in-visu* descriptions. These types of approaches can be considered optimal to obtain a more objective and empirically accurate look into the data at hand.

## 2.1 Notions of *Shape* and *Form*

Two words that are commonly used in the study of morphology are "shape" and "form"; two terms that need clarification and definition before more complex notions of geometry can be described. In the visual arts, and many fields that have been developed since, the term *shape* refers to a flat representation of an element, which can usually be deconstructed into different *shapes*, such as simple collections of lines, triangles, circles, or other polygonal geometries. *Form*, on the other hand, is often defined as the three-dimensional composition of these elements (Rowland, 1966). In topology, Borsuk (1975) provided a complex definition of shape that describes a more global view of topological spaces, and the maps between them. Nevertheless, here we will be using the definitions of Kendall (1977, 1984), and Goodall (1991), that are more specific to the empirical and mathematical processing of "real" shape from a morphological perspective, as opposed to a topological one;

**Definition 1** *Shape* What is left when the differences which can be attributed to translations, rotations, and dilations have been quotiented out.

**Definition 2** *Form* What is left when the differences which can be attributed to translations and rotations have been quotiented out.

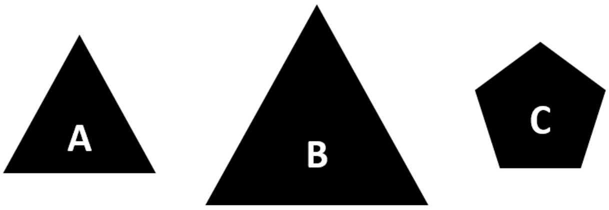
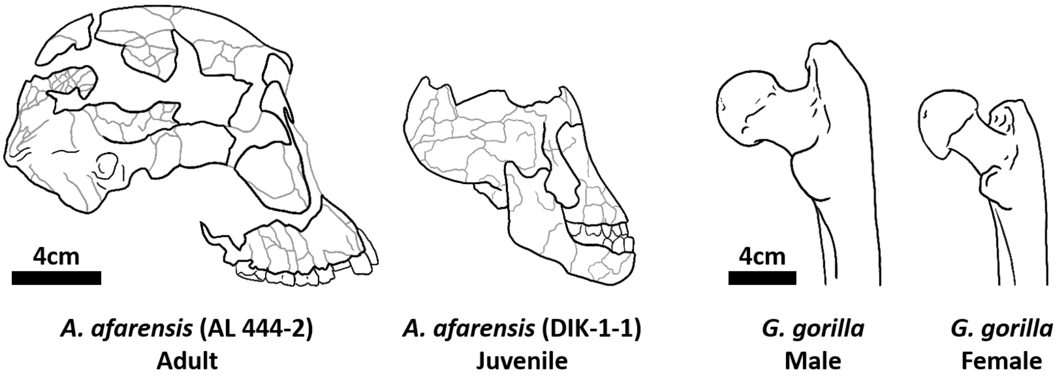


***Figure 2.1:*** *Figure presenting differences in shape and form. Triangles A and B are of the same shape yet have different form, while pentagon C is of a different shape and form.*

Shape is thus considered as a representation of the pure morphology of an element, excluding its size, position, and orientation. Form, on the other hand, includes size (Fig. 2.1). Appendix A.1 provides a more specific mathematical definition of these concepts in equations A.1 and A.2.

The importance of separating between shape and form is fundamental, especially in biological contexts. In general, shape is more informative on the internal constitution or structure of an organism (Jolicoeur and Mosimann, 1981), as the variable size is usually conditioned by many other variables that may not be of interest to the researcher. For example, the size of an animal is often strongly conditioned by external factors, such as ecological context and temperature (Bergman, 1847; Allen, 1877; Peters, 1983). Similarly, size can also influenced by the behavioural and physiological attributes of animals and organisms (McMahon and Bonner, 1983; Schmidt-Nielsen, 1984; Jungers, 1985; Jungers et al., 1995). From a different perspective, an animal's size is influenced to different degrees by other variables, such as sexual dimorphism (Sundberg, 1989; Formicola and Franceschi, 1996; Franklin et al., 2006, 2007; Plavacan, 2012; del Bove et al., 2020), and development with age (Sundberg, 1989; Enlow and Hans, 1996; Rice, 1997; Lieberman et al., 2002; Cobb and O'Higgins, 2004; Mitteroecker et al., 2004a; Bastir and Rosas, 2004; Bastir et al., 2006; Aramendi, 2015; Freidline et al., 2015). Figure 2.2 provides an example of these conditioning variables.



***A. afarensis*** **(AL 444-2) Adult**

***A. afarensis*** **(DIK-1-1) Juvenile**

***G. gorilla*** **Male**

***G. gorilla*** **Female**

***Figure 2.2:*** *Examples of cases where the variable 'size' can have great influence on morphology, including differences in age, and sexual dimorphism. Examples of differences in age include two skulls of* Australopithecus afarensis *individuals, including the male adult skull (AL 444-2) recovered from Hadar (Ethiopia), in comparison with the three-year-old "Salem" (DIK-1-1) individual, from Dikika (Afar, Ethiopia). Examples of sexual dimorphism compare the differences in femoral proximal epiphyses of a male and female Gorilla.*

Each of the aforementioned points will condition the weight size has on morphology. This relationship is typically studied and described in terms of two main types of shape-size relationships (Jungers et al., 1995);

**Definition 3** *Isometry* Shape is preserved among entities of different sizes.

**Definition 4** *Allometry* Shape changes as a function of size.

In light of these relationships, researchers are conditioned to initially assess the weight size may have on morphological patterns, and in the case of detecting allometry, perform their research without scaling their data, or at least utilise size to understand and assess how these variables influence their study.

The issue behind ignoring this relationship is especially evident from a methodological and statistical perspective. As will be discussed in the following section, many techniques for the study of shape and form are conditioned by the high correlation linear distances evidently have with an individual's size (Bookstein et al., 1985). Likewise, the weight this variable has on many types of multivariate analyses, such as Principal Components Analyses (PCA), overwhelms other variables that may be equally, or more, important (Jolicoeur, 1963; Sundberg, 1989; Jungers et al., 1995).

To demonstrate this point, if we were to take the extreme example of comparing the humeri of two male Tragelaphine African bovids (Fig. 2.3), *Tragelaphus imberbis* (≈ 92-108 kg) and *Taurotragus derbianus* (≈ 900 kg), despite them both being male individuals of the same taxonomic tribe, little could be said about morphological differences and similarities between the two beyond the extreme differences in size (Jungers et al., 1995). If we were to exclude the variable size from this analysis, however, we would likely begin to detect more interesting patterns in African bovid biomechanics.

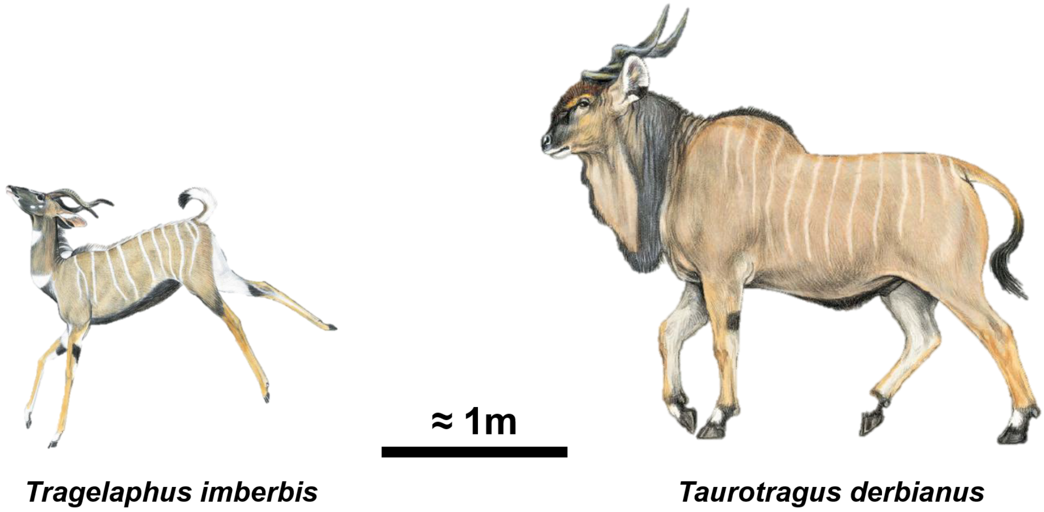


***Figure 2.3:*** *A comparison in average size of Tragelaphus imberbis and Taurotragus derbianus. Drawings are from Kingdon (2015)*

## 2.2 Methods in the Study of Morphological Data

For many years, morphological studies were performed using the measurement of linear distances, curvilinear distances, ratios, and angles (Atchley et al., 1976; Atchley and Anderson, 1978; Dodson, 1978; Humphries et al., 1981; Howells, 1989; Jungers et al., 1995, *inter alia*). Through combinations of these variables, analysts could then perform multivariate statistical analyses in order to assess and understand possible underlying patterns (Slice, 2005). In many cases, these methods have been refered to as *Traditional Morphometrics* (Marcus, 1990). Nevertheless, while linear measurements can be informative for many applications, they also have some limitations. As pointed out by Bookstein et al. (1985), these types of measurements are highly sensitive to variations in size. While many methods of size correction have been proposed, most provide different results with little agreement. Similarly, Humphries et al. (1981) stated that size cannot be estimated by using a single linear measurement; the same measurement can be obtained from different points, without necessarily sharing geometric or structural meaning. Finally, an oval and a teardrop might have different shapes, but could very easily have the same measurement of maximal length and width, which would provide misleading information about their true morphology (Adams et al., 2004).

Towards the end of the XXth century we begin to find the use of Cartesian coordinate systems for the extraction of morphological variables, leading to some of the largest advances in this field (Bookstein, 1978, 1986a; Bookstein et al., 1985). The use of Cartesian coordinates in this sense is generally dominated by the term *landmark*-based analyses, where landmarks are;

**Definition 5** *Landmark* A homologous point of correspondence on each object, of anatomical or geometric significance, that matches between and within populations.

The use of landmarks overcomes many of the aforementioned issues, as seen through the use of the terms “homologous”, and “anatomical or geometric significance” (see Appendix A.1). Using this definition, an individual can only be studied if presenting all of the referential loci, while loci have to be defined to a sufficient degree so that they can be easily located on each individual (Corner et al., 1992b; Yezerniac et al., 1992; Robinson and Terhune, 2017). While it is true that further research has found different means to overcome the issue of missing landmarks in some specimens (Gunz, 2005; Gunz et al., 2005a, 2009), these methods can be sensitive to the data used as input and are not always applicable to certain specimens. In this sense, it is generally stressed that all landmarks must be present in order to study an individual. In light of these specifications, this type of landmark can also be referred to as a *fixed* landmark, and can be recorded in both 2D or 3D.

Three main types of fixed landmarks exist (Dryden and Mardia, 1998), with their categories being defined as follows;

**Definition 6** *Anatomical Landmark* Landmarks of biological and anatomical meaning.

**Definition 7** *Mathematical Landmark* Landmarks that are located based on some mathematical or geometrical property.

**Definition 8** *Pseudo-Landmark* Landmarks that are constructed based on, or around, already present landmarks or local elements of the object under study.

As can be seen, these definitions rely heavily on biological features, and are thus quite restrictive to those interested in studying the shape and form of different types of elements. From this perspective, we can develop Definition 6 to be referred to as *Scientific* landmarks, where domain-based knowledge of general points of interest can also be used to define the landmark.

Another more specific definition of landmarks provided by Bookstein (1991) considers;

**Definition 9** *Type I* Landmarks that occur at the intersection or joints between tissue and bone, and that are easily recognisable to the trained eye.

**Definition 10** *Type II* Landmarks that are defined by local properties, such as points of maximal curvature, that usually present some biological significance.

**Definition 11** *Type III* Landmarks that are constructed across the element under study, and that are of no biological significance, such as centroids, or points between Type I and II landmarks.

When comparing both of these sets of definitions, Type I landmarks can be considered to be both anatomical and mathematical landmarks, while Type II landmarks can essentially be considered anatomical, mathematical and pseudo-landmarks. Type III landmarks are exclusively pseudo-landmarks. Each of these definitions show how different landmarks have their advantages and disadvantages, with the degree of subjectivity and human-induced error likely to increase for Type III as opposed to Type I landmarks (see analysis and references therein of Courtenay et al. (2020a)). Type III landmarks, however, are common, and can be useful for the definition of of interest on more complex surfaces where biological meaning is either scarce or completely absent.

Once a set of Cartesian coordinates has been defined, a number of different approaches exist to convert this data into more meaningful variables of morphological significance. The first considers the development of the aforementioned traditional morphometric studies by using landmark coordinates to extract measurements and ratios. This type of analysis is generally known as Euclidean Distance Matrix Analyses (EDMA), and uses the calculation of inter-landmark distances in order to characterise the morphology of the element under study (Lele, 1991; Richtsmeier et al., 1992, *inter alia*). Another approach performs an initial triangulation of landmark coordinates, describing morphology as a function of each triangle's interior angles (Rao and Suryawanshi, 1996, 1998). Nevertheless, the most innovative and powerful advances in landmark-based analyses are from the field of Geometric Morphometrics (GMMs).

**Definition 12** *Geometric Morphometrics* The suite of methods for the acquisition, processing, analysis, and visualisation, of morphological variables, that retain *all* of the geometric information contained within the data.

The objectives of GMMs are to analyse landmark coordinates directly, as opposed to the distances between them (Slice, 2005). From this perspective, GMMs focuses on the analysis and characterisation of landmark displacements consequent to a common cause, while trying to determine what this cause may be (Bookstein, 1991). Raw coordinates thus provide perspective into the position, size and orientation of the configuration, and present a more efficient means of analysing both *shape* and *form* (Fig. 2.4).

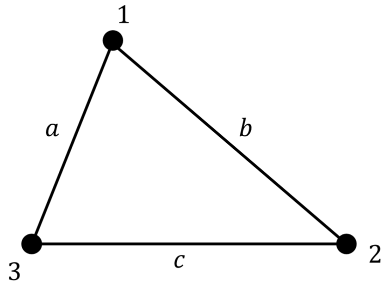


***Figure 2.4:*** *Diagram explaining the theory behind shape and form variation using landmark theory, where landmarks 1-3 are sufficient in describing the pure morphological shape of an individual, while distances a, b and c are more indicative of form from a traditional morphometric perspective.*

One way to analyse the total variation of the landmark configurations is by superimposing all landmarks onto a common reference system. While multiple techniques exist for this superimposition procedure (see Appendix A and references therein), we generally refer to this process as Generalised Procrustes Analysis (GPA) (Appendix A.2).

The advantages of superimposing coordinates are multiple, as this technique allows for the direct comparison of landmark configurations, thus quantifying minute displacements of individual landmarks in space (Bookstein, 1991; Richtsmeier et al., 2002). Moreover, the superimposition procedure is also affected by the relationship landmarks have with each other in the context of the configuration, a feature that can be exploited in greater detail, as will be discussed in Courtenay et al. (Under Review-b). This phenomenon, known as the *Pinocchio effect* (Fig. 2.5), can be described by how the displacement of only a single landmark can cause all landmarks to shift in the final superimposed configuration (Chapman, 1990; Walker, 2000; Hallgrimsson et al., 2015; Klingenberg, 2021; Courtenay et al., Under Review-b, Sup. File 2).

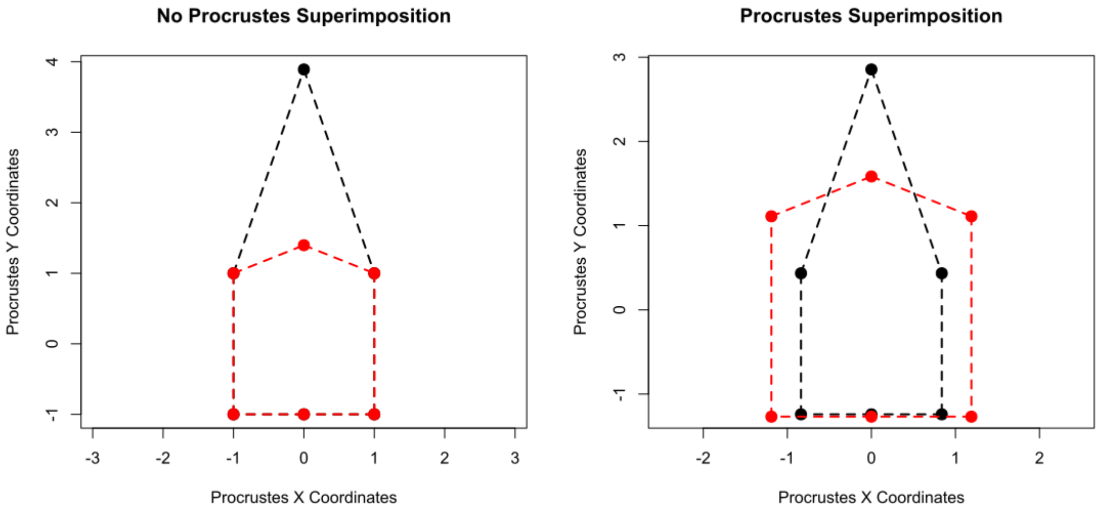


***Figure 2.5:*** *A demonstration of the* Pinocchio effect *from Courtenay et al. (Under Review-b), Supplementary File 2. Left panel: the original two icons, where only a single landmark presents a displacement along the y-axis. Right panel: the two icons after procrustes superimposition in shape space.*

GMM analyses are also highly useful for the visualisation of shape and form changes, a concept that is lost in most traditional morphometric analyses. The Thin Plate Spline (TPS), for example, is a mathematical model proposed by Bookstein (1989), and developed from Thompson (1917), that employs a two-dimensional grid to visualise and express both local and global deformations when comparing two landmark configurations (Fig. 2.6, A.9, and Appendix A.4). TPS transformation grids are thus a valuable tool for the visualisation of shape and form changes across any feature space, providing more comprehensible meaning to the morphological variables obtained. With the additional advantage of using landmarks, as opposed to simple measurements, we can also localise shape changes to specific anatomically or scientifically defined points of interest, producing results that are easier to interpret and understand.

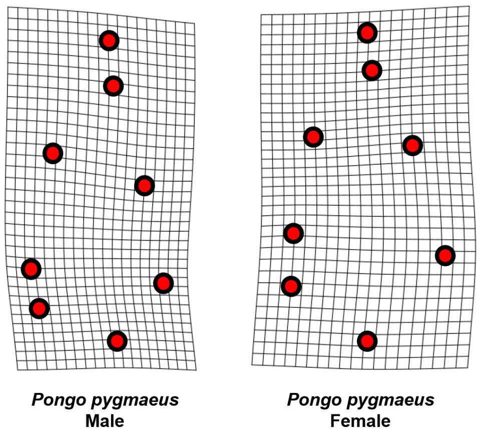


***Figure 2.6:*** *Two Thin Plate Spline warpgrids visualising shape change between male and female orangutan (Pongo pygmaeus) individuals. Data from O'Higgins and Dryden (1993).*

One of the early limitations, however, of landmark-based approaches, was the ability to objectively quantify curves and surfaces where morphological information is valuable, but no homologous points of interest can be defined (Bookstein, 1997). For this purpose, the proposal of a new type of landmark, known as a sliding *semilandmark*, has proven a valuable means of reducing the excessive use of Type III landmarks, and is a more precise means of digitising curves and surfaces (Bookstein, 1991, 1997; Gunz et al., 2005b; Gunz and Mitteroecker, 2013).

**Definition 13** *Sliding Semilandmark* A computational landmark placed on a curve or surface which is allowed to slide with respect to neighbouring landmarks so as to establish geometric homology.

Originally proposed for 2D samples (Bookstein, 1991, 1997), and later developed in 3D (Gunz et al., 2005b), semilandmarks are a means of modeling from homologous curves and surfaces, that can then be combined with information from fixed landmarks in a final complete landmark model (Fig. 2.7). Methodologically, the process of sliding semilandmarks consists in designing a template that can be projected onto a specimen, where the semilandmarks are then *slid* across a plane tangent to the surface so as to minimise differences between the reference and target elements (i.e. find homology). This minimisation can be performed either by reducing bending energy (Bookstein, 1991; Gunz et al., 2005b), or by reducing Procrustes distances during superimposition procedures (Rohlf and Slice, 1990). The former, however, is generally the most accepted and produces the best results (Gunz and Mitteroecker, 2013). Moreover, sliding is an iterative process (Bookstein, 1997; Gunz et al., 2005b; Gunz and Mitteroecker, 2013), where semilandmarks slide simultaneously and are constrained by neighbouring fixed landmarks that define the starting and finishing points of the sliding process.

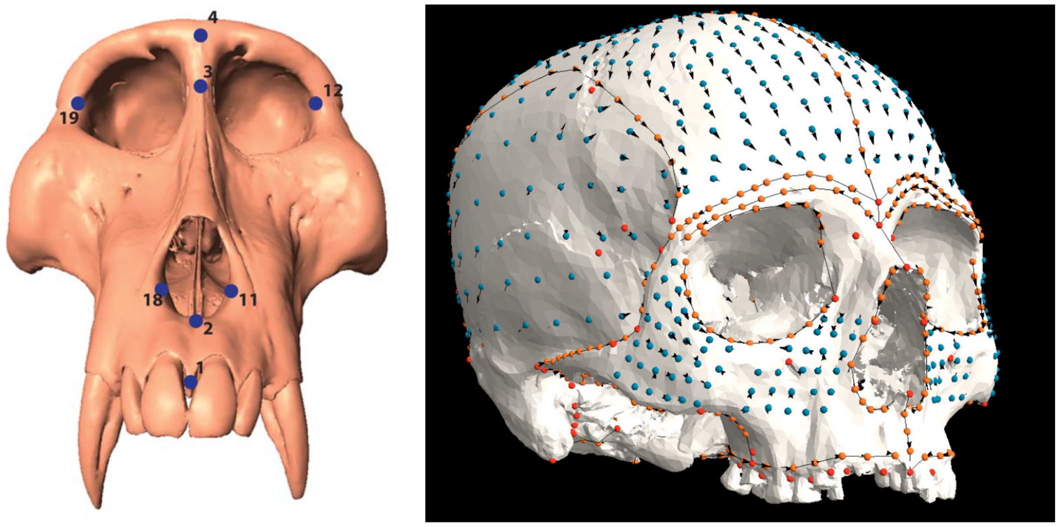


***Figure 2.7:*** *Examples of both landmarks and semilandmarks from case studies published by Robinson and Terhune (2017) (left) and Mitteroecker and Bookstein (2011) (right). Left Panel: examples of fixed landmarks, including Type I (LM3: Nasion), Type II (LM18: Right Alare) and Type III (LM4: Glabella) landmarks. Right Panel: Examples of sliding semilandmarks (blue and orange dots).*

It is important to point out, however, that the digitisation of a curve or surface using semilandmarks does not gaurantee, nor imply, equidistance. While equidistance may seem a logical means of computing the position of semilandmarks, this does not necessarily lead to geometric or biological significance, thus may not imply homology (Bookstein, 1997; Gunz et al., 2005b; Gunz and Mitteroecker, 2013). Semilandmarks are therefore projected in such a manner that they optimize their position with reference to both the surface, and the average shape of the sample (usually the average of Procrustes shape coordinates), thus establishing geometric correspondence and homology (Gunz and Mitteroecker, 2013).

Nevertheless, regardless of the digitisation technique, GMMs are not applicable to all types of elements. If homology between points cannot be established, then a GMM approach will not work.

A different technique for the extraction of morphological variables from (namely) 2D data, where no homologous loci of biological or geometric significance can be identified, consists in the fitting of a mathematical function, or curve, to the empirical curve or outline being studied (Rohlf, 1990). These techniques can often be found referred to as *outline analyses*. Many techniques exist to perform this type of analysis, however the most common and popular is known as Fourier Analyses (FA). FA describes shape (and sometimes form) as a series of periodic functions along the curve or outline of the object being studied (see Appendix B).

FA exists in three main variants; Fourier Radius Variation (FRV, Appendix B.3), Fourier Tangent Angle (FTA, Appendix B.4), and Elliptic Fourier Analyses (EFA, Appendix B.5). Each have their advantages and disadvantages, while some are more conditioned by the complexity of the outline being analysed. FRV and FTA, for example, are better suited for simple outlines, where radii of the outline do not intersect. EFA is more adaptable to complex outlines that present convexities and concavities. While each of these functions are technically suited for the study of closed outlines, they are still able to perform well on open outlines if the objectives of the study are to produce discriminant functions. More complex evolutionary studies, however, should use other types of functions (see Rohlf, 1990).

The use of fourier descriptors for the analysis of shape thus utilises Fourier series to deconstruct each outline into a series of *harmonics*, the sum of which can be used to reconstruct the entire shape (Fig. 2.8). Each harmonic can then be described by a series of coefficients which serve as morphological variables for further statistical analyses.

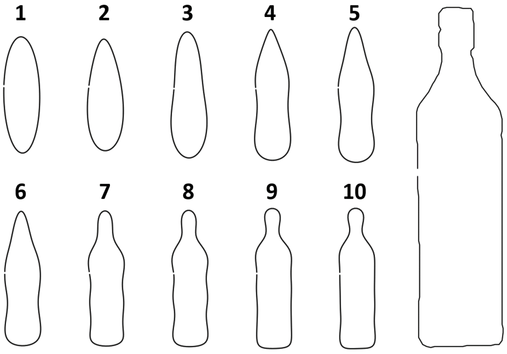


***Figure 2.8:*** *Diagram presenting the reconstruction of a bottle's outline based on the cumulative contributions of the first 10 harmonics.*

## 2.3 Statistical Approaches to Processing Morphological Data

Similar to the way the collection of morphological data has evolved in the last century, so has the way this data is processed.

Morphological data is typically analysed in a number of ways and from a number of different perspectives. The most common approaches part from a Frequentist perspective (Blackith and Reyment, 1971; Carroll and Green, 1997), however Bayesian statistics can also be found in morphometric analyses (e.g. Otárola-Castillo et al., 2017). Within the branch of Frequentist statistics, data is then often described univariately, either through descriptive statistics or hypothesis testing, or multivariately, using multivariate hypothesis tests.

A brief overview of statistical analyses in morphometric studies highlights the use of a specific set of very popular univariate tests for the comparison of single variables that describe samples. These mostly include the $t$-test, $F$-test, and Analyses of Variance (ANOVA) (Pearson, 1895; Student, 1908; Fisher, 1925; Chambers et al., 1992). The latter is frequently found in the form of Type II ANOVA, and typically accompanied with a Tukey's pairwise post-hoc analysis (Tukey, 1949). When comparing data multivariately, ANOVA is usually developed into a Multivariate Analysis of Variance (MANOVA) (Lawley, 1938; Hotelling, 1951). When comparing the relationships between sets of variables, correlations are often calculated using Pearson's $r$, while linear regressions are also frequently used (Pearson, 1895).

As for ordination, one of the most popular techniques for analysing multiple variables simultaneously is through the use of general scatter plots, or the eigendecomposition of datasets using Principal Component Analyses (PCA) (Jollife, 2002; Bishop, 2006).

In GMM analyses, a number of other tools are also available. For example, common practice in GMMs is the use of Procrustes distance calculations to measure similarities and morphological affinities between entire landmark configurations (Gower, 1975; Rohlf and Slice, 1990; Goodall, 1991; Kent, 1994; Rohlf, 1996; Dryden and Mardia, 1998; Rohlf, 2000; Slice, 2001). Using this measure, tests such as Goodall (1991)'s $F$ test have been shown to be particularly powerful and stable for the identification of similarities and differences between individuals (Rohlf, 2000). Developing from Goodall's $F$ test, GMMs present a series of tools that are particularly useful for assessing shape-size relationships as well (Mitteroecker et al., 2004b; Drake and Klingenberg, 2008; Adams and Nistri, 2010). Using multiple linear regressions, the $F$ test can be used similar to ANOVA to assess the significance of shape-size relationships. From this perspective, the degree of influence a measurement of size has on a set of shape variables can be used to assess how changes in size affect changes in shape (Adams and Nistri, 2010). For GMMs, the centroid size (and its logarithm) is typically used for this test.

In terms of ordination, Cannonical Variance Analysis (CVA) and between-group PCA (bgPCA) are also popular ordination tools to understand inter and intra-group variance (Campbell and Atchley, 1981; Albrecht, 1992; Boulesteix, 2004; Mitteroecker and Bookstein, 2011; Rohlf, 2021). CVA is based on the decomposition and visualisation of sample covariance matrices, thus visually describing the statistical separation among groups. bgPCA, on the other hand, is mathematically more similar to PCA, where data is projected onto the principal components of group means (Barker and Rayens, 2003; Boulesteix, 2004; Cardini et al., 2019; Rohlf, 2021). Both CVA and bgPCA are tools applicable to any type of data, nevertheless, in morphology it is seen most frequently when applied to GMM data. Needless to say, one of the most common approaches to processing GMM data is first to decompose form and shape variables into lower dimensions via PCA or Relative Warp Analysis (RWA) (Bookstein, 1989, 1991; Rohlf, 1993, 1996, 1999).

PCA reduces coordinate data into a set of more manageable correlated variables that can be represented by much fewer dimensions, while accounting for as much of the original distribution's variation as possible (Jollife, 2002; Bishop, 2006). This is computed via the eigendecomposition of data, mostly using Single Value Decomposition (SVD). RWA, on the other hand, calculates the eigendecomposition of bending energy matrices derived from Thin Plate Splines (Bookstein, 1989; Rohlf, 1993, 1996, and Appendix A.4), performing a similar task to PCA, however deriving morphological variables from a different source. Moreover, RWA has an additional $\alpha$ parameter that can be used to stress local or global morphological changes (Bookstein, 1991; Rohlf, 1993, 1996, 1999; Rohlf et al., 1996; Walker, 1996). While these are both popular techniques for pattern recognition and dimensionality reduction in GMMs, PCA is generally more widely used. Once either PCA or RWA has been performed, the scores (i.e. dimensions) presenting the highest morphological variance are then separated and subjected to a series of univariate and multivariate statistical tests, as described above.

While each of these tests are perfectly valid and useful in a number of applications, they are all parametric, and are thus often unlikely to capture the true nature of variances in sample distributions. We define parametric and nonparametric tests as;

**Definition 14** *Parametric Test* A statistical test that is conditioned by assumptions about the underlying properties of a distribution.

**Definition 15** *Nonparametric Test* A statistical test that is either not conditioned by assumptions about the underlying properties of a distribution, or if these assumptions are made, the precise properties of the distribution are unknown.

In these definitions, the "assumption about the underlying properties of a distribution" is commonly that the data is normally or homogeneously distributed, i.e. the data fits a Gaussian distribution. Gaussian distributions, characterised by their "bell"-shaped Probability Density Function (PDF), can be defined by a series of very specific properties. Firstly, measures of central tendency of the distribution, such as the mean ($\mu$) and median ($\tilde{x}$), are approximately equal ($\mu \approx \tilde{x}$). Second, the PDF is symmetric, such that the area under the PDF marked by the first standard deviation ($\sigma$) represents $\approx 68.27\%$ of the information, the second standard deviation accounts for $\approx 95.45\%$, and the third standard deviation accounts for $\approx 99.73\%$ (Fig. 2.9). Finally, both the skewness and kurtosis parameters of the distribution are equal to 0.

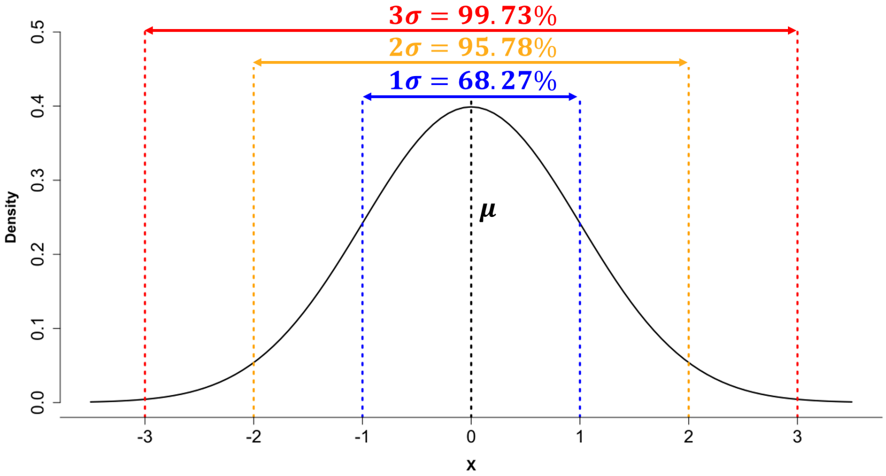


***Figure 2.9:*** *Visualisation of a simple Gaussian distribution (*$\mu = 0,\ \sigma = 1$*), marking both the location of the mean (*$\mu$*) and the ammount of information captured by each degree of the standard deviation (*$\sigma$*)*

While other types of statistical distributions exist, and are equally as common with their own particularities, in most cases these distributions eventually approximate a Gaussian distribution. For example, the popular $t$-distribution is conditioned by the degree of freedom parameter, while higher degrees of freedom cause the $t$-distribution's PDF to approximate a Gaussian distribution.

While it is true that many sources of morphological data tend to display some degree of homogeneity, especially data that has been pre-processed using PCA (Diaconsis and Freedman, 1984), this assumption cannot be considered a global law of morphology. In many cases, especially when sample sizes are small in relation to the general population, data is less likely to appear homogeneous. From this perspective, nonparametric statistical tests are typically more reliable when sample sizes are small. Likewise, in non-biological morphological studies, homogeneity is less common (Courtenay et al., 2020a, Under Review-b). In palaeoanthropology, for example, the additional likelihood that a dataset is noise-free is very low. While nonparametric tests can also be found in morphometric literature, such as the Mann-Whitney $U$ test (Mann and Whitney, 1947), they are generally used to a lesser extent.

When deviations from normality occur, a number of the aforementioned characteristics of the Gaussian PDF are no longer representative of the empirical distribution under study (Fig. 2.10). For example, the mean and median of a skewed distribution are not the same ($\mu \neq \tilde{x}$), while $\sigma$ no longer reflects the same amount of information on either side of the mean. Statistical tests, such as ANOVA that rely on sample means (Pearson, 1895; Student, 1908; Fisher, 1925; Chambers et al., 1992), are thus unreliable under these conditions. In light of this, if our data is skewed, noisy, or presents a PDF with a complex and unusual shape, how can this data be processed?

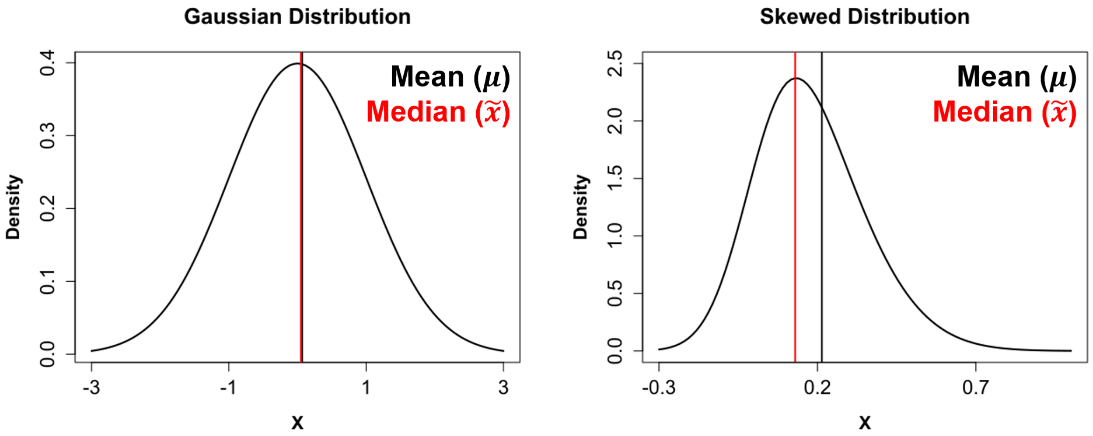


***Figure 2.10:*** *Diagram demonstrating the difference between mean and median as a measure of central tendency according to the homogeneity of the distribution.*

In these cases, the increasingly popular field of robust statistics, that includes the use of nonparametric testing, is the most accurate solution;

**Definition 16** *Robust Statistics* The branch of statistics that is designed to be resistant to deviations from assumptions about the underlying distribution of data.

From a robust statistical perspective, descriptive statistics such as the mean and standard deviation are replaced by other metrics less sensitive to the presence of outliers. For central tendency this is usually the median. For deviation, the Median Absolute Deviation (MAD), and its normalised version (NMAD), are frequently used (Höhle and Höhle, 2009; Hasan et al., 2011; Rodríguez-Gonzálvez et al., 2014; Ariza-López et al., 2019; Herrero-Huerta et al., 2018; Rodríguez-Martín et al., 2019a; Courtenay et al., 2020a, 2021c). As will be discussed in Courtenay et al. (2020a), the Square Root of the Biweight Midvariance ($\sqrt{BWMV}$) is also a reliable measure of sample dispersal (see also Nocerino et al., 2017; Rodríguez-Martín et al., 2019b; Courtenay and González-Aguilera, 2020; Courtenay et al., 2020a, 2021c). When calculating confidence intervals, robust statistical approaches use the calculation of quantiles (Höhle and Höhle, 2009), thus providing a non-symmetric means of describing the upper and lower confidence interval according to a set probability range. For hypothesis testing, we can replace tests such as ANOVA with the Kruskal-Wallis nonparametric rank test (Kruskal and Wallis, 1952; Vargha and Delaney, 1998), for example. In the case of MANOVA, instead of using the Hotelling-Lawley test statistic (Lawley, 1938; Hotelling, 1951), we can use the more robust Wilk's Lambda (Coombs et al., 1996). Pearson's $r$ for correlation can be substituted by Kendall's $\tau$ (Kendall, 1955), and linear regressions can be replaced with other non-linear functions, such as polynomial functions, or substituted with robust or generalised regression models.

As will be discussed in Courtenay et al. (Under Review-b), other statistical methods, such as PCA, can also be made robust, removing the assumption of linearity among data. From this perspective, non-linear Dimensionality Reduction (DR) can be considered a valuable new contribution to morphological studies.

Finally, in terms of interpreting these results, a more robust approach to statistical hypothesis testing should take into consideration other factors as well, such as the probability a $p$-value could lead to a mistaken conclusion (Colquhoun, 2017; Held and Ott, 2018; Colquhoun, 2019; Benjamin and Berger, 2019). In most traditional statistical applications throughout the XXth century, scientists, especially those without a statistical background, have used the (in)famous $p < 0.05$ as a criteria to accept or reject a hypothesis regarding their data (Wasserstein and Lazar, 2016; Kennedy-Schaffer, 2019; Wasserstein et al., 2019). Recent debate in the statistical community (Wasserstein and Lazar, 2016; Wasserstein et al., 2019), however, has pushed for the renewal of acceptance criteria, stating $p < 0.05$ to be a value associated with a high probability of reporting a False Positive (Colquhoun, 2019). The present body of work supports this, and proposes the use of $p < 0.003$, providing a detailed justification behind this observation in Appendix C.

## 2.4 Classification and Artificially Intelligent Algorithms

### 2.4.1 Traditional Morphometrics and Geometric Morphometrics

A very common goal in morphological analyses is classification. Once variables have been extracted and analysed statistically, analysts tend to seek a means in which to use this data to classify new samples. This could be useful for many purposes, including the identification of new species, identifying anomalies, or as a helpful tool that the analyst can exploit to make more empirically informed decisions. The idea behind classification is to take a set of observations ($x$), of known origin (i.e. presenting class labels $y$), and construct a mathematical function, $f(x)$, that maps these variables to their origin ($f(x) = y$) (Fig. 2.11). Additionally, for scientific rigour, classification tasks tend to stress the importance of calculating the probability that the observations originate from their corresponding class. In other words, given a new observation, if we were to use $f(x)$ to predict its origin, we would like to know the probability that this prediction is correct.

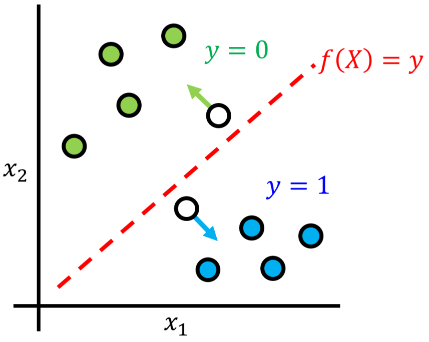


***Figure 2.11:*** *Description of how a classification algorithm constructs a discriminant function ($f(X)$) to assign new observations (white dots) to either one of the groups (blue and green dots).*

Currently the most popular used classification algorithms in morphometrics are the Linear Discriminant and Partial Least Square Discriminant functions. Use of these algorithms in a study is typically referred to as Linear Discriminant Analysis (LDA) and Partial Least Square Discriminant Analysis (PLSDA), respectively. LDA utilises a maximum likelihood classification rule to assign individuals to a group. This is done by maximising the ratio of between-group sum of squares to within-group sum of squares (Fisher, 1936; Mitteroecker and Bookstein, 2011). PLSDA, on the other hand, attempts to regress sample information so as to maximise the covariance between independent variables and class information (Barker and Rayens, 2003).

LDA is similar to CVA as Fisher (1936)'s linear function is equivalent to the first Canonical Variate from CVA (Mardia et al., 1997). Nevertheless, CVA is mostly used for ordination purposes (Mitteroecker and Bookstein, 2011). From this perspective, CVA can be considered a more "visual" discriminant function, seeing how separation among samples in just a single CV score is indicative that the distributions are separate (Mitteroecker and Bookstein, 2011). This particular property has been exploited in the past with the use of statistical distances, such as the Mahalanobis distance, to perform a more statistically oriented type of classification (Albrecht, 1992; Klingenberg and Monteiro, 2005). LDA, however, provides a direct means of constructing a decision boundary that can be used to calculate the probability of class association.

Both CVA and LDA are closely related to analyses such as PCA (Fig. 2.12), as they construct feature spaces that are derived from the eigendecomposition of data. In contrast to PCA, however, CVA and LDA take into account group labels during this procedure, constructing decision boundaries in feature spaces according to group covariance matrices. One alternative to CVA is the projection of data onto the principal components of the group averages, commonly referred to as bgPCA. As opposed to CVA, the PC scores from bgPCA are now orthogonal, and are also equivalent to the feature space used by PLSDA in the construction of decision boundaries (Barker and Rayens, 2003; Boulesteix, 2004).

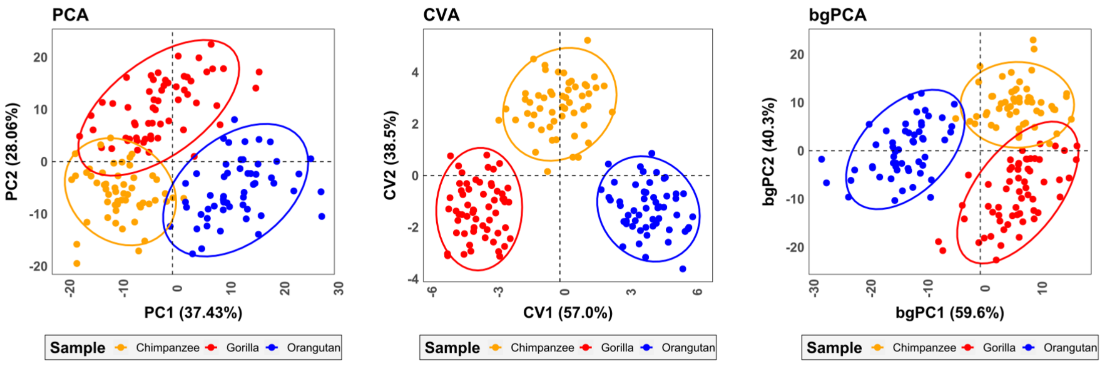


***Figure 2.12:*** *Examples of ordination and discrimination of 3 different species studied using geometric morphometrics from O'Higgins and Dryden (1993). Left panel - traditional ordination of data using Principal Component Analysis (PCA). Middle panel - Canonical Variate feature space (CVA). Right panel - Between-group Principal Component feature space (bgPCA). Linear Discriminant Analyses construct decision boundaries from a similar feature space to CVA, while Partial Least Discriminant Analyses construct these boundaries from bgPCA.*

A strong limitation both LDA and PLSDA have in common are their underlying parametric nature. LDA, for example, makes the assumption that the PDF of each group's multivariate distribution is Gaussian in nature. PLSDA, on the other hand, uses orthogonal and linear transformations to construct a feature space for discrimination, a process which is also parametric (Barker and Rayens, 2003; Boulesteix, 2004; Mitteroecker and Bookstein, 2011).

### 2.4.2 Basic Notions of Artificial Intelligence

Over the years, a field that has become a growing protagonist in classification tasks is the field of Artificial Intelligence (AI), and the use of Artificially Intelligent Algorithms (AIAs). AIAs are highly versatile, and in many cases, much more robust with less parametric restrictions. Likewise, they have proven highly successful over the years in the processing of morphological data (Dobigny et al., 2002; Baylac et al., 2003;

Lorenz et al., 2015; Soda et al., 2017; Cuthill et al., 2019; Courtenay, 2019; Courtenay et al., 2019b, 2020b; Quenu et al., 2020; Aramendi, 2021).

AI is the field of science dedicated to the development of computer systems able to perform tasks that would normally require biological intelligence. From this perspective, the goal of AI is to construct algorithms that replicate our own cognitive functions, such as seeing, hearing, and decision making. AIAs are thus tools that can be used in devices that can interpret external data, and make decisions that optimise the chances of achieving a goal (Nilsson, 1998; Poole et al., 1998; Russel and Norvig, 2003; Legg and Hutter, 2007; Kaplan and Haenlein, 2018).

The basis of all AI is the concept of learning. As with human cognition, learning can be summarised by the ability to take experience, and utilize it in order to improve its performance on similar experiences in the future (Mitchell, 1997). The basic learning process thus consists in data retrieval (*input*), whether this be through memory, observations, or experience, which is then translated into a broader representation of data (*abstraction*), with a final output where the abstract data is used to form the basis behind an action (*generalisation*). This concept can be used to find a mathematical function that maps the input to the output, known as a *model*. Models represent explicit descriptions of data through structured patterns, and the process of learning these patterns is known as *training*.

The actual training of an AIA is performed using a series of mathematical concepts that employ a combination of linear algebra, calculus, probability theory, and statistics (Bishop, 1995, 2006; Goodfellow et al., 2016). Through the combination of these tools, AIAs adjust their own internal parameters, known as *weights*, that are used to minimise the error (the *loss*) of the algorithm (Zhang, 2004b; Duchi et al., 2011; Kingma and Ba, 2015). Nevertheless, while reduction of error is important, a fundamental component of any AIA is the concept of *generalisation* (Goodfellow et al., 2016). An algorithm is only powerful if it is able to apply the information it has learnt to external and foreign problem solving tasks. If an algorithm is found to systematically reach imprecise conclusions, then this is often referred to as model *bias* (Cover and Thomas, 1991). On the other hand, if algorithms are found to be overly sensitive to small fluctuations in data (such as noise), then this is referred to as a model's *variance* (*ibid*). The bias-variance tradeoff (Breiman, 1996a,b), therefore, is a fundamental means of detecting or controlling whether an algorithm is *underfitting*, or *overfitting* (Fig. 2.13);

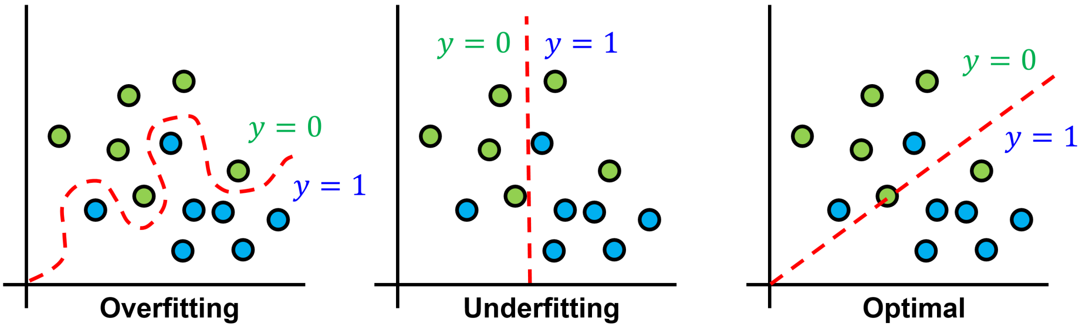


***Figure 2.13:*** *Examples of mathematical functions that have either under or overfit on a set of data for classification.*

**Definition 17** *Overfitting* Overfitted models are those that fit a function too closely to a particular dataset, and are, therefore, unable to adapt to unknown observations reliably.

**Definition 18** *Underfitting* Underfitted models are those that are unable to adequately capture the underlying structure of data.

Underfitting and overfitting can be produced under many different circumstances. To provide analogies with the previously discussed concepts of parametric and nonparametric statistical models, an algorithm may underfit due to an incorrect selection of an AIA (e.g. using a linear model to perform non-linear tasks). Overfitting, on the other hand, is often produced when the data used to train the algorithm poorly represents the true domain, or the dataset is too small.

Detecting this phenomenon is often difficult. Nevertheless, the most common means of assessing an algorithm's performance is through their evaluation during and after training. AIAs are typically evaluated using subsets of the dataset, known as *training*, *testing* and *validation* sets (Goodfellow et al., 2016; Chollet, 2017). The training data is provided to the algorithm alongside the validation set, using the training data to fit the model, and then performing self-evaluation on the validation set so as to learn from its mistakes and adjust the internal weights (Goodfellow et al., 2016). The test set is kept separate at all times, and only used once the process of training has finished so as to evaluate the real performance of the AIA (see Appendix D.3).

The means in which the algorithm is exposed to the validation set may vary, with a number of different algorithms available for this process. One of the most common techniques is known as *cross-validation*; a resampling technique that uses different portions of the data to provide an algorithm with a train and validation set (Stone, 1974, 1977). For more traditional approaches, such as PLSDA and LDA, Leave-One-Out Cross Validation (LOOCV) is normally used. LOOCV consists in fitting the algorithm over a number of iterations, and with each iteration a single observation is removed and separated for validation, while all other observations are used for training. This method is powerful for assessing how robust an algorithm is to outliers. In most AI applications, however, the $k$-fold cross-validation algorithm is more popular (Bengio and Grandvalet, 2004). In $k$-fold cross-validation, the training data is split into $k$ sets, the smaller set is used for validation, while the larger set is used for training. For each iteration of cross-validation, the algorithm is fit to the training data, and adjusts its own weights by assessing the error when used to predict the validation data. Finally, in more advanced applications, algorithms are exposed to a fixed percentage of data for both training and validation, while over each iteration (known as an *epoch*), the data is typically shuffled (Bishop, 1995).

Training and validation are used for the algorithm to adjust its own parameters, however, a true test of the AIA's ability to process new data is by exposing it to the test set. The algorithm is thus used to classify the test set, and a confusion matrix is calculated to evaluate the number of times the algorithm has correctly or incorrectly classified the data (see Appendix D.3). From this perspective, powerful algorithms should be able to perform well on both the training and test sets. In cases where algorithms have learned the training data well, but fail to perform on the test set, this is a strong indication of overfitting.

### 2.4.3 Machine, Deep, and Computational Learning

Many learning paradigms exist in AI (e.g. Caruana, 1997; Ben-David et al., 2010; Zhang and Ma, 2012; Ravichandiran, 2018a; Kuow and Loog, 2019; Wirsansky, 2020), however the most common three means of training an algorithm are known as *Supervised*, *Unsupervised*, and *Reinforcement* learning (Ravichandiran, 2018b; Bonaccorso, 2019). The present Doctoral Thesis focuses primarily on supervised approaches, however some unsupervised approaches have also been explored throughout this body of research (e.g. Courtenay and González-Aguilera, 2020; Courtenay et al., 2021b).

**Definition 19** *Supervised Learning* The process of training an algorithm to learn on data while explicitly defining the objectives of the task at hand; given a particular input, we wish to map this input to a specific output ($f(x) = y$).

**Definition 20** *Unsupervised Learning* The process of training an algorithm to learn on data without explicitly defining the objectives of the task at hand; given a particular input, the algorithm is required to detect patterns or learn the overall representation of the data without external help (typically represented as $f(x) = x'$).

The main difference between the two is the type of output that is produced. Supervised approaches have a defined output ($y$) that the algorithm has to predict. Supervised algorithms are very popular, especially for classification and regression tasks. In unsupervised learning, however, most algorithms are asked to reconstruct the input, thus producing a version of the input as output ($x'$). The internal properties of these algorithms can then be used for pattern recognition to detect clusters or underlying features of the data. In other unsupervised approaches, the algorithm simply tries to detect patterns based on the density of information, or a defined mathematical function that can be used to propose a means of evaluating error. The means in which error is then reduced for both supervised and unsupervised approaches is often performed using "back propogation" (Fig. 2.14), where the error produced by the algorithm is passed back and used to adjust the internal algorithm's weights (Zhang, 2004b; Duchi et al., 2011; Kingma and Ba, 2015).

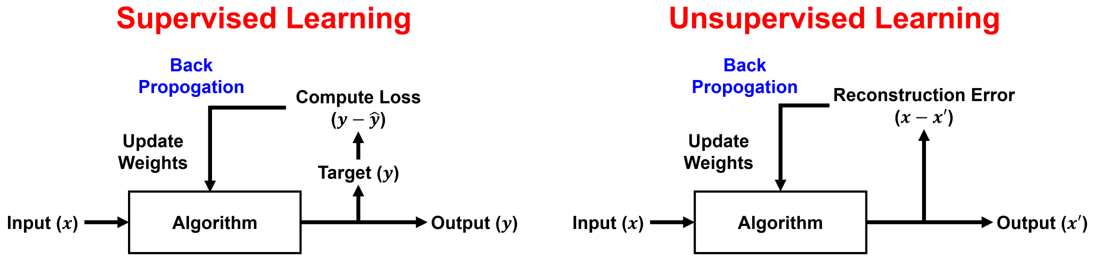


***Figure 2.14:*** *A figurative description of how supervised and unsupervised algorithms learning work.*

Nevertheless, the two most common words that are associated with the term "learning" in AI are the terms *Machine* and *Deep Learning* (ML & DL). ML can simply be defined as;

**Definition 21** *Machine Learning* The field of computer science focused on the development of algorithms for transforming data into intelligent actions.

Machine learning algorithms can be divided into a number of different categories (Fig. 2.15). Some of the most common types of algorithms are Decision Tree based algorithms, distance based algorithms, probabilistic algorithms, Discriminant Functions, and Neural Networks (NNs).

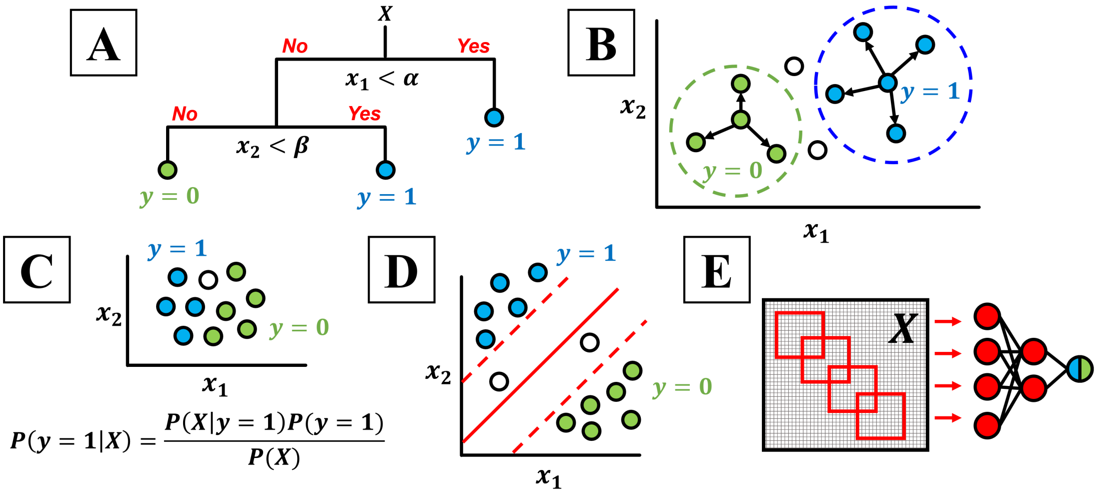


***Figure 2.15:*** *Graphic examples of AI classification algorithms, including (A) a Decision Tree, (B) a Nearest Neighbour Classification algorithm, (C) a Naïve Bayes Classifier, (D) a Discriminant Function using a Support Vector Machine with maximised margins, and (E) an example of a Convolutional Neural Network. For each example, variables $x_1$ and $x_2 \in X$ are used to predict the label y. In (A) variables α and β indicate any threshold or acceptance criterion, and in (C) the formula uses Bayes Theorem to predict the probability (P()) of the label y given values of $x \in X$*

Decision tree based algorithms, such as the Conditional Inference Tree (CTREE: Hothorn et al. (2006)), the C5.0 decision tree algorithm (C5.0: Quinlan (1992)), and more complex variants such as the Random Forest (RF: Ho (1995)) or Gradient Boosting (GB: Friedman (2002)) algorithm, generally present a flowchart-like structure, whereby decisions are made by following each branch of the tree according to a set of criteria, until reaching a final leaf or "node" that defines the final decision to be made by the tree (Fig. 2.15a). Distance based algorithms, such as the *k*-Nearest Neighbour algorithm (KNN: Breiman (1951)), are typically used for unsupervised learning tasks (e.g. clustering). Nonetheless, distance based algorithms can also be leveraged for classification by predicting class labels for observations based on the group they are closest to (Fig. 2.15b). Probabilistic algorithms, such as Naïve Bayes (NB: Zhang (2004a)), utilise some form of probability theory, typically Bayes' theorem, to calculate the probability of association of an observation to a group (Fig. 2.15c).

From the perspective of discriminant functions, both the aforementioned LDA and PLSDA algorithms from traditional morphometrics can be trained using ML based approaches. Nevertheless, one of the most powerful algorithms for classification tasks of this nature is the Support Vector Machine (SVMs: Cortes and Vapnik (1995)). SVMs are a highly powerful modeling algorithm, useful for many types of AI tasks. Unlike typical discriminant functions, SVMs use "soft maximised-margins" (Fig. 2.16), conditioned by a cost (*c*) parameter, that define the decision boundary. The term "maximised-margins" introduces two main concepts; (1) a *margin* is the perpendicular distance between the closest data point and the decision boundary, while (2) the objective of an SVM is to *maximise* this margin. A decision boundary is then considered to be *soft* if the algorithm is allowed to account for outliers. From this perspective, the *c* parameter of an SVM defines how lenient the decision boundary may be. This is a fundamental means of avoiding overfitting.

The main power of the SVM, however, is found in the concept of a *kernel trick*, which is used to overcome traditional limitations imposed by linearity (Bishop, 2006). A kernel is a mathematical function that performs a pre-transformation of the data (Fig. 2.16a), projecting data-points into an implicit feature space (Fig. 2.16b), where the construction of a linear function is easier to compute. If we were to flatten this implicit feature space back to the original dimensions of the dataset, then the decision boundary constructed using the kernel will appear to be non-linear (Fig. 2.16d). This decision boundary is typically referred to as a *hyperplane*, as it is essentially a linear function that exists across the multiple dimensions of both the original and the constructed domain. A number of kernel functions exist for SVMs, however two of the most popular are the Polynomial function, and the Radial Basis Function (Bishop, 2006).

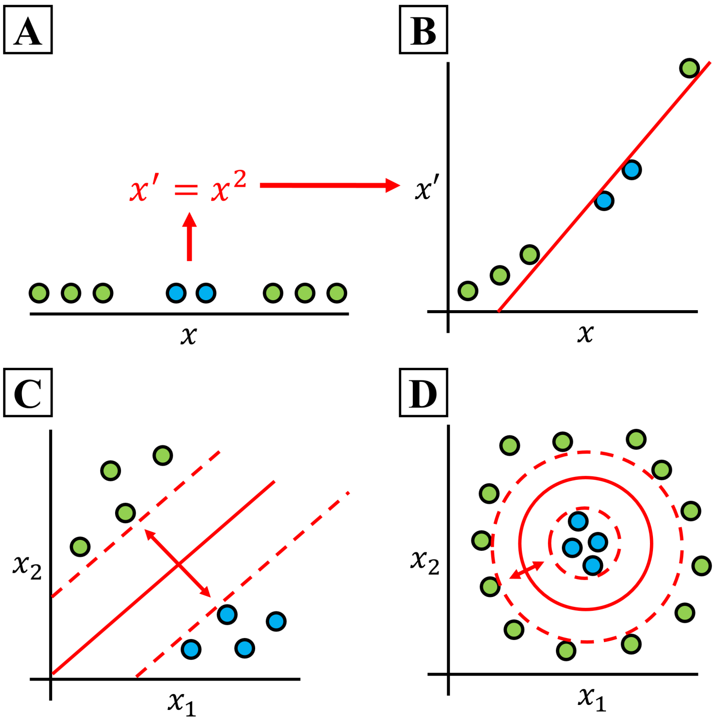


***Figure 2.16:*** *Theoretical examples of non-linear Support Vector Machine maximised margin decision boundaries and kernel functions. (A) A simple univariate dataset where constructing a linear decision boundary to separate between the two groups is not possible. (B) A transformation of variable x from part (A), using $x^2$ as a kernel, i.e. a basic polynomial kernel. A decision boundary separating the two groups can now be easily constructed in the new feature space. (C) A typical linear example of an SVM with maximised-margins. (D) An example of a non-linear SVM with maximised-margins implementing a kernel.*

The final main type of algorithm is the Neural Network (NN). NNs, originally called Multilayer Perceptrons (McClelland and Rumelhart, 1986; Rumelhart et al., 1986), are models that attempt to replicate the functionality of a biological brain. They do so by using a simple mathematical formula that represents a neuron, which can then be linked to other nodes in a "Network" (see Appendix D). From this perspective, having hundreds of interconnected nodes can be considered a computational analogy of a biological brain. So as to truly replicate the brain, NNs additionally have a series of *activation functions*, that define how information flows through the network (see Appendix D.1 and D.2). These activation functions are the most efficient means of designing a non-linear algorithm that is capable of sorting through information. NNs are also typically referred to in relation with the word *architecture*, which defines the way layers of neurons are arranged (Fig. 2.17). From this perspective, the use of different architectures reveal NNs to be highly versatile, but can also be very complicated to design, implement and train. NNs additionally have hundreds of different internal parameters that can be used to fine tune and improve learning performance, and are thus considered one of the most powerful AIAs to exist.

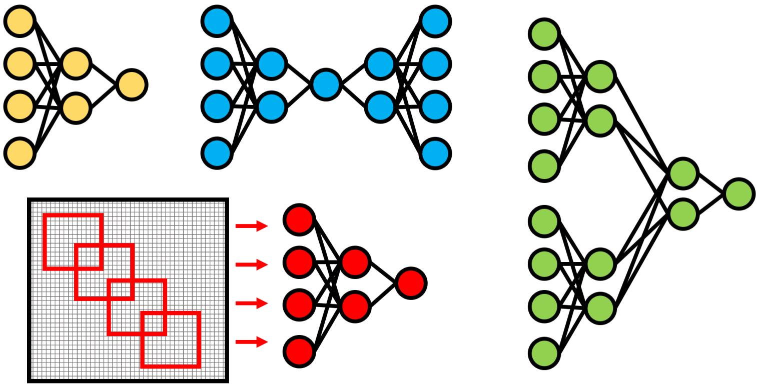


***Figure 2.17:*** *Some figurative examples of Neural Network architectures. The yellow NN represents a simple Feed Forward Neural Network. The blue NN represents an Autoencoder, which is typically used for unsupervised tasks such as dimensionality reduction. The red NN represents a Convolutional Neural Network, which is popular for the processing of image data. The green NN represents a Siamese Neural Network; a useful algorithm for comparing pairs of observations. For ease of visualisation, no Bias neurons (see Appendix D) have been included in this figure.*

NNs, however, are not typically used in ML, and when they are, they are relatively small (only a few neurons in a single layer). From this perspective we begin to define the concept of DL, where the NN is the protagonist.

The difference between ML and DL is often confuse, with DL presenting a number of possible definitions (Brownlee, 2016). DL is a subfield of ML, primarily defined by the use of NNs. ML on the other hand, tends to envelope the much wider range of different algorithms previously described. The word “Deep”, for many, refers directly to the “depth” of the NN, i.e. the number of layers included in the algorithm. For others, NNs are only considered deep if they include a particular architecture, known as a Convolutional Neural Network (CNN). Nevertheless, the question is: how many layers does a NN need to be considered deep?

Here we will refer mostly to DL in light of the definition from Goodfellow et al. (2016)’s iconic book, aptly titled *Deep Learning*:

> "*Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined in relation to simpler concepts, and more abstract representations computed in terms of less abstract ones.*" - Goodfellow et al. (2016), Page 8

This quote touches directly on a component frequently mentioned in DL literature, which is the concept of high- and low-level features. Bengio (2012) defines DL to be algorithms that “*seek to exploit the unknown structure in the input distribution [. . . ] often at multiple levels, with higher-level learned* *features defined in terms of lower-level features*". Similarly, LeCun et al. (2015) state DL algorithms to present "*. . . multiple levels of representation, obtained by composing simple but non-linear models that each transform the representation at one level [. . . ] into a representation at a higher, slightly more abstract level. . .* ". Analogous to how the human brain is capable of deconstructing a complex element into multiple smaller and simpler components (Fig. 2.18), DL algorithms aim to convey and extract the most important features of data where the majority of meaning is held, and use these extracted features to make decisions. This is especially important if we consider the amount of noise typically found within data, where AIAs are more likely to be able to extract the essential features that convey meaning.

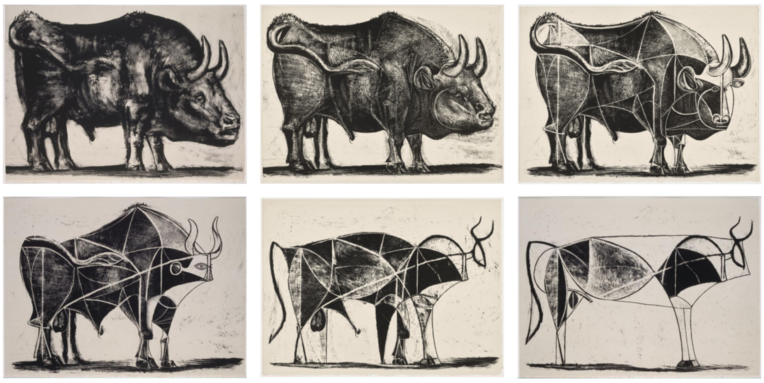


***Figure 2.18:*** *Examples of different levels of abstraction, as displayed by six of the lithographs from Pablo Picasso's Le Taureau, stages II to VII (left to right, top to bottom). Here Picasso begins with a more realistic representation of the bull, followed by the extraction of basic features that still convey and depict the general essence of the animal, despite being represented by gradually less features. Images were obtained from the Musée National Picasso-Paris' online collection (https://www.museepicassoparis.fr - Accessed: 07/07/2022).*

As will be seen throughout this Doctoral Thesis, the words *Deep* and *Machine* will be used often. Nevertheless, over the course of this research, a more general term has become more apparent, which encompasses elements of both ML and DL; *Computational Learning* (CL). Our proposal for the use of the term CL over ML and DL stems mostly from the concept of an algorithm's *Hyperparameters*.

Hyperparameters are internal parameters that each algorithm has which are used to fine-tune and control the learning process. These should be differentiated from traditional AI *parameters*, which are typically the internal weights of an algorithm. Additionally, hyperparameters are defined by the user, while parameters are adjusted by the computer during the learning process. SVMs, for example, have three main hyperparameters; the cost ($c$) of the hyperplane, the type of kernel to be used, and the parameter that defines how much influence the kernel has on the new dimension (typically referred to as $d$, $\alpha$, or $\gamma$, depending on the kernel) (Rahimi and Recht, 2007; Wiering et al., 2013; Jacot et al., 2020; Tancik et al., 2020; Courtenay et al., 2022b). The NN, however, has hundreds of hyperparameters, of which the "depth" (e.g. number of neurons or layers), is just one of them (Brownlee, 2019).

While a simple SVM can be programmed in just a few lines of code, more complex SVMs (e.g. Courtenay et al., 2021b, 2023) can also be defined customising every element of the algorithm's internal functioning. Likewise, the NN from Courtenay et al. (2021b) consists in 118 lines of functional Python code, however the simplest of NNs can also be defined using a single line of R code as well. From this perspective, even if a NN is not used in a study, this is not to imply that the complexity of the AIA is any less than that of a "deep" NN. In light of this, a different type of "depth" can be proposed, defining the depth of the algorithm in terms of its hyperparameters. So as to avoid the assumption that DL is exclusive to the use of NNs, we propose the use of the term CL, that encompasses all types of AIAs, without the need to specify how "deep" the programmer goes into the internal functioning of the algorithm.

### 2.4.4 Limitations of AI

AI popularity has grown exponentially only over the last 10 years, especially due to its impressive above-human level performance in many CV based tasks. Nevertheless, as with any major trend in scientific research, the eventual discovery of a method's limitations can cause major problems in a widespread use of these techniques, leading to the need for research in the overcoming of these limitations (Fenn and Rasinko, 2008). While a chapter on the limitations of AI could confront this topic from a number of different angles, the present Doctoral Thesis will focus primarily on a limitation known as the "Curse of Dimensionality".

The "Curse of Dimensionality", a term originally coined by Bellman (1957, 1961), refers to the difficulties of modelling from complex and high dimensional data, especially when used to describe a small set of observations. Given a high number of variables, each with their own statistical properties and complexities, analysts need increasingly larger sample sizes in order to capture the true variability of the population. Nevertheless, as is common knowledge in most fields of science, large sample sizes are very hard to obtain.

AIAs are notorious for needing very large datasets, especially in the case of NNs. The pioneering study by LeCun et al. (1998), for example, reached $> 95\%$ accuracy with a dataset of 58,527 monochrome $28 \times 28$ px images (the MNIST dataset[1]). Between the years 2009 and 2012 pivotal research in AI was carried out using two very large datasets, one containing 60,000 coloured $32 \times 32$ px images (the CIFAR-10 dataset[2]), and the second containing 15 million coloured $469 \times 387$ px images (the ImageNet dataset[3]). In most of these cases, authors struggled to reach 80% accuracy in image classification (Deng et al., 2009, 2010; Krizhevsky and Hinton, 2009; Krizhevsky, 2010; Krizhevsky et al., 2012). Nevertheless, considering the complexity of these images, as well as the number of image classes in each dataset, these studies can still be considered some of the most important advances in CV to date. In Szegedy et al. (2014a), the authors publish more innovations in NN architectures that are able to reach 90% classification with 1.35 million images from the ImageNet dataset. Similarly, Redmon et al. (2016) reach $> 90\%$ accuracy for object detection and recognition when using 9 million coloured video frames (the ImageNet10k dataset[4]).

While in each of these cases the input to algorithms are images, as opposed to measurements, these present perfect examples of how much information many algorithms require in order to learn a specific task. In many applications the human eye is able to almost immediately detect that two images are different. For an AIA, however, this task is much more difficult, as most AIAs still struggle to capture fundamental components of human cognition, such as common sense (George et al., 2017), true generalised knowledge across different tasks (Finn et al., 2017; Ravichandiran, 2018a), the leveraging of prior knowledge (Fink, 2004; Santoro et al., 2016), and foresight (Finn and Levine, 2017). This forms a fundamental component of biological cognition and conditions how we make decisions. In the case of numeric data such as measurements and coordinates, the complexity of an algorithm is dependent on the underlying statistical properties of the data; high overlap between complex multivariate and non-parametric distributions implicates the need for

[1] http://yann.lecun.com/exdb/mnist/

[2] https://www.cs.toronto.edu/~kriz/cifar.html

[3] https://www.image-net.org/

[4] http://lear.inrialpes.fr/~akata/ImageNet_res.php

increasingly more complex algorithms in order to model from this information (Diakonikolas et al., 2017, 2019).

From the perspective of classification tasks in traditional morphometrics, much debate has gone into the required sample sizes needed for different types of analyses (Cardini and Elton, 2007; Cardini et al., 2015; Bookstein, 2017). LDA, for example, is known to require that the size of the smallest group be larger than the number of predictor variables (Mitteroecker and Bookstein, 2011). This is due to how, given a sufficiently large number of variables, Canonical Variates separate groups regardless of the actual probability distribution (*ibid*). PLSDA, on the other hand, was long considered a better algorithm for the analysis of smaller datasets (Barker and Rayens, 2003; Mitteroecker and Bookstein, 2011). Recent investigation into the algebraic properties of bgPCA, however, has led researchers to retract this statement (Bookstein, 2019), considering PLSDA to be equally problematic (see Fig. 2.19). From this perspective, the ability of bgPCA to produce the spurious separation of small groups in ordination tasks is likely to imply an exaggerated accuracy of PLSDA (Cardini et al., 2019; Cardini and Polly, 2020).

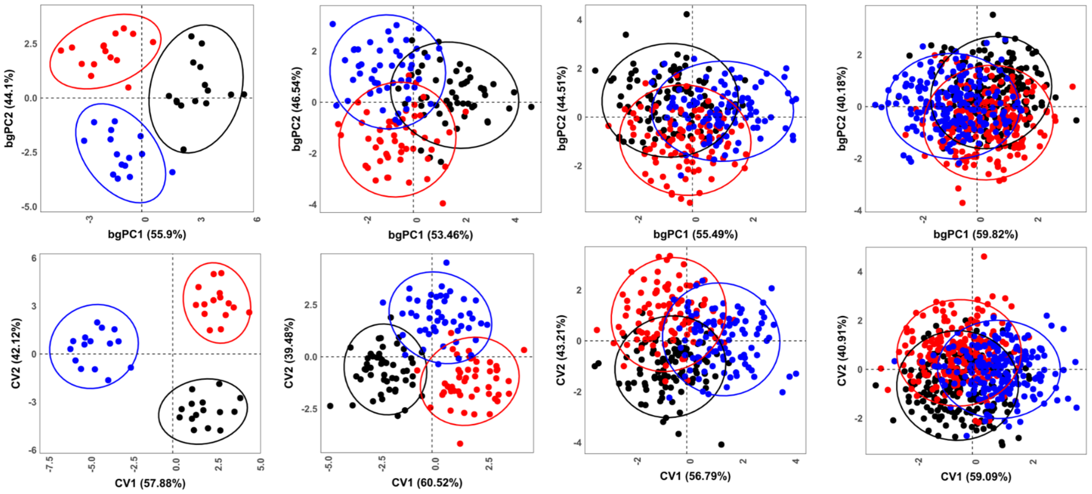


***Figure 2.19:*** *An example of how the curse of dimensionality has an effect on traditional morphometric and Geometric Morphometric tests such as bgPCA (upper panels) and CVA (lower panels). This in turn will effect their algebraic classification counterparts; PLSDA and LDA. The phenomenon described in this figure consists in first simulating a theoretical three-group set of identical centagons, described by 100 landmarks in two dimensions (200 morphological variables in total). The number of individuals per group on all accounts are lower than the required sample sizes for these types of analyses. As can be seen, although differences between groups should not exist, the curse of dimensionality for these particular ordination tasks produces the spurious separation between groups.*

As a basic rule of thumb, the minimum sample size needed for most classification tasks in traditional morphometrics, therefore, must be larger than the number of measurements taken (Bookstein, 2017, 2019). Unfortunately, this rule is slightly harder to fulfil in GMMs. The number of predictor variables in GMMs is $\mathbb{R}^{p \times k}$, where $p$ is the number of landmarks, and $k$ the number of dimensions. In the example of a simple two-dimensional triangle, morphological variables lie in a $\mathbb{R}^{3 \times 2}$ feature space, therefore over 6 individuals are needed per group. Even when performing analyses on variables pre-processed via PCA, in many cases the number of individuals is still lower than the number of variables (e.g. the cases of Wu et al., 2010; Freidline et al., 2012; Daver et al., 2018; Détroit et al., 2019), while the selection criteria for the optimal number of PC scores is usually unclear (Jollife, 2002; Bishop, 2006). For analyses such as EDMA, the number of predictor variables can be much greater, considering a simple interlandmark distance matrix for a single individual to be of size $\mathbb{R}^{p \times p}$.

Beyond dataset size, however, the importance of dataset quality is evidently equally as important. For an algorithm to successfully learn how to perform a task and generalise weights to be robust to new observations, the data provided must capture, to the best of its abilities, the natural variability of the population. Evidently this is a hard task to perform, as in most cases the true nature of the population is unknown, and sampling bias is likely to introduce noise.

Over the years, some research into how and why NNs fail to perform tasks has led to a number of interesting lines of research. Su et al. (2019), for example, prove how the modification of just a single pixel causes algorithms to incorrectly classify images (Fig. 2.20), even if their original performance reached $\approx 87\%$ accuracy when trained on the CIFAR-10 dataset. In this study, authors showed how the perturbation of a single pixel can cause a drop in accuracy as low as 19.82%. Similarly, in both Szegedy et al. (2014b) and Goodfellow et al. (2015), the authors present how a slight, imperceptible perturbation to the input of a neural network causes an algorithm to fail (Fig. 2.20).

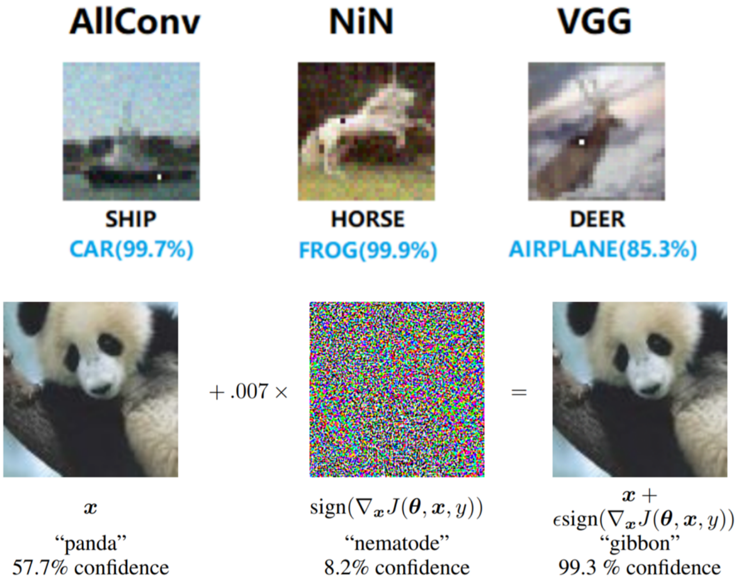

AllConv
NiN
VGG
SHIP
CAR(99.7%)
HORSE
FROG(99.9%)
DEER
AIRPLANE(85.3%)
$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

***Figure 2.20:*** *Examples of studies that have shown the limitations of NNs by fooling them during image classification. Upper panels: examples of how eliminating a single pixel causes NNs to fail; a ship is predicted to be a car with 99.7% confidence, a horse a frog (99.9%), and a deer an airplane (85.3%). Source: Su et al. (2019). Lower panels: examples of how introducing a fine layer of slight Gaussian noise causes algorithms to fail; a panda is predicted to be a gibbon with 99.3% confidence. Source: Goodfellow et al. (2015)*

In each of these cases, it can be seen how, even when a large dataset is provided and found to train an algorithm efficiently, the introduction of a slight amount of noise causes algorithms to produce erroneous classifications. This proves how not all algorithms are necessarily robust unless they are provided with the correct type of data that allows them to be. Algorithms are thus unable to capture the true variability of the domain and fail to generalise when exposed to new observations.

One of the means of overcoming some of these issues is known as *Data Augmentation*;

**Definition 22** *Data Augmentation* A technique used to simulate new data by increasing the density and diversity of the original data provided.

In most CL literature, data augmentation is typically used in CV applications as a means of slightly transforming a set of images, so that they appear to be slightly different, and thus allow algorithms to learn more robust features. This technique typically includes the rotation, shearing, flipping, and adjustment of contrast/brightness of an image (Goodfellow et al., 2016). Beyond image data, however, data augmentation techniques aim to simulate new information based on statistical and probabilistic criteria. From this perspective, algorithms are used to model from a distribution, and thus try and increase the variability of the sample based on the information that is already there. This approach can additionally be performed using AI itself, in the context of unsupervised learning algorithms. Means in which to overcome problems with sample size, however, will be discussed in greater detail in Courtenay and González-Aguilera (2020) amd Courtenay et al. (Under Review-a) (Chapter 3, Section 2).

# Chapter 3

# Scientific Publications

This chapter is dedicated to the scientific publications produced over the course of this Doctoral Thesis. A total of nine articles are presented, seven accepted and published, with an additional two under review, confronting multiple topics in relation with the extraction of morphological data (Courtenay et al., 2020a), statistical processing of this information (Courtenay et al., 2021a), simulation of new data for the purpose of data augmentation (Courtenay and González-Aguilera, 2020; Courtenay et al., Under Review-a), as well as the construction of classification algorithms (Courtenay et al., 2020b, 2021b). In addition to these methodological studies, we present two case-studies, one application using Geometric Morphometrics Courtenay et al. (2023), and a second using Fourier Descriptors to analyse "forms without landmarks" (Courtenay et al., 2022a). Finally, as a proposal for improvements in future research, we describe a new mathematical model for the analysis of Geometric Morphometric data (Courtenay et al., Under Review-b).

Prior to each article, a Spanish translation of the title and abstract have been provided, as well as the links to all associated Supplementary Materials and code.

*Publications: Section 1*

# Obtaining Data and Robust Statistics

*Spanish Translation of Title and Abstract*

# Alcanzando nuevas resoluciones en el análisis morfológico de las marcas de dientes de carnívoros: Una nueva actualización metodológica para la tafonomía digital.

La investigación actual en los campos de la arqueología y paleontología se caracteriza por un crecimiento exponencial de la integración de nuevas tecnologías. No obstante, aunque dichos avances son de gran importancia para las múltiples líneas de investigación relacionadas con estos dos campos, su mejora y actualización debería ser una prioridad. Aquí presentamos un análisis del error inter e intraobservador en los análisis morfométricos de las marcas de dientes tipo depresión, fossa o "pit", generadas por carnívoros. Para ello, aplicamos múltiples herramientas de la estadística robusta a muestras experimentales de carnívoros modernos. Primero, con el objetivo de comprender la influencia de los errores de medición en la recopilación de datos morfométricos, realizamos una evaluación estadística sobre las coordenadas de *landmarks* 3D en bruto, las coordenadas totalmente superpuestas, y las coordenadas parcialmente superpuestas. Los modelos 3D fueron obtenidos mediante escáneres de luz estructurada. Las muestras experimentales utilizadas para realizar este análisis incluyen depresiones de dientes producidas por perros y lobos en diferenctes contextos, originalmente utilizadas en el análisis de ataques al ganado extensivo. Los resultados de este estudio remarcan la importancia del tipo de *landmark* a la hora de estudiar el error inter e intraobservador. También demuestra el valor que pueden tener los modelos de *semilandmarks* deslizantes, por encima de los *landmarks* fijos tipo III. Además de esto, los datos revelan la importancia de la experiencia del observador a la hora de tomar los datos, junto con un aumento del error cuando se trabaja con coordenadas totalmente superpuestas, debido al *"efecto Pinocho"*. Gracias a este estudio se ha podido actualizar los modelos de morfometría geométrica para el estudio de marcas de dientes tipo depresión. Este modelo híbrido de *landmarks* fijos tipo II y *semilandmarks* deslizantes reducen considerablemente el error inducido por el observador, generando un método más fiable desde el punto de vista métrico. Estos resultados pueden considerarse un avance fundamental para los estudios sobre carnívoros, con impacto en la investigación arqueológica, paleontológica y ecológica actual, así como en otras ciencias forenses.

*Supplementary Information and Links*

**Supplementary Information available from:**
https://doi.org/10.1371/journal.pone.0240328.s001
https://doi.org/10.1371/journal.pone.0240328.s002
https://vimeo.com/409256777

**Code and data available from:**
https://github.com/LACourtenay/GMM_Measurement_Accuracy_Tools

PLOS ONE

RESEARCH ARTICLE

# Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy

**Lloyd A. Courtenay**[1]*, **Darío Herranz-Rodrigo**[2,3], **Rosa Huguet**[4,5,6], **Miguel Ángel Maté-González**[1,7,8], **Diego González-Aguilera**[1,7], **José Yravedra**[2,3]

**1** Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Ávila, Spain, **2** Department of Prehistory, Complutense University, Madrid, Spain, **3** C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Madrid, Spain, **4** Institut Català de Paleoecologia Humana I Evolució Social (IPHES), Tarragona, Spain, **5** Department d'Historia i Historiao de l'Art, Universitat de Rovira I Virgili (URV), Tarragona, Spain, **6** Unit Associated to CSIC, Departamento de Paleobiologia, Museo de Ciencias Naturales, Madrid, Spain, **7** Gran Duque de Alba Institution, Diputación Provincial de Ávila, Ávila, Spain, **8** Department of Topographic and Cartography Engineering, Higher Technical School of Engineers in Topography, Geodesy and Cartography, Technical University of Madrid, Madrid, Spain

* ladc1995@gmail.com

**Citation:** Courtenay LA, Herranz-Rodrigo D, Huguet R, Maté-González MÁ, González-Aguilera D, Yravedra J (2020) Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy. PLoS ONE 15(10): e0240328. https://doi.org/10.1371/journal.pone.0240328

**Editor:** Efthymia Nikita, The Cyprus Institute, CYPRUS

**Received:** April 24, 2020

**Accepted:** September 24, 2020

**Published:** October 8, 2020

**Data Availability Statement:** All data used are held in a public repository and directly cited in the text. Data as well as any code used can be found at https://github.com/LACourtenay/GMM_Measurement_Accuracy_Tools while an additional supplementary video has been provided via https://vimeo.com/409256777

**Funding:** L.A.C. is supported by the Spanish Ministry of Science, Innovation and Universities with a FPI Predoctoral Grant (Ref. PRE2019-

## Abstract

Modern day investigation in fields of archaeology and palaeontology can be greatly characterised by an exponential growth of integrated new technologies, nevertheless, while these advances are of great significance to multiple lines of research, their evaluation and update over time is equally as important. Here we present an application of inter and intra-observer analysis in taphonomy based geometric morphometrics, employing robust non-parametric statistical analyses for the study of experimental carnivore tooth pit morphologies. To fully understand the influence of measurement errors in the collection of this data, our statistical assessment was performed on fully superimposed, partially superimposed and raw landmark coordinates collected from 3D surface scanning. Experimental samples used to assess these errors includes wolf and dog tooth pits used in modern day ecological livestock predation analysis. Results obtained from this study highlight the importance of landmark type in the assessment of error, emphasising the value of semi-landmark models over the use of ambiguous Type III landmarks. In addition to this, data also reveals the importance of observer experience for the collection of data alongside an interesting increase in error when working with fully superimposed landmarks due to the "Pinocchio Effect". Through this study we are able to redefine the geometric morphometric models used for tooth pit morphological analyses. This final hybrid Type II fixed landmark and semi-landmark model presents a significant reduction in human induced error, generating a more metrically reliable and replicable method that can be used for data pooling in future inter-institutional research. These results can be considered a fundamental step forward for carnivore inspired studies, having an impact on archaeological, palaeontological, modern-day ecological research as well as applications in other forensic sciences.

089411) associated to project RTI2018-099850-B-I00 and the University of Salamanca. R.H. is supported by MICINNFEDER PGC2018-093925-B-C32, the AGUAR project number SGR 2017-1040; the Universitat Rovira I Virgili (2014, 2015 and 2016 PFR-URV-B2-147) and CERCA Programme/ Generalitat de Catalunya. D.H.R. is supported by the Ministry of Science, Innovation and Universities, under the contract REF: PEJ2019-005420-A as part of the i+D+I Garantía Juvenil. Finally L.A.C., M.Á.M.G and D.G.A. are supported by the General Foundation of the University of Salamanca under the Plan TCUE 2018-2020; Project "WOLF_FOOTPRINT" (PC-TCUE18-20_2013); lead researcher D.G.A. Funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## 1. Introduction

The identification and analysis of the Bone Surface Modifications (BSMs) produced by different taphonomic agents is a fundamental component in modern archaeological, palaeontological and, more recently fields of applied forensic sciences. Among the different types of BSMs, in prehistory tooth marks present valuable information that greatly contribute to the understanding of site formation processes, palaeolandscapes and their associated ecologies. Under this premise, many authors have paid particular attention to the different tooth marks that different animals may produce, whether this be for the consumption of meat or other practices of biological origin such as osteophagia or general foraging activities.

Among tooth mark studies, mammalian carnivores are a particularly important protagonist in archaeological research [1–5], considering their competition for resources with hominins throughout the Pleistocene. Interestingly herbivores [6–8], omnivores [9, 10], insects [11], and birds [12, 13], are also known modifiers of osteological remains, causing some confusion in the classification and interpretation of BSMs. Rodents such as the porcupine (*Hystrix sp.*) are an additional well documented bone accumulator [4, 14–16], proving to be problematic agents in the interpretation of faunal accumulations. Likewise, chondrichthyans [17, 18] as well as large [19, 20], and very large reptiles [21–23], are also of great importance especially in paleontological research. Last but not least, human tooth modifications to bone are also a compelling line of investigation with great value in archaeology [24, 25].

Through these studies, current understanding of carnivore BSMs enumerate 4 main types of alterations produced by teeth [1, 5]. These include rounded circular depressions caused by the direct imprint of the cuspid on bone (*pits*); elongated depressions with a circular base produced by the dragging of teeth across the cortical surface (*scores*); circular holes produced by the direct penetration of the tooth through the cortical walls (*puncture*); and the progressive deletion of large portions of bone through continuous chewing (*furrowing*).

To date a great deal of investigation has focused their efforts on the systematic study of pits and scores. This type of research has gone through numerous different phases, starting with a simple qualitative *in-visu* characterisation of these traces [1–5, 17, 18, 22, 23], to a more in depth integration of quantitative analyses [10, 21, 25–29]. In many cases, the use of metric variables such as the length and width of tooth marks have proven valuable variables in the interpretation of archaeological and paleontological sites [19, 30–34, *inter alia*]. Nevertheless, a common limitation noted by most is the ability to go beyond the size of the carnivore chewing and specifically identify the agent(s) present in a site.

New possibilities have recently been made available through Geometric Morphometric Methods (GMM). GMM are a powerful multivariate statistical toolset for morphological analyses [35, 36], originating primarily from fields of biology, evolutionary sciences, systematics, and physical anthropology [37–44]. Outside of anatomy, however, GMM has had other successful integrations into other aspects of archaeology, such as the study of lithic materials [45, 46, *inter alia*].

For geometric morphometric analysis, the identification and quantification of points of interest, known as *landmarks*, are required. These points can be defined as "a point of correspondence on an object that matches between and within populations" [47: pg. 3]. One advantage of landmarks is that they can either be recorded as 2D or 3D coordinates in space, providing a direct means of visualising variations between morphologies through the distortion or displacement of each landmark on each individual [48, 49]. The configuration of these points in space additionally present the advantage of preserving the full geometry of the specimens being studied [37]. Finally with the introduction of powerful 3D modelling equipment and methodological protocols, the capturing of morphological data is much more precise.

In BSM taphonomy, GMM were initially introduced as a means of quantifying cut mark morphologies for both 2D derived data [50] and entire 3D models [51]. These advances have obtained exceptional resolutions, presenting means of identifying the lithic tool and raw material used to produce each cut mark in archaeology [50–52, *inter alia*], while also arising in forensic research for sharp force trauma identification [53, 54, *inter alia*]. Likewise, initial efforts in the quantification of tooth score and pit morphologies reached up to 80% classification rates differentiating between carnivore agencies [55–58], with Machine Learning algorithms obtaining > 95% accuracy [58, 59].

Regardless of this success, however, it is important to consider the value of the landmark models being used for geometric morphometric data collection. To name one primary issue, landmarks can be further divided into three different types; easily locatable landmarks (Type I) and landmarks whose location are more ambiguous (Types II and III). Type I landmarks in biology are usually anatomical features of a precisely defined nature (e.g. tendon insertions), while Type II landmarks are defined by local properties (e.g. maximum curvature) and Type III landmarks are constructed points across the structure (e.g. a centroid) [35, 47: pg. 3–4]. From this concept, each type of landmark possesses a certain degree of error and subjectivity that must be addressed [60, 61]. Possible issues can be extrapolated to include; the experience of the researcher placing the landmark and analyst inconsistencies throughout measuring sessions [62–69]; the natural variability and preservation rate of the landmark itself [62, 70–73]; the landmark's definition and how easy it is to pinpoint [60–62, 67, 70]; and finally the precision of the measuring device being used [67–69].

Similarly, variability according to inter- and intra-observer error additionally has a significant effect on the comparison of information produced in different studies [67]. The ability to share data is a fundamental component of modern day research [74], especially in areas where accessibility to certain types of data is costly. This is a considerable issue when considering data derived from carnivore feeding, where in many cases large sample sizes are hard to obtain or access to wild animals may not be easy for researchers who reside on different continents [75]. Moreover, special care must be taken when working with animals, considering the conditions required to ensure correct welfare practices [76]. Under this premise, the ability to share data must be considered fundamental in order for research to advance.

While some initial approximations have been made to assess the reliability of geometric morphometric models in carnivore BSM analyses [55: pg. 97, 57: S2 Appendix], an in-depth study is yet to be carried out. For example, the current model for tooth pit analyses is known to produce highly complex dataset of high dimensionality. One particular issue with these datasets also includes the low proportion of cumulative variability represented by each variable [59]. Moreover, tooth pit models are based primarily on Type III landmarks, some of which are questionably much harder to locate than others. To ensure the reliability of this methodological approach, and where possible improve its precision, a detailed assessment is necessary.

The present study thus demonstrates how a detailed revision and update of Aramendi et al. [55]'s original model can facilitate future analyses of carnivore tooth pits with greater statistical reliability and accuracy. This in turn can be considered an important obstacle to supersede in order to make carnivore tooth mark data readily available to other researchers. Such advances are of great value to the study of carnivore BSMs; applicable to archaeological, palaeontological and modern day ecological research.

## 2. Methods

### 2.1. Tooth mark samples and digitisation

A total of 60 carnivore tooth pits were studied and compared for the purpose of this study. All marks were collected on the long bone diaphyses of large sized animals, including tibiae and radii. Diaphyses were chosen considering their greater density than epiphyses and are thus more likely to survive during carnivore feeding. Samples were produced in a controlled setting and include tooth marks produced by *Canis lupus signatus* wolves (pits = 30) and an Irish Setter gundog (pits = 30). Tooth mark samples from both carnivores were obtained in a controlled setting during the feeding of these animals. For the case of wolves, these samples were obtained from semi-captive animals (>1000m$^2$) from the Cabárceno natural park (Cantabria, Spain, http://www.parquedecabarceno.com) following standard feeding protocol established by the park. Wolves were left with the bones for 1 week before the samples were recollected by park employees. In the case of gundog tooth samples, bones were provided using product obtained from a local butcher. In this case, bones were left with the dogs for less than a day in an equally controlled environment. All samples were donated for use in the study, and no permits were required for their collection. Once collected, if necessary bones were cleaned in boiling water prior to analysis.

Each of these samples were originally studied by Yravedra et al. [58], identifying significant differences in tooth pit morphology for the identification of livestock predation in Europe. These bone surface modifications can thus be used as a control sample considering how morphological variations are already known in detail.

Digital reconstructions (Fig 1) of each mark were performed using the DAVID SLS-2 Structured Light Surface Scanner located at the TIDOP Research Group of the Higher Polytechnic School of Ávila (University of Salamanca, Spain). Equipment was calibrated using a 15 mm marker board while both the projector and the camera were equipped with additional macro-lenses (x2 to x10) for optimal resolution at microscopic scales [77]. Digitisations using this equipment are carried out in ca. 1 minute producing a point cloud density of up to 1.2 million points.

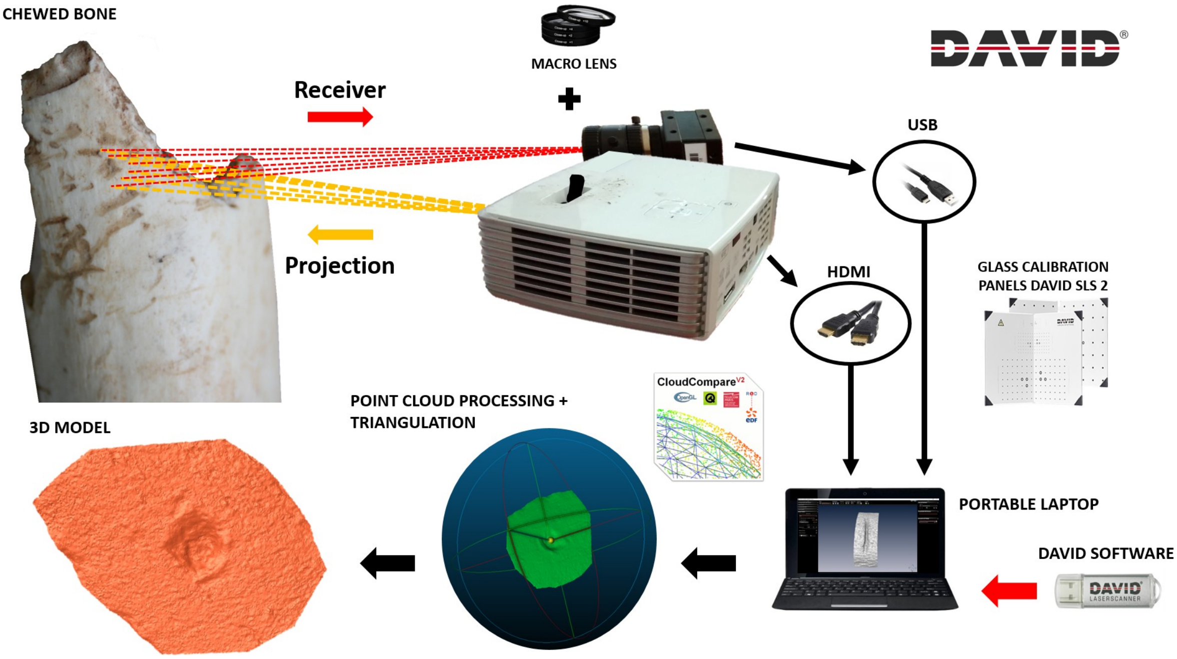


**Fig 1. A graphical description of the digital reconstruction protocol used in the present study; producing 3D models of carnivore tooth marks on bone according to Maté-González et al. [77].**

https://doi.org/10.1371/journal.pone.0240328.g001

## 2.2. Geometric morphometric data collection

For the purpose of this study, landmark data collection was performed by three separate individuals (A1-3) with varying degrees of experience. The first analyst (A1) is an experienced taphonomist, familiar with the different types of bone surface modifications produced by carnivores, yet has no prior experience working with geometric morphometric data. This analyst was chosen as a true analogy of a researcher in taphonomy who may be interested in applying GMM to their own analyses. The second analyst (A2), on the other hand, is a specialist in remote sensing and the handling of three-dimensional data, having worked in the past with geometric morphometric data. Finally, the third analyst (A3) is both a specialist in taphonomy and has the greatest extent of experience working with geometric morphometric data.

The main landmark configuration used for this study was that proposed by Aramendi et al. [55], consisting in a mixture of 17 fixed Type II and Type III 3D landmark points on the exterior and interior surfaces of each pit (Fig 2A). Each of the analysts performed collection of landmark data in separate locations, all using the free Landmark Editor software v. 3.0.0.6 [78]. Before data collection began, each of the participants were provided with the original paper [55] describing the landmark configuration to be used, alongside some general instructions on how to use the software. Throughout the process, none of the participants were allowed contact during the procedure so as to avoid influencing their placement of points.

To ensure an identical orientation of the pit by each observer, thus facilitating any consequent analyses, the only additional instruction provided was on how to identify the position of landmark n°1. Landmark n°1 was thus defined as the point along the axis marking the maximum length furthest away from the perpendicular axis marking the maximum width.

Once 3D landmark data had been collected by each of the analysts, coordinates were formatted and imported into R v. 3.5.1. (https://www.r-project.org/) for further statistical processing.

## 2.3. Inter- and intra-observer analyses

As described by Cramon-Taubadel et al. [65], four main approaches exist for the comparison of inter and intra-observer errors. The first, and arguably most common approach, consists in projecting repeated measurements onto a common coordinate system [79]. This is most commonly performed through Generalised Procrustes Analyses (GPA) [41, 79, 80], consisting in a series of superimposition procedures, including scaling, rotation and translation. If after GPA the measured landmarks fall into a suitable range of variation, then the configuration can be considered reliable. The second method employs Euclidean distances between repeat measurements and the centroid of the configuration to assess the relative repeatability of each landmark [38]. The third approach consists in the repeated digitisation of a specimen that is held in a constant orientation, thereby directly assessing the error of a landmark coordinate through variation in landmark location [81]. The final approach consists in the partial superimposition of landmark configurations so as to reduce the impact of the "*Pinocchio effect*" [82], a phenomenon described by the distortion of measurements influenced by scaling procedures in GPA from the first and second approaches [65].

For the purpose of this study, a mixture of these approaches were employed on the analysis of observers studying different carnivore tooth pits. For the first part of these analyses, Euclidean distance calculations between points placed by each analyst were performed prior to any superimposition procedures. This was carried out on the raw coordinates obtained directly from the 3D models, providing a metrically accurate true approximation in millimetres to the margin of error of each landmark [81]. Additional calculations for inter-analyst variability in raw feature space was performed by calculating the Euclidean distance between the point placed by each analyst and the centroid between all three analysts [69].

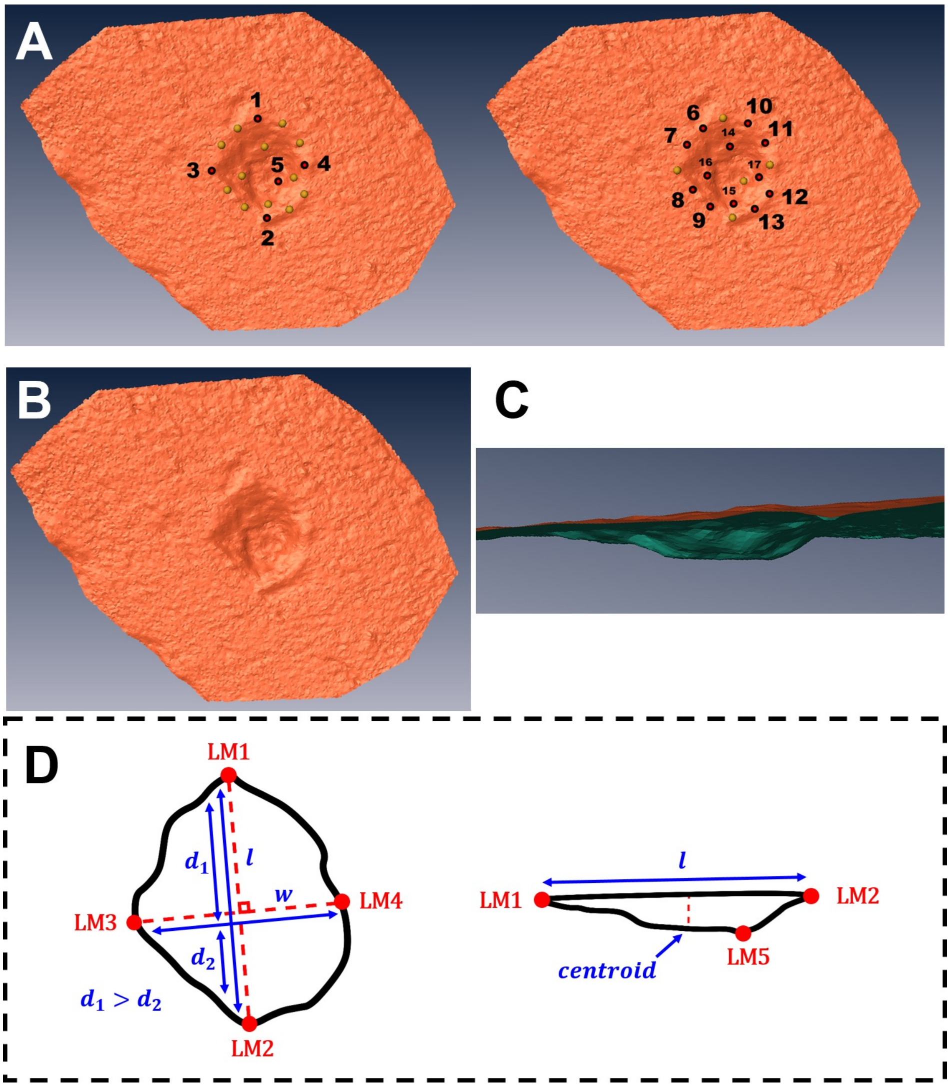


**Fig 2. A detailed graphical description of the landmark models employed within this study.** (A) The original 17-Landmark model [51], placed on the pit presented in (B) and (C). (D) The detailed description of the precise locations of landmarks 1 to 5 in the revised version of this model.

https://doi.org/10.1371/journal.pone.0240328.g002

This was followed by the full superimposition of landmarks through GPA (*shape space*) and the calculation of mean reference configurations for each of the analysts separately, the carnivore samples, and for the entire sample as a whole. Principal Components Analyses (PCA) were then performed to examine the distribution of specimens in morphospace. If error among analysts is low, PCA should present a tight clustering of individuals regardless of the analyst recording the data. From the derived PC scores, Multivariate Analyses of Variance (MANOVA) tests were performed to statistically assess these variances. For MANOVA, either the "Hotelling-Lawley" or "Wilk's Lambda" formulae were used for homogenously and inhomogeneously distributed coordinates respectively. For homogeneity testing, both numerical tests and Quantile-Quantile (Q-Q) plots were used on landmark point clouds, PCA results and distance calculations to avoid issues induced by sample size [83].

From the fully imposed feature space, Procrustes distances were also calculated between a reference shape and the configuration of landmarks placed by each of the analysts. Here the mean configuration of all individuals was used as the reference shape. From these Procrustes distances, further statistical tests and metrics were used to evaluate the amount of error and variability produced. While most traditional studies of intra and inter-observer measurement errors represent within-group variance as an estimation of the within-observer mean squares or mean distances [e.g. 63, 66, 67], in this study robust statistical metrics were employed [83], using the Biweight Midvariance (BWMV) (Eqs 2–4), calculated via the median of errors and their consequent Median Absolute Deviation (MAD) (Eq 1).

$$MAD = m(|x_i - m_x|) \tag{1}$$

$$BWMV = \frac{n\sum_{i=1}^{n} a_i(x_i - m)^2(1 - U_i^2)^4}{\left(\sum_{i=1}^{n} a_i(1 - U_i^2)(1 - 5U_i^2)\right)^2} \tag{2}$$

$$a_i = \begin{cases} 1, & if|U_i| < 1 \\ 0, & if|U_i| \geq 1 \end{cases} \tag{3}$$

$$U = \frac{x_i - m}{9MAD} \tag{4}$$

MAD is a calculation of the median ($m$) of the absolute deviations from the data's median ($m_x$) which is normally reported as a normalized value (NMAD). Normalisation of MAD is calculated through multiplying by the constant 1.4826. This is carried out to provide a scale factor of the true domain's standard deviation, using a 0.75 percentile as reference (Eq 5). Under this premise, MAD values are scaled by a factor of (Eq 5):

$$\frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} \approx 1.4826 \tag{5}$$

Where $\Phi^{-1}$ is the inverse cumulative distribution function of a standard Gaussian distribution [84].

NMAD has the distinct advantage of being a more robust measurement that is resilient to outliers in a dataset [83]. This is particularly useful for the analysis of datasets that do not follow a Gaussian distribution. The BWMV (reported as its square-root) is a further non-parametric measurement that additionally presents robustness of efficiency, proving a valuable substitute in non-parametric calculations [85]. NMAD and the square-root of the BWMV have proven to be important variables in accuracy control and system assessment analysis in remote sensing applications [83, 85–89]. Because of this, these metrics were employed as a substitute for standard deviation error calculations where Gaussian distributions were not detected, while the central tendency is reported as the median. Nevertheless, where deemed necessary by the normality of the dataset's distribution, further mean and standard deviation calculations were also reported.

Likewise, while most traditional analyses of observer measurement errors employ the use of Type II Analysis of Variance (ANOVA) calculations or Tukey's pairwise post-hoc comparisons [63, 64, 66–68, 90, 91, *inter alia*], these tests are statistically and mathematically conditioned by the underlying assumption of normality in the data's distribution [92, 93]. In cases where distances and errors do not fit a Gaussian distribution, parametric ANOVA tests are unlikely to truly capture the true nature of variances in sample distributions [94]. Because of this, when non-Gaussian distributions were detected, the present study employed the use of an adaptation of ANOVA known as a non-parametric Kruskal-Wallis robust statistical test to assess differences in distribution medians [95; pg. 585, 587 & 598]. This test employs a rank-based H statistic, similar to the more traditional $x^2$ or F statistic, that is able to "detect the kinds of difference of real interest" while making "only general assumptions. . . about the kind of distributions from which the observations come" [95; pg. 585].

Further testing considered the use of *Repeatability Measures* (RM); a calculation of the proportion of variance due to true variation among individuals [63, 66]. RMs can be defined as calculations involving the sum of squared distances ($d^2$) through (Eq 6):

$$RM = \frac{\left(\frac{1}{n-I}\sum_{j=1}^{J} d_j^2\right)}{\left(\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{J} d_{ij}^2\right)} \tag{6}$$

where $n$ is the total number of individuals in all of the samples, $I$ is the total number of samples and $J$ is the number of individuals in each of the samples. $d$ values used to calculate this metric were the Procrustes distance. The RM metric is reported as a value ranging from 0 to 1; the latter indicating variance in measurements being attributable to the analyst under study ($A_j$), rather than the true variation within the dataset [63].

From Procrustes distance calculations, further analyses included the generation of Unweighted Pair Group Method with Arithmetic Mean (UPGMA) trees. UPGMA results were analysed according to tree topologies, testing to see whether tooth pits clustered together according to the animal producing the tooth mark or the observer recording landmark data. Neighbour Joining (NJ) tree algorithms were also contemplated, nevertheless in this case NJ proved harder to read and was therefore discarded

Following this, two unsupervised Machine Learning algorithms were trained on fully superimposed coordinate data to test and see whether landmarks could be clearly defined through pattern recognition techniques. These analyses were performed in Python v.3.7 (http://www.python.org), using the SciKit-Learn package. Python applications were additionally implemented into the rest of the R workflow using the reticulate package and the RStudio development environment (https://rstudio.com/).

The first algorithm used was a Density-Based Spatial Clustering Algorithm with Noise (DBSCAN). DBSCAN employs a series of non-parametric mathematical and logical theories that differentiate between points that can be classed as reachable from its clustering core, and those that are considered as outliers or *noise points* ([96]: $\{P\epsilon D|\forall|i: p \notin C_i$ where point $p$ in dataset $D$ is not associable to any of the clusters $C$). DBSCAN requires the definition of 2 hyperparameters; the ε value defining the neighbourhood of a point and the minimum number of points (MinPts) that are required to form a cluster (i.e. the density of points). MinPts was set to 30, considering each of the analysts were required to digitise 30 individuals per sample for this study. $\varepsilon$ values were established in accordance with both the corresponding MinPts value, as well as the nature of the dataset's distribution. This value can be objectively defined mathematically using k-distance graphs according to the nearest neighbour [96], while the "elbow-joint" method can be used to establish the optimal $\varepsilon$ value for a dataset through calculating the point of maximum curvature in the plotted $k$-distance graph. The final $\varepsilon$ value for this study was thus calculated at 0.04 using the "kneed" algorithm in Python [97].

The second pattern recognition algorithm used was the non-parametric Mean-Shift algorithm (MS). MS, much like DBSCAN, works through calculating areas of high point density with the use of a kernel function, also allowing for the mathematical detection of outliers [98–100]. MS only requires the definition a single hyperparameter, known as the bandwidth, which is used to define the size scale of the algorithm's internal kernel function. In this study, the bandwidth was established using Scikit-Learn's bandwidth estimation functions, employing a calculated quantile value of 0.05 to fine tune this estimation.

The objectives of using pattern recognition algorithms consisted in defining the degree of separation between landmarks. Landmarks that can be clearly defined by all three analysts should theoretically appear clearly differentiable by both algorithms with as little noise as possible. The results produced by these algorithms were evaluated by considering the number of identified clusters, the number of noise points, and the number of points that were correctly or incorrectly grouped with other points within each cluster.

The final part of this analysis performed calculations on partially superimposed landmarks, carried out by removing the scaling procedure from GPA (*form space*). This ensures that each of the configurations are centred and superimposed without creating statistical noise from the scaling process and thus factoring in the possible distortion that could be created by the "*Pinocchio effect*" in GMM [65, 82]. On partially superimposed data, PCA and Procrustes distance calculations were performed. From these distances, UPGMA trees were additionally calculated.

### 2.4. Semilandmark models

Once the accuracy and reliability of each landmark had been assessed, trials were performed removing landmarks most prone to inter and intra-analyst error and replacing them with semilandmarks (S1 Appendix) [35, 73, 101]. Semilandmarks are computed equidistant points that slide along the surface of the 3D model [73]. This overcomes many issues in accuracy when dealing with features that can only be represented by Type III fixed landmarks. Considering these landmarks are most prone to error and often incorrectly placed [61], substituting these points with semilandmarks presents a more efficient means of objectively quantifying complex morphological features such as curves and irregular surfaces [68].

For the process of placement of these semilandmarks, patches were drawn over each pit in the Landmark Editor v.3.0.0.6 software and different numbers of semilandmarks were computed. Trials were performed to find the optimal number of semilandmarks for tooth pit analysis, including 5x5, 6x6, 7x7, 8x8, 9x9 and 10x10 configurations.

The final part of this study consisted in both a full and partial GPA of the final landmark data to assess the variances in shape and form space captured by the newly updated landmark configurations. PCA were then performed to represent these differences graphically and to reduce the dimensionality of the dataset for further multivariate statistical processing. Comparison of PCA results reported the dimensionality of the consequent feature space ($\mathbb{R}^n$), the proportion of variance represented by the first 2 PC scores as well as the cumulative proportion of variance represented by the first 10 PC scores. These were followed by MANOVA calculations to assess the degree of separation between wolf and dog samples according to the landmark configuration used. Finally, a Canonical Variance Analysis (CVA) was performed, alongside calculations of Mahalanobis and Procrustes distances and permuted (n = 999) *p*-values to evaluate the degree of separation between the samples.

## 3. Results

### 3.1. Inter- and intra-observer errors

Throughout the study, clear patterns emerged indicating those landmarks that were easier to locate and those that were more difficult. In all three trials, A1-3 noted difficulties on locating

Landmarks (LM) 14 through to 17, while LM1-5 were considered the most straightforward. A1-3 also noted that LM6-13 were easy to locate once LM1-4 had been clearly established. Nevertheless, *in-visu* inspection of each of the measurements taken on the 3D models documented that some analysts were more meticulous upon placing these points than others. Interestingly, A1 and A2 noted that wolf tooth marks were harder to process, observing that the 3D models obtained from this carnivore presented weaker topographical variations and were thus harder to analyse. Dog tooth marks on the other hand were noted as much deeper, easier to recognise and therefore digitisation procedures were easier overall. While the variations between these marks may be due to a number of variables dependent on the samples used, these go beyond the scope of the present study and are pending further future research.

**3.1.1. Raw measurement calculations.** Raw errors obtained directly from the 3D models indicate an error range of 1.95 mm across the entire data set (Table 1: Min error documented = 0.095 mm, Max = 2.045 mm). Mean and Median errors for all individuals were reported at 0.343 mm and 0.303 mm respectively, with a NMAD of 0.119 mm and square root of the BWMV at 0.121 mm. In all cases, no significant differences were detected when comparing the animal being processed by each analyst (Fig 3: Chi$^2$ = 0.027, $p$ = 0.8682). Nevertheless, in accordance with the personal observations by A1 and A2, errors calculated on wolf tooth mark samples still present a slightly more varied distribution as opposed to dogs (Fig 3: Wolf Skewness = 4.28, Dog Skewness = 1.61). This indicates that wolves may be harder to process.

When separating samples according to the observer (Fig 3), a tendency can be observed for larger errors to be recorded when comparisons include the least experienced individuals, nevertheless, kruskal-wallis testing indicates no significant differences (Chi$^2$ = 0.98, $p$ = 0.61).

Further experimentation with the number of landmarks included reveal that as LM14-17 are removed, error calculations drop to less than half (Table 1: Median = 0.146 mm, NMAD = 0.077 mm, BWMV = 0.070 mm). Likewise, when only considering LM1-5, errors documented on wolf and dog samples are considered the most equivalent (Chi$^2$ = 0.0018, $p$ = 0.9658). Moreover, when considering each of the LMs separately (Table 2), LM14 presents the highest recorded error among analysts (Median = 0.352 mm, NMAD = 0.292 mm, BWMV = 0.175 mm) while LM5 the lowest (Median = 0.175 mm, NMAD = 0.122 mm, BWMV = 0.075 mm).

When analysing inter-analyst variability according to each individual landmark (Fig 4, S1-S4 Tables of S1 File), errors begin to appear indicating a significant negative correlation between the experience of the analyst and the error produced (Kendall's τ = -0.03, p = 0.017), proving the error to be significantly lower in analysts of more experience working with GMM than those in the process of learning. Nevertheless, differences in inter-analyst error are only

**Table 1. Overall documented metric errors according to the landmarks included in the study and the animal being analysed.** BWMV values are reported as the square root of the BWMV. All measurements are recorded in mm. For frame of reference; the studied pits range from 1.176 to 4.448mm in total length and 0.8382 to 3.5128mm in total width.

| | | Min | Sample Mean | Standard Deviation | Median | NMAD | BWMV | Max |
|---|---|---|---|---|---|---|---|---|
| LM 1–17 | Absolute | **0.095** | 0.343 | 0.227 | **0.303** | 0.119 | 0.121 | **2.045** |
| | Wolf | **0.101** | 0.363 | 0.291 | **0.305** | 0.128 | 0.130 | **2.045** |
| | Dog | **0.095** | 0.322 | 0.134 | **0.303** | 0.109 | 0.113 | **0.966** |
| LM 1–13 | Absolute | **0.048** | 0.172 | 0.084 | **0.146** | 0.077 | 0.070 | **0.464** |
| | Wolf | **0.053** | 0.172 | 0.090 | **0.143** | 0.075 | 0.084 | **0.448** |
| | Dog | **0.048** | 0.171 | 0.078 | **0.158** | 0.067 | 0.070 | **0.464** |
| LM 1–5 | Absolute | **0.002** | 0.172 | 0.107 | **0.139** | 0.083 | 0.092 | **0.586** |
| | Wolf | **0.002** | 0.177 | 0.120 | **0.139** | 0.081 | 0.097 | **0.586** |
| | Dog | **0.009** | 0.166 | 0.094 | **0.140** | 0.085 | 0.087 | **0.509** |

https://doi.org/10.1371/journal.pone.0240328.t001

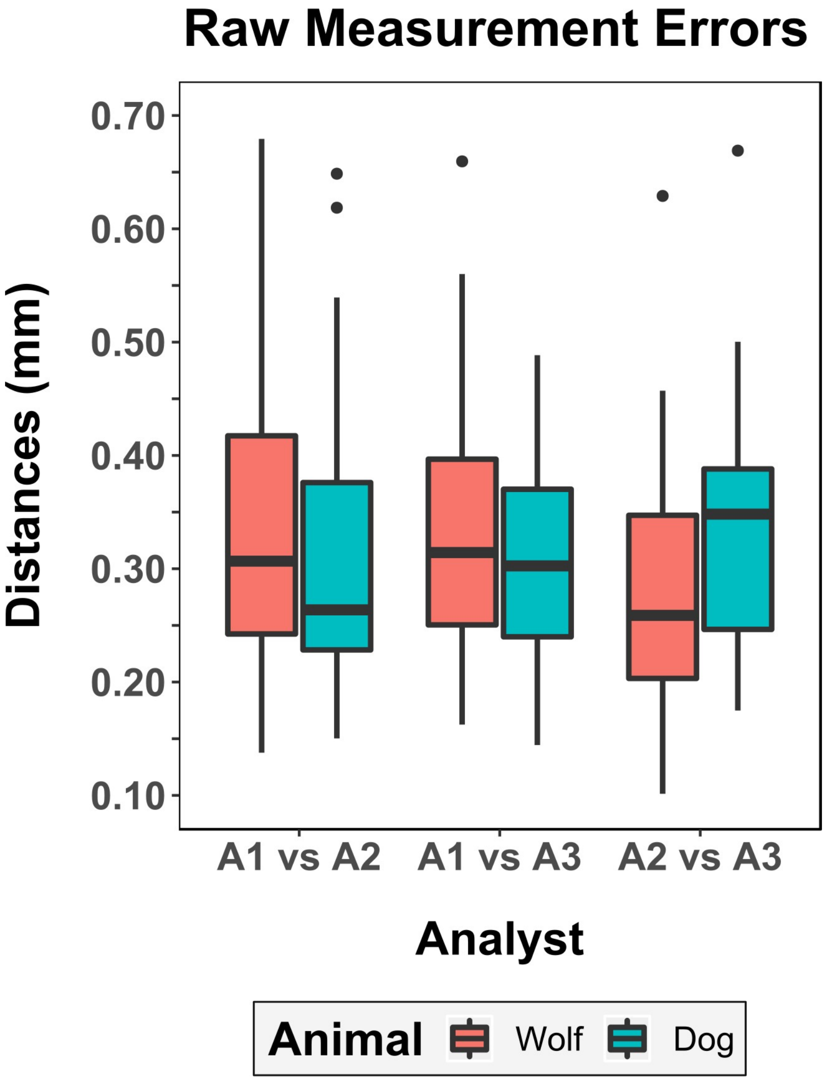


**Fig 3. Boxplot representing the distribution of metric errors according to the analyst placing landmarks and the animal being studied.**

https://doi.org/10.1371/journal.pone.0240328.g003

revealed to be significant when considering wolves separately (S3, S4 Tables of S1 File; Chi$^2$ = 11.6, $p$ = 0.003), while landmark placement appears much more consistent among analysts when analysing dogs (S3, S4 Tables of S1 File; Chi$^2$ = 0.36, $p$ = 0.8).

**3.1.2. Fully superimposed calculations.** Fully superimposed coordinates present highly non-Gaussian spatial distributions on all accounts (Shapiro $w > 0.78$, $p < 2.2\text{e-}16$), while distribution of landmark centroids reveal interesting differences among the distribution of landmarks according to the analyst (Figs 5 and 6). In most cases, the greatest differences can be detected for LM14-17. Moreover, LM14-17 can clearly be seen to present the highest differences among analysts by observing point cloud distributions (Fig 5) as well as differences in centroid location (Fig 6). Likewise, LM1 to LM5 appear to be the most consistent.

PCA results on superimposed coordinate data (Fig 7 and S1 Fig of S1 File) reveal an inhomogeneous (Shapiro $w$ = 0.99, $p$ = 0.0002) wide dispersal of points with no clustering occurring according to the pit being studied (S1 Fig of S1 File). Clear differences among the analysts

**Table 2. Metric errors according to the landmark being placed.** BWMV values are reported as the square root of the BWMV. All measurements are recorded in mm. For frame of reference; studied pits range from 1.176 to 4.448mm in total length and 0.8382 to 3.5128mm in total width.

| Landmark | Median | NMAD | BWMV |
|---|---|---|---|
| LM1 | 0.282 | 0.191 | 0.108 |
| LM2 | 0.279 | 0.216 | 0.121 |
| LM3 | 0.324 | 0.211 | 0.125 |
| LM4 | 0.342 | 0.229 | 0.137 |
| LM5 | 0.175 | 0.122 | 0.075 |
| LM6 | 0.279 | 0.162 | 0.103 |
| LM7 | 0.287 | 0.176 | 0.111 |
| LM8 | 0.315 | 0.216 | 0.123 |
| LM9 | 0.271 | 0.214 | 0.116 |
| LM10 | 0.321 | 0.233 | 0.121 |
| LM11 | 0.288 | 0.221 | 0.132 |
| LM12 | 0.339 | 0.209 | 0.135 |
| LM13 | 0.297 | 0.172 | 0.114 |
| LM14 | 0.352 | 0.292 | 0.175 |
| LM15 | 0.26 | 0.189 | 0.122 |
| LM16 | 0.278 | 0.203 | 0.123 |
| LM17 | 0.315 | 0.202 | 0.13 |

https://doi.org/10.1371/journal.pone.0240328.t002

can be observed with A1 and A2 occupying a vast area of the represented feature space while A3 presents the least amount of variation (Fig 7). MANOVA calculations obtained from this data indicate significant differences between observers ($p$ = 0.001). While significant differences are also detected between wolf and dog samples ($p$ = 0.002), the latter $p$ values are slightly

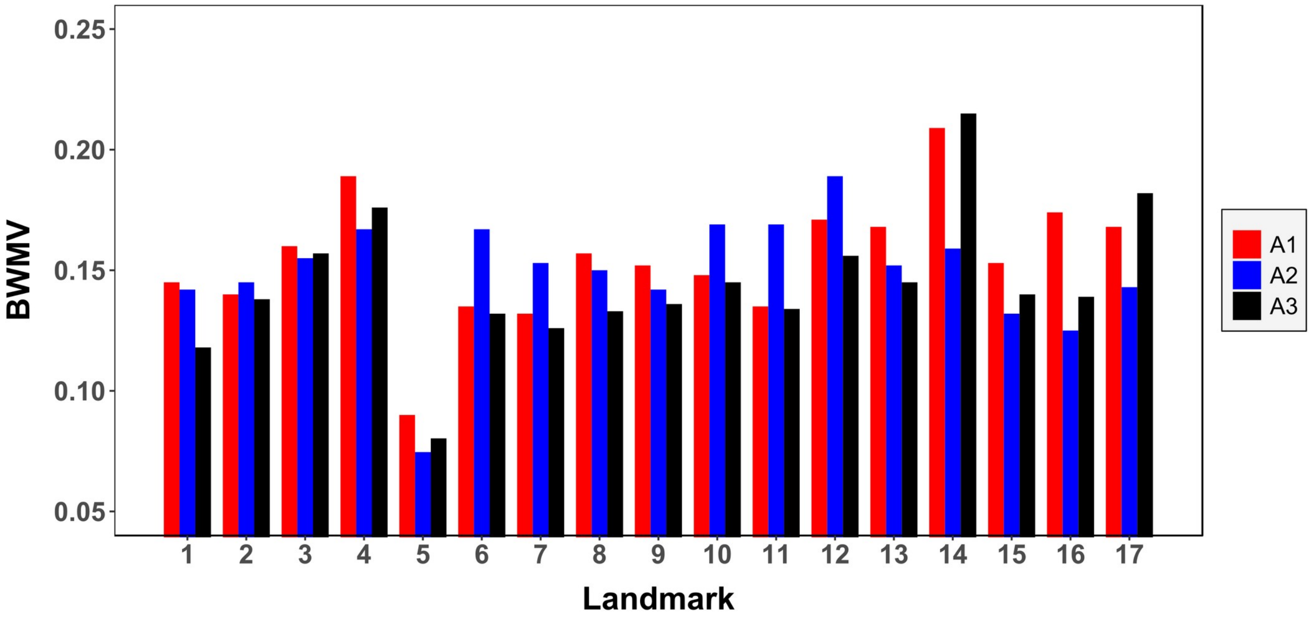


**Fig 4. Inter-analyst raw measurement errors for each of the landmarks.** Measurements represent the square root of the BWMV of Euclidean distances (mm) from each of the points placed by the analyst to the landmark's absolute centroid by all three analysts.

https://doi.org/10.1371/journal.pone.0240328.g004

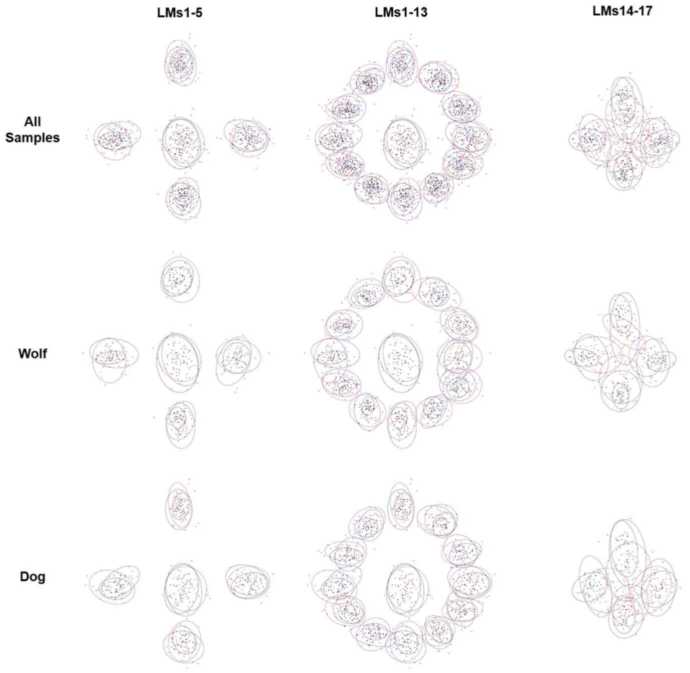


**Fig 5. Fully superimposed landmarks in shape space and 95% Confidence ellipses calculated for spatial distributions (A1 = Red, A2 = Blue, A3 = Black).**

https://doi.org/10.1371/journal.pone.0240328.g005

higher indicating the observer to have a stronger weight over variations detected in PCA than the carnivore.

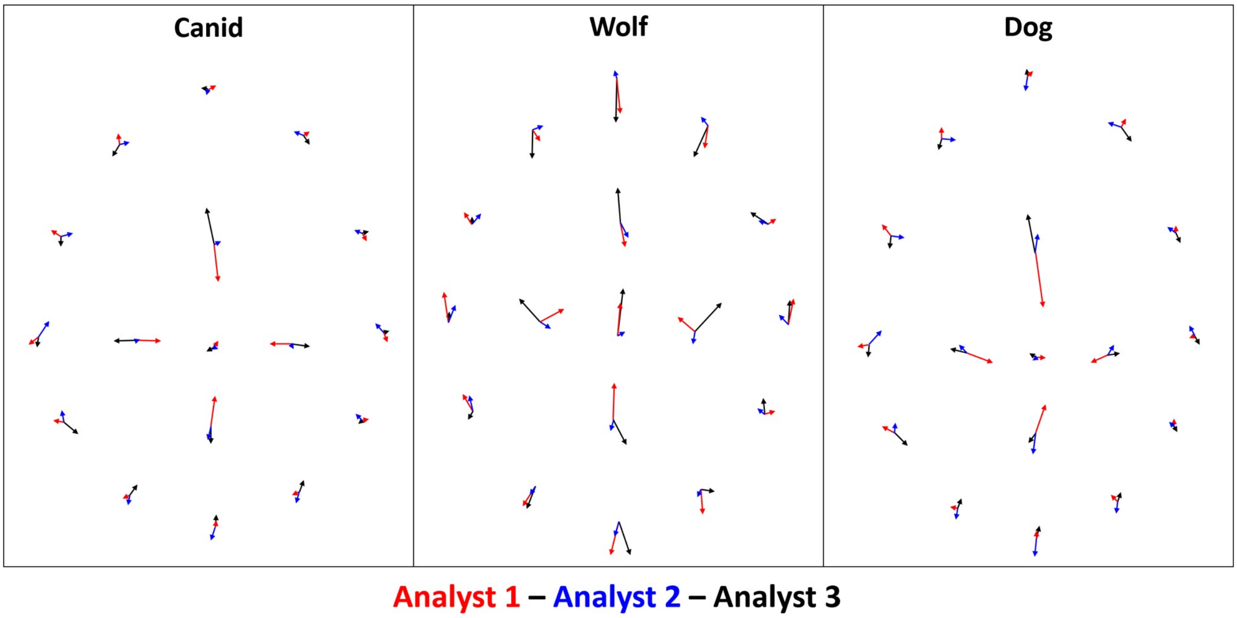


**Fig 6. Deviations of landmark centroids calculated from mean configurations in fully superimposed shape space.** The origin of each arrow marks the overall centroid for the corresponding landmark in the absolute mean configuration. The end of each arrow marks the position of the same landmark's centroid according to the analyst.

https://doi.org/10.1371/journal.pone.0240328.g006

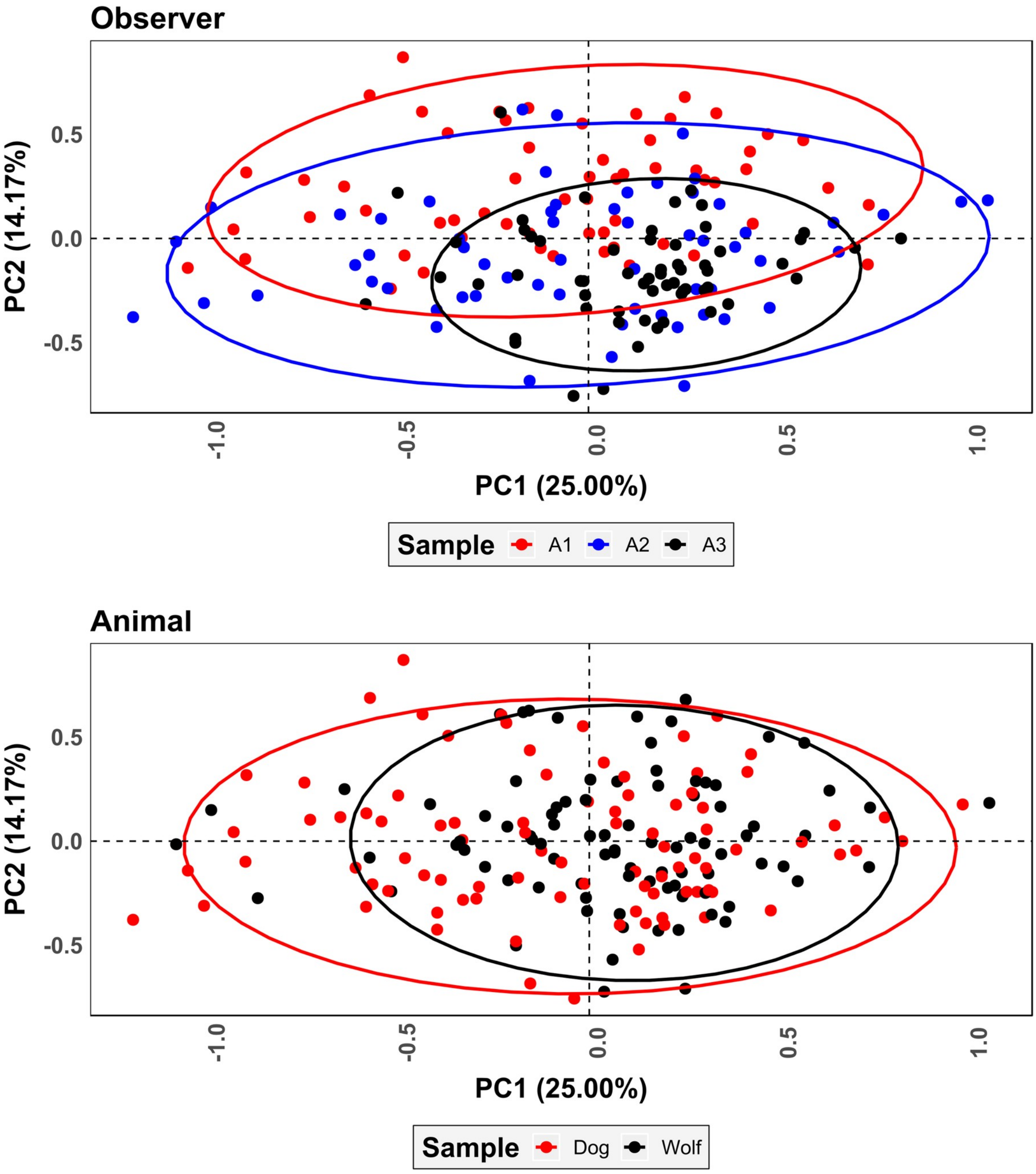


**Fig 7. Scatter plot of principal components analysis results for fully superimposed landmark coordinates.** Ellipses represent 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0240328.g007

Procrustes distance calculations (Table 3) reveal a mixture of significant and insignificant differences among the different analysts (Table 4), while comparisons with A1 and A3 show the greatest degrees of variation. These differences are equally evident when comparing carnivore samples separately. Overall distances highlight A3 to be the analyst closest to the absolute mean configuration presenting the least amount of intra-analyst variability (RM = 0.251). Nevertheless, this highlights how analysts with greater experience in GMM are significantly separable from those beginning (Table 3). RMs in general (A1 = 0.403, A2 = 0.357, A3 = 0.251)

**Table 3. Median, NMAD and square root of the BWMV values for Procrustes distances in fully superimposed shape space between each of the tooth pits processed by the analysts and the mean configuration.**

| | Analyst | Median | NMAD | BWMV |
|---|---|---|---|---|
| Overall | A1 | 0.230 | 0.071 | 0.062 |
| | A2 | 0.205 | 0.041 | 0.051 |
| | A3 | 0.165 | 0.028 | 0.038 |
| Wolf | A1 | 0.219 | 0.080 | 0.067 |
| | A2 | 0.197 | 0.047 | 0.057 |
| | A3 | 0.155 | 0.028 | 0.034 |
| Dog | A1 | 0.233 | 0.061 | 0.054 |
| | A2 | 0.208 | 0.044 | 0.052 |
| | A3 | 0.171 | 0.024 | 0.037 |

https://doi.org/10.1371/journal.pone.0240328.t003

indicate that a great degree of variation detected within the samples is due to the analyst rather than the true variation produced by the carnivores. This is especially worrying considering how this variation increases according to the experience of the analyst. UPGMA tree topologies additionally reflect this, with absolutely no grouping according to the carnivore under study (S2 Fig of S1 File), and a slight agglomeration of A3 individuals grouped towards the middle of the tree's branches.

Pattern recognition results classify a high number of points as noise (Fig 8 and S3 Fig & S5 Table of S1 File; DBSCAN = 74, MS = 250), with the least amount of noise detected when only considering LM1-5 (Fig 8E and 8F). Likewise, the highest misclassification rates appear when LM14-17 are included in the models (S5 Table of S1 File; DBSCAN = 2110, MS = 167). Visually, while MS is still able to clearly define the 17 landmarks (Fig 8B), misclassification and noise rates are considerably higher. DBSCAN on the other hand appears much more susceptible to confusion created by density, only clearly identifying points of interest when only LM1-5 are included (Fig 8E). In both cases, pattern recognition helps identify 5 clear points of interest that can be considered key landmarks (Fig 8E and 8F).

**3.1.3. Partially superimposed calculations.** Partially superimposed coordinates present highly non-Gaussian spatial distributions on all accounts (Shapiro $w > 0.94$, $p < 1.99\text{e-}10$). Procrustes distance calculations (Table 5) reveal insignificant differences among all three analysts (Table 6), while comparisons with A1 and A3 show the greatest degrees of variation. This is observed when comparing both carnivore samples together as well as separately (Table 6).

**Table 4. Kruskal-Wallis Chi$^2$ and p-values comparing Procrustes distances in fully superimposed shape space between each of the analysts and the overall mean configuration.** Calculations are provided for both carnivores as well as wolf and dog samples separately. Significant p-values under the standard alpha (α) threshold of 0.05 are marked in bold.

| | | Chi$^2$ | *p*-Value |
|---|---|---|---|
| Overall | A1 vs A2 | 1.56 | 0.21 |
| | A1 vs A3 | 20.23 | **6.9e-06** |
| | A2 vs A3 | 16.24 | **5.6e-05** |
| Wolf | A1 vs A2 | 0.42 | 0.52 |
| | A1 vs A3 | 9.00 | **0.0027** |
| | A2 vs A3 | 7.97 | **0.0047** |
| Dog | A1 vs A2 | 0.50 | 0.48 |
| | A1 vs A3 | 10.29 | **0.0013** |
| | A2 vs A3 | 7.89 | **0.0050** |

https://doi.org/10.1371/journal.pone.0240328.t004

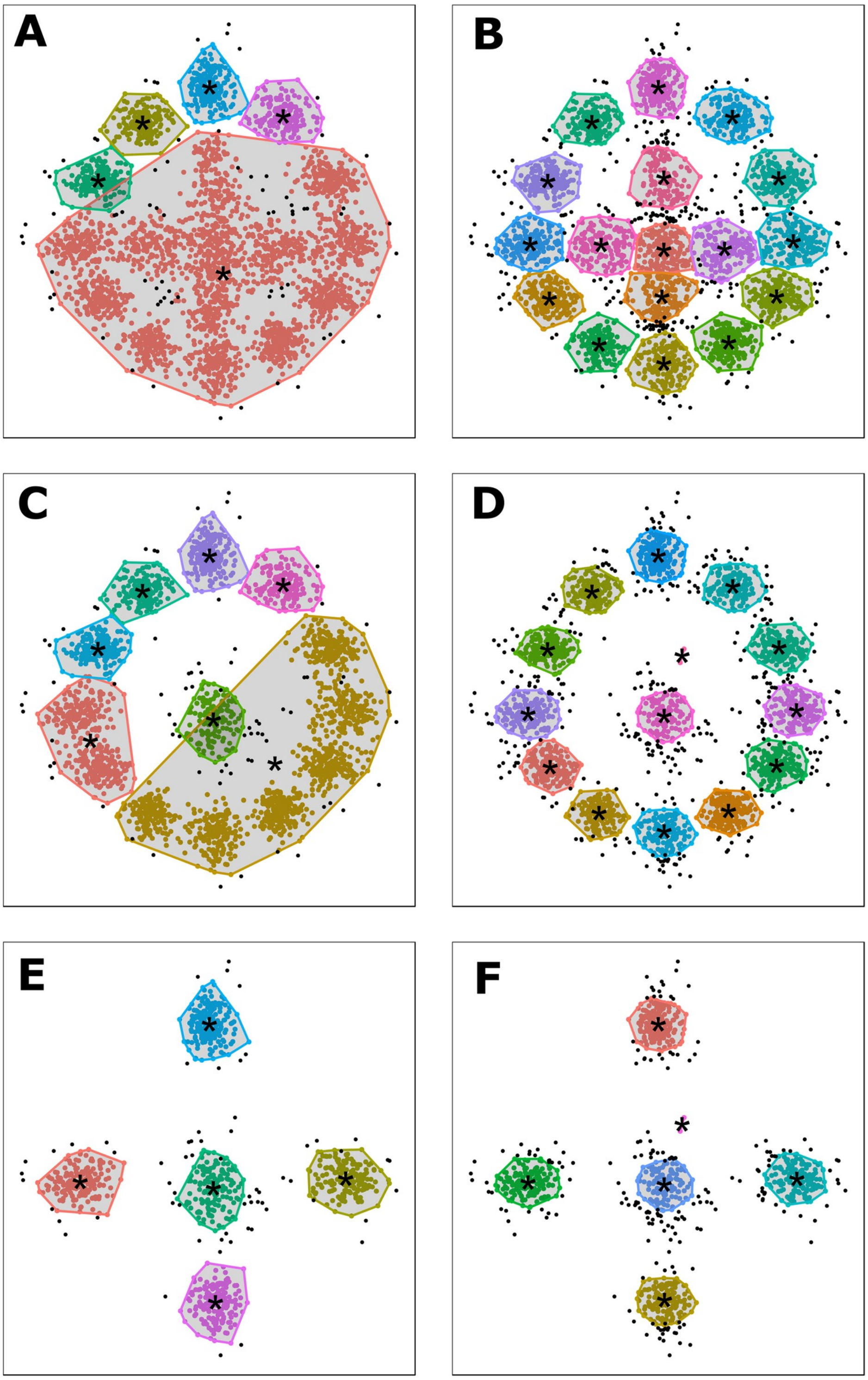


**Fig 8.** Unsupervised DBSCAN (A, C, E) and MS (B, D, F) pattern recognition results trained on fully superimposed landmark coordinates. Detected clusters are marked by their corresponding colours and calculated convex hulls. Points classified as noise are marked in black. The centroid or point of interest for each cluster is marked by a single black $^*$. (A, B) LM1-17; (C, D) LM1-13; (E, F) LM1-5.

https://doi.org/10.1371/journal.pone.0240328.g008

**Table 5. Median, NMAD and square root of the BWMV values for Procrustes distances in partially superimposed shape space between each of the tooth pits processed by the analysts and the mean configuration.**

| | Analyst | Median | NMAD | BWMV |
|---|---|---|---|---|
| Overall | A1 | 1.056 | 0.375 | 0.410 |
| | A2 | 0.977 | 0.394 | 0.420 |
| | A3 | 0.962 | 0.362 | 0.329 |
| Wolf | A1 | 0.976 | 0.450 | 0.460 |
| | A2 | 1.083 | 0.445 | 0.383 |
| | A3 | 0.962 | 0.417 | 0.335 |
| Dog | A1 | 1.040 | 0.318 | 0.375 |
| | A2 | 0.959 | 0.351 | 0.447 |
| | A3 | 0.946 | 0.382 | 0.383 |

https://doi.org/10.1371/journal.pone.0240328.t005

Considering each of the analyst's deviation from the mean reference configuration (Table 5), A3 is seen to present the smallest differences and the greatest consistency when placing landmarks (RM = 0.301). A2 (RM = 0.328) on the other hand is frequently seen separate while A1 presents the least amount of repeatability (RM = 0.383). Considering all three RM scores, a moderate degree of variation is detected within the sample due to the analyst rather than the true variation produced by the carnivore. UPGMA tree topologies additionally reflect this, with no grouping according to the carnivore under study (S4 Fig of S1 File).

PCA results on partially superimposed coordinate data (Fig 9 and S1 Fig of S1 File) yet again reveal an inhomogeneous (Shapiro $w = 0.90$, $p = 2.2e\text{-}16$) distribution of points. As opposed to fully superimposed data (Fig 7), however, clustering according to the pit being studied is stronger (S1 Fig of S1 File). At no point, however, is the grouping of points as strong as should be expected, with great variation within the represented feature space being product of the analyst. Once again A3 occupies a much more restricted proportion of feature space, however the difference between distributions is not as extreme. MANOVA calculations indicate equally significant differences between observers as in fully superimposed shape space ($p = 0.001$), however differences between animal samples remain significant regardless of the observer ($p = 0.001$).

## 3.2. Overall evaluation of the 17-landmark model

Comparison of results obtained through raw coordinate data as well as fully and partially superimposed landmarks are sufficient in drawing two main conclusions about the 17

**Table 6. Median, NMAD and square root of the BWMV values for Procrustes distances in partially superimposed shape space between each of the tooth pits processed by the analysts and the mean configuration.**

| | Analyst | Median | NMAD | BWMV |
|---|---|---|---|---|
| Overall | A1 | 1.056 | 0.375 | 0.410 |
| | A2 | 0.977 | 0.394 | 0.420 |
| | A3 | 0.962 | 0.362 | 0.329 |
| Wolf | A1 | 0.976 | 0.450 | 0.460 |
| | A2 | 1.083 | 0.445 | 0.383 |
| | A3 | 0.962 | 0.417 | 0.335 |
| Dog | A1 | 1.040 | 0.318 | 0.375 |
| | A2 | 0.959 | 0.351 | 0.447 |
| | A3 | 0.946 | 0.382 | 0.383 |

https://doi.org/10.1371/journal.pone.0240328.t006

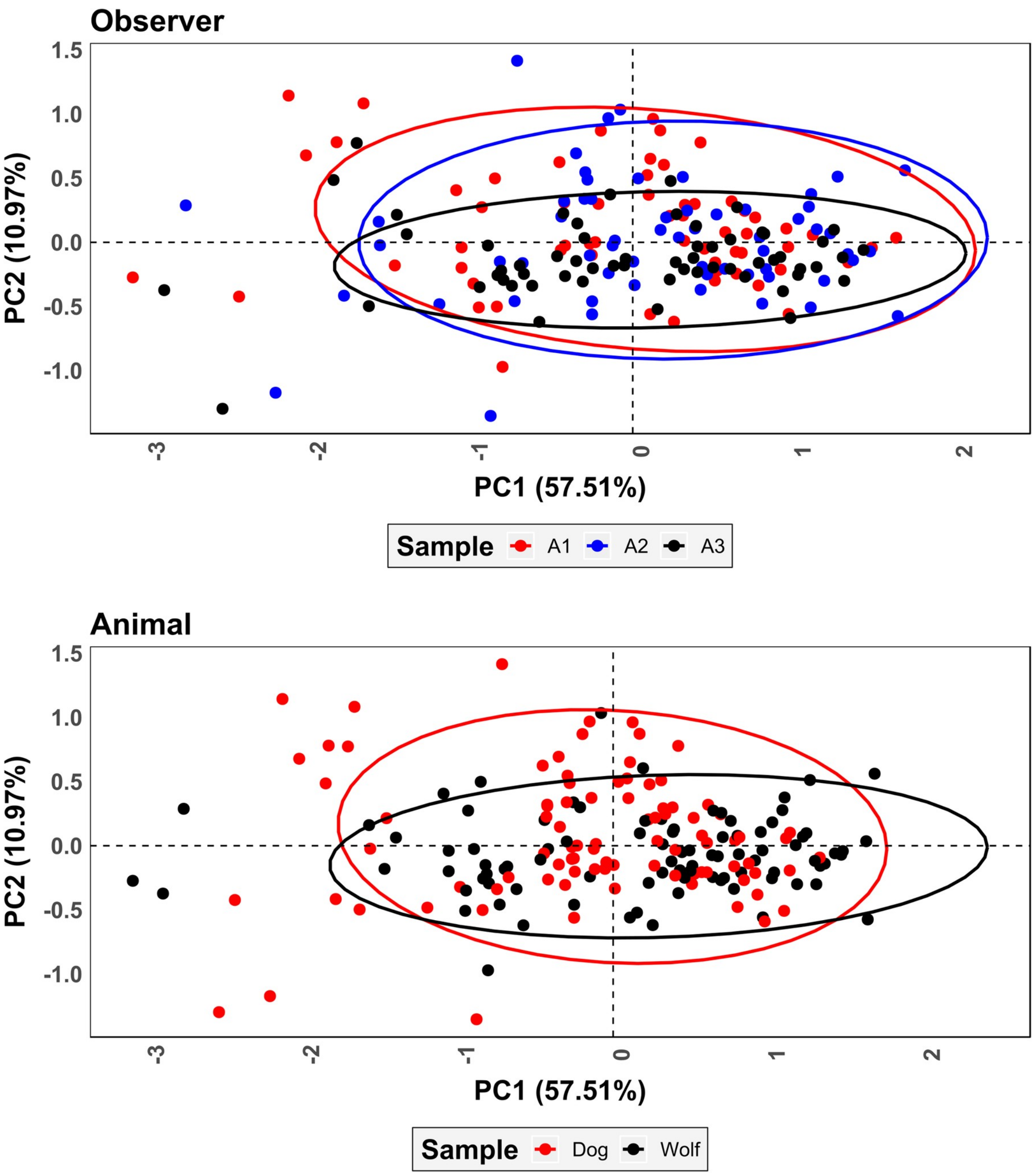


**Fig 9. Scatter plot of principal components analysis results for partially superimposed landmark coordinates.** Ellipses represent 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0240328.g009

landmark model proposed by Aramendi et al. [55]; (1) experience in GMM is a highly conditioning factor in the amount of error produced within the model and (2) error is observed to increase upon the inclusion of LM14-17.

While the observed error does not necessarily correlate with the type of landmark being used (Kendall's $\tau$ = -0.26, $p$ = 0.14), Type III landmarks such as LM6-13 tend to present lower degrees of error because their location is influenced by their spatial relationship with Type II LM1-4. If we were to then exclude LM6-13, the weight of this correlation substantially

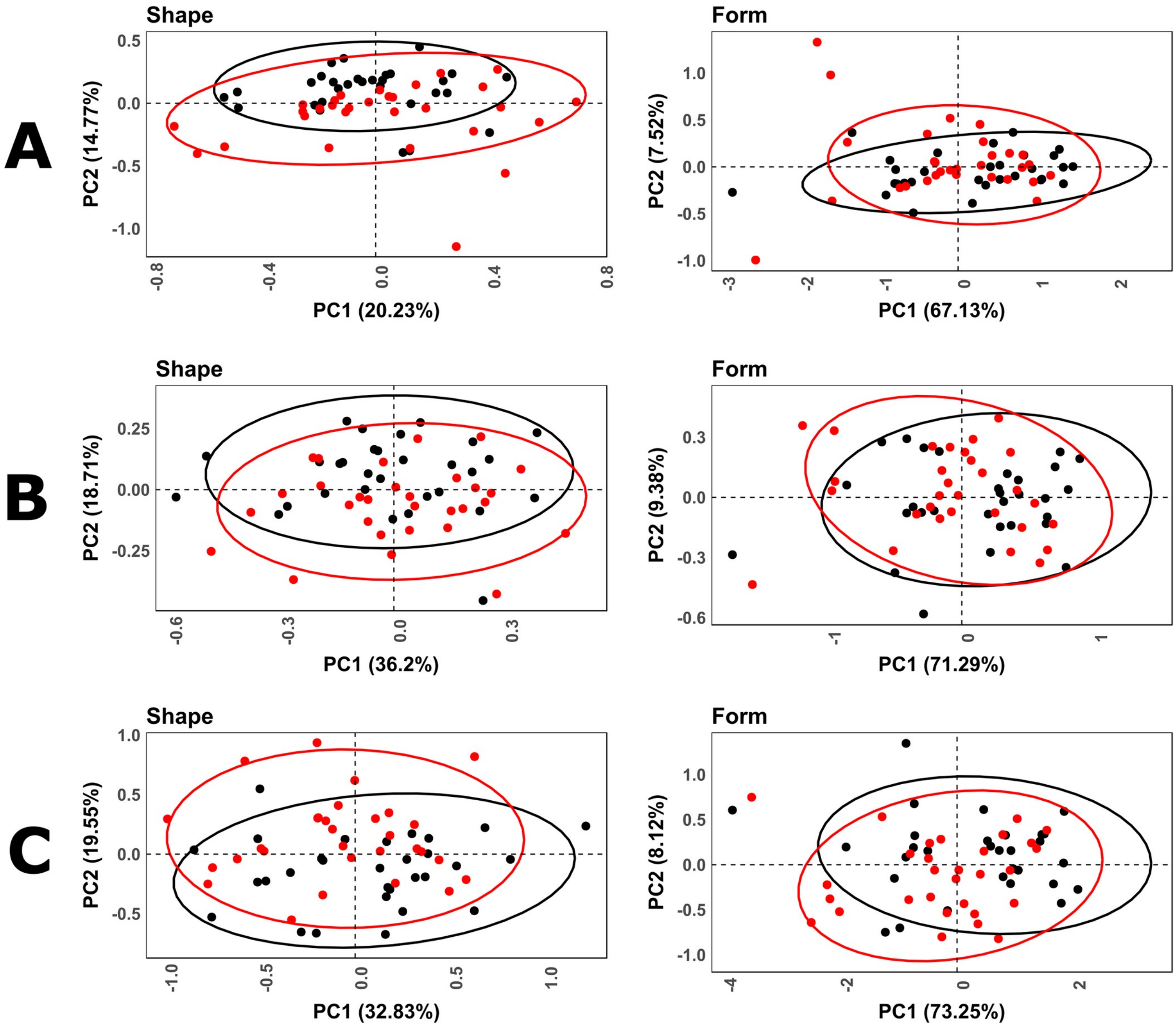


**Fig 10. Scatter plots for principal components analysis results comparing different landmark configurations for (Red) dog and (Black) wolf tooth pits.** (A) 17-Landmark model in (Left) shape and (Right) form. (B) 5-Landmark model in (Left) shape and (Right) form. (C) 30-landmark model in (Left) shape and (Right) form. Ellipses represent 95% confidence intervals.

https://doi.org/10.1371/journal.pone.0240328.g010

increases ($\tau = 0.45$, $p = 0.09$), arguing LM14-17 to be the greatest source of error. This logically agrees with observations made by multiple authors on the value and accuracy of Type III as opposed to Type II landmarks [60, 61, 72].

From a different perspective, while the relationship between experience and repeatability is relatively clear and logical to assume [62–68], upon analysing RMs alone no reliable $p$ values can be used to quantify this correlation for this case study ($\tau = -1$). When including the metric errors and overall distances between configurations, however, correlations can be successfully calculated supporting the aforementioned conclusions (Kendall's $\tau = -0.03$, p = 0.017).

Further in depth analysis of overall results presents A3 to be the most consistent in placing landmarks regardless of the sample. Nevertheless, other underlying factors may be

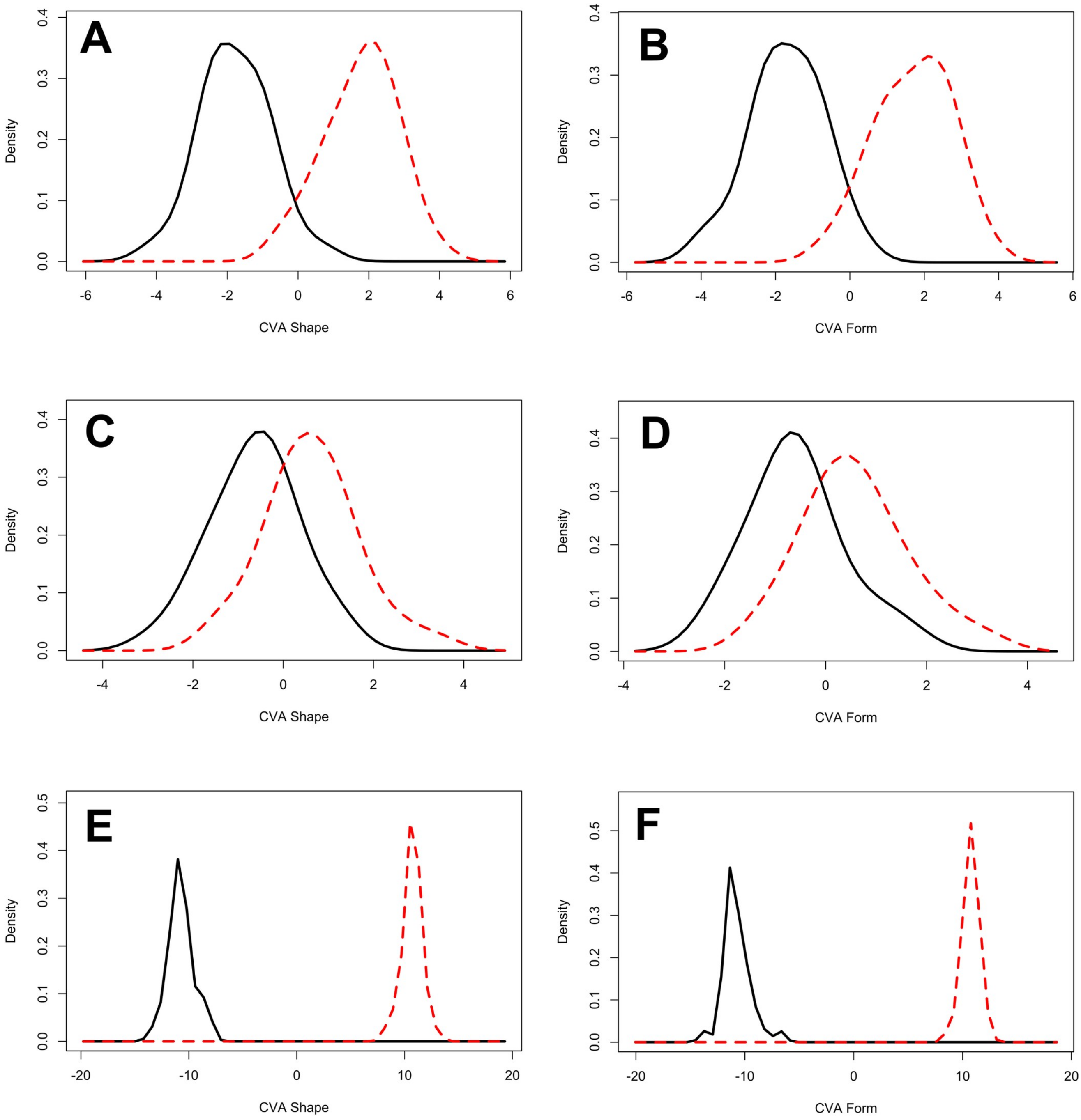


**Fig 11.** Distribution density plots for Canonical Variance Analyses in (A, C, E) shape and (B, D, F) form comparing (Red) dog and (Black) wolf tooth pits. (A, B) 17-landmark model. (C, D) 5-Landmark model. (E, F) 30-Landmark model.

https://doi.org/10.1371/journal.pone.0240328.g011

conditioning some of the results that are harder to quantify. For example, A1 and A2 took little time when placing landmarks (ca. < 1 min), taking a total of approximately 3 hours to process the entire sample. A3 on the other hand took longer processing each pit (ca. > 1 min), and took over twice as long to process the entire sample (approximately 6.5 hours). While the time taken to process a model might not be the most important variable, here it can be argued that elements of an observer's personality are likely to take some effect in the results, as more meticulous individuals can be seen to take longer and may handle the sample differently. These observations, however, are more subjective and harder to model, stressing the need for analysts to take exceptional care when processing each individual pit so as to avoid these types of error.

Observations regarding the "*Pinocchio Effect*" present notable differences between fully superimposed, partially superimposed and raw data. With regards to geometric morphometric data, PCA appears to be the test where differences become most notable (S1 Fig of S1 File), with drastic changes to feature space appearing when including the scaling procedure in GPA.

Finally, considering the vast amount of data produced in this study, deductions can be made supporting the decision of removing LM14-17 from future analyses so as to reduce statistical errors. On the other hand, while LM6-13 do not produce significant errors in analyses, under the premise of reducing as much statistical noise as possible that may be dependent on the analyst, we have decided to remove all Type III LMs from the rest of this study as well. According to this decision, the definitive fixed Type II landmarks that will be used in our final model will be LM1-5.

## 3.3. Landmark model comparisons

PCAs (Fig 10A) obtained by the original 17-Landmark model present a non-polarized morphospace with relatively high overlapping of wolf and dog samples in both shape and form space. The defined feature space described by each of these PCAs presents a high number of dimensions in both shape ($\mathbb{R}^{44}$) and form ($\mathbb{R}^{51}$), representing low overall variability in the first 2 dimensions (shape = 35.39%, form = 74.65%). The first 10 dimensions represent a much higher cumulative proportion of variance (shape = 83.78%, form = 94.39%), with MANOVA results highlighting significant statistical differences between carnivore tooth samples in shape ($p = 0.001$), but not in form ($p = 0.059$). CVA results obtained from this model present a much clearer separation between samples (Fig 11A and 11B), with both significant Procrustes (D = 0.072, $p = 0.002$) as well as Mahalanobis distances (D = 3.55, $p < 0.0001$).

Removal of Type III landmarks with the 5-Landmark model presents a slight increase in overlapping in PCA (Fig 10B) of carnivore samples in a reduced number of dimensions for both shape ($\mathbb{R}^{8}$, PC1&2 = 55.50%) and form ($\mathbb{R}^{15}$, PC1&2 = 80.67%). CVA additionally presents a significant reduction in both Procrustes (D = 0.064, $p = 0.04$) and Mahalanobis (D = 1.26, $p = 0.009$) distances with high overlapping across graphs (Fig 11C and 11D). Likewise, MANOVA tests indicate differentiation between carnivores to be less clear when using only LM1-5 (shape $p = 0.029$, form $p = 0.039$).

Of the 6 semilandmark models tried and tested, immediate improvements in results were noted with the model employing a 5x5 net of semilandmarks with the 5 fixed Type II landmarks (30 in total, S1 Appendix). PCA results (Fig 10C) display similar graphical results to the original 17-landmark model, with a relatively highly overlapping non-polarized morphospace. Nevertheless, inclusion of semilandmarks display a clear increase in the cumulative degree of variation represented by each graph in both shape ($\mathbb{R}^{59}$, PC1&2 = 53.06%, PC1-10 = 92.60%) and form ($\mathbb{R}^{60}$, PC1&2 = 81.38%, PC1-10 = 97.91%). These results directly indicate that semilandmark based models are able to increase the chance of finding structure within the data by minimizing bending energy and thus reducing noise [72]. This in turn increases the possibility of finding statistical separation between the samples. These results are also seen through CVA results where clear separations between wolf and dog samples are noticeable (Fig 11E and 11F), with equally as significant Mahalanobis (D = 5.88, $p < 0.0001$) and Procrustes (D = 0.067, $p = 0.012$) distance values. While CVA results are likely to be overestimate separation considering sample size and the increased number of variables when including semilandmarks [102, 103], other statistical tests such as those obtained through MANOVA are also able to highlight significant differences between samples in both shape ($p = 0.004$) and form ($p = 0.007$).

PCA results do not vary to a significant degree upon increasing the number of semilandmarks, with the total accumulative variance increasing by only 0.002% across the first 10 PC scores using 36 semilandmarks. Beyond 49 semilandmarks, percentage of represented variance even begins to decrease and MANOVA results also display poorer separation between samples. Likewise, CVA results on the other hand deteriorate with the inclusion of more semilandmarks, with less clear differences between groups. This is to be expected when considering the sample sizes available, and thus number of semilandmarks should be reduced [102, 103]. This concludes that the optimal statistical model for discerning between carnivore agencies remains to be the 30-landmark model presented here (S1 Appendix).

## 4. Discussion and conclusions

Throughout the years, experimentation has become a fundamental component in archaeological and palaeontological research. Needless to say, with the ever-growing realisation of the impact true analogy has on experimental results, creating parallelisms with the fossil record is often difficult. Product of this is an increasing importance in the availability of experimental samples. This is especially apparent in geographic regions where the present day wild fauna is drastically different to that which existed over 10,000 years ago. In areas such as Europe, access to wild carnivores such as hyenas, lions and leopards is particularly difficult, highlighting the need for collaborative efforts in order to study the archaeological/palaeontological register effectively.

In archaeology/palaeontology, a possible response to this situation has been the collection and pooling of morphometric datasets via online resources [104, 105 *inter alia*], while some datasets are available through supplementary materials [e.g. 56–59]. Nevertheless, pooling of data from different sources is often conflictive, and induced errors are likely to increment if data collection is not strictly controlled [67]. Moreover, in many cases making 3D models readily available often requires a certain degree of computational costs and obtaining permissions to share these files can often be problematic. This latter point is unfortunately due to the additional issue of competitivity amongst many research teams, making scientific progress much more politically, rather than empirically, oriented.

In fields of data science, the sharing of information can be considered indispensable, contributing to some of the most important advances in applied sciences. This is especially relevant when confronting issues of generalisation in predictive modelling [106]. With the advent of *Domain Adaptation* in computational learning [106, 107: pg. 526–531], the development of artificially intelligent Transfer Learning [108, 109], and the creation of One-Shot/Zero-Shot models [110–113], new possibilities for collaborative efforts in data science are now available to solve many research questions. Similar significant advances through *Multimodal* and *Federated Optimization* strategies present new possibilities for the collaborative integration of different datasets [114–116]. Sharing information and training classification models over multiple datasets in this manner presents the distinct advantage of protecting components of data privacy and overcoming issues presented by dataset centralisation [117].

Through more open-minded collaborations among teams, and greater efforts to make these learning strategies available, it may be possible for archaeological and palaeontological researchers to obtain similar success rates to other fields of science [118–120].

Nevertheless, in order for this to happen, careful attention must be paid to the statistical quality of the datasets being used. While tooth pit analyses have been able to reach $> 95\%$

classification using GMMs [58, 59, 121], each of these studies have the distinct advantage of landmark data being collected by a single experienced individual [personal communication and direct participation]. Furthermore, while the original 17-landmark model was able to successfully draw conclusions from each of these studies (with the inclusion of [58, 59, 121, 122]), the integration of data produced here may increase the precision of these results. The results presented within this research thus represent a means of removing as much analyst-induced subjectivity as possible in the study of carnivore tooth pits. This is achieved through replacing weak landmark points with more precise computational semilandmarks. Nevertheless, as seen here, LM1-5 still present an important margin of error that must be confronted.

It is common knowledge that in-person training is the most valuable tool for obtaining optimal results in any task. Such an observation has been proven true not only for geometric morphometric research [121], but is also applicable to evolutionary theory [123, 124]. Likewise, the descriptive quality of any metric model is fundamental in its reproducibility. Under this premise, so as to ensure reproducibility and accuracy, a visual guide to the proposed methodological approach has been provided and is available at https://vimeo.com/409256777. This is further accompanied by a detailed description of the fixed landmarks (LM1-5) employed in this study (S1 Appendix), and with graphical representation in Fig 2D. Needless to say, on a practical note the time and care taken to process a 3D model is clearly an important component to consider. Although subjective, personal observations of the different analysts processing the models revealed clear differences between some individuals based on how thorough they were when handling the 3D models. These differences became increasingly obvious as time went by in each work session. While experience is a clear variable of importance, we strongly advise that even the most experienced observers carry out digitisation procedures as meticulously as possible over numerous digitisation sessions and taking full advantage of the tools in most 3D modelling software (e.g. measurement, zoom, pan, and rotating tools). While it is true that repeatability measures are likely to be higher in computerised procedures [66, 69], which is one of the clearest advantages to the methodological approach presented here, work ethic is still a valuable variable that needs to be controlled at all times.

Focusing specifically on the present study, human-induced landmark errors of the new proposed model can be conclusively defined at 0.139 +/- 0.092 $\in$ {0.002:0.586} mm (Median +/- $\sqrt{}$BWMV $\in$ {Min:Max}). This presents a total drop in error of 164 μm from the original model which is extrapolated mainly from the statistically noisy landmarks LM14-17. As can be seen by the error range intervals, differences between samples are highly skewed, supporting our use for a robust statistical approach in defining the final model [83]. While the impact this margin of error has on statistical results is arguably not likely to affect the model's ability to separate between carnivore samples (as seen through MANOVA results), it is important to highlight the power that *form* has over *shape* in geometric morphometric analysis [47].

While Procrustes analyses are a valuable tool for most applied geometric applications [35, 65, 125, 126], it has been seen throughout our results how full Procrustes superimposition produces a notable statistical distortion that cannot be ignored in taphonomic GMM. This is direct evidence of the power the "*Pinocchio effect*" can have on geometric morphometric data (Current study: S1 Fig of S1 File; [65, 82]). Nevertheless, recent research into the power of computational learning algorithms highlight the value of *form* data in the predictive modelling of tooth pits [59]. This supports the removal of scaling processes for tooth pit classification tasks, additionally providing an interesting reflection on *form* and *shape* theory in modern morphometrics [127]. While it is also arguable that the increased presence of a *Pinocchio effect* is product of the microscopic scale being used [64, 66, 70], we insist on our recommendations that *form* data be used when working with decentralized pooled data for classification tasks on tooth pits.

Finally, the variability present in canid tooth marks is of increasing interest, considering the statistical noise they are frequently noted to produce [28, 56–59]. As seen here both qualitatively and quantitatively, wolf tooth marks in general are harder to process and frequently produce an increase in error (Deviations from absolute error = {+4 μm : +14 μm}, Table 1). This is primarily due to their more superficial nature and great variability which is an interesting topic to note and should be explored in detail in the future [58], till then, caution is certainly advised when working with this animal.

In conclusion, this study presents a detailed revision of the current methodological approach available for discerning carnivore agencies via tooth pit morphology. Through this methodological update, we present a means of reducing human induced error in taphonomic geometric morphometric data collection; facilitating the pooling of inter-institutional datasets for the training of classification models. While the theoretical and statistical reflections presented here are extensive, if these protocols are strictly followed and their true implications considered, tooth pit morphological analyses can have a promising future in applied scientific research. Under this premise, we predict lines of investigation of this type will be an encouraging development relevant in palaeontological, archaeological, forensic and even modern-day ecological sciences.

## Supporting information

**S1 Appendix.**
(PDF)

**S1 File.**
(PDF)

## Acknowledgments

Firstly, we greatly appreciate the meticulous reviews and comments made by one anonymous reviewer, Dr. Ros Coard and the editorial staff of PLoS ONE. We truly believe that their comments and suggestions have improved the quality of the present paper. Next, we would like to thank the members and staff of the Musée de l'Homme in Paris, including Florent Detroit, Antoine Balzeau and a special thanks to Martin Friess for their introduction to topics discussed and presented within this paper. Next, we would like to thank the helpful comments by Abel Moclán, Isabel Cáceres and Toni Pineda on earlier version of these studies. We would also like to acknowledge the help and support provided by the staff of the Cábarceno Natural Park and Santiago Borragan for providing the samples used in our experiments. Finally, we would also like to thank the TIDOP Group from the Department of Cartographic and Land Engineering of the Higher Polytechnics School of Avila, University of Salamanca, for the use of their tools and facilities.

## Author Contributions

**Conceptualization:** Lloyd A. Courtenay, Rosa Huguet.

**Data curation:** Lloyd A. Courtenay.

**Formal analysis:** Lloyd A. Courtenay.

**Funding acquisition:** Diego González-Aguilera.

**Investigation:** Lloyd A. Courtenay.

**Methodology:** Lloyd A. Courtenay, Darío Herranz-Rodrigo.

**Project administration:** Rosa Huguet, Diego González-Aguilera, José Yravedra.

**Resources:** Miguel Ángel Maté-González, José Yravedra.

**Software:** Lloyd A. Courtenay.

**Supervision:** Rosa Huguet, Diego González-Aguilera, José Yravedra.

**Validation:** Lloyd A. Courtenay, Darío Herranz-Rodrigo.

**Visualization:** Lloyd A. Courtenay.

**Writing – original draft:** Lloyd A. Courtenay.

**Writing – review & editing:** Lloyd A. Courtenay, Darío Herranz-Rodrigo, Rosa Huguet, Diego González-Aguilera, José Yravedra.

## References

1. Haynes G. Evidence of carnivore gnawing on pleistocene and recent mammalian bones. Paleobiol. 1980; 6(3), 341–351.
2. Haynes G. A guide for differentiating mammalian carnivore taxa responsible for gnaw damage to herbivore limb bones. Paleobiol. 1983; 9(2), 164–172.
3. Binford LR. Bones: Ancient Men and Modern Myths. New York: Academic Press Inc; 1981.
4. Brain CK. Hunters or the Hunted? An Introduction to African Cave Taphonomy. Chicago: University of Chicago Press; 1981.
5. Blumenschine RJ. Percussion marks, tooth marks and the experimental determinations of the timing of hominid and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania. J. Hum. Evol. 1995; 29, 21–51.
6. Johnson DL, Haynes CV. Camels as taphonomic agents. Quat. Res. 1985; 24, 365–366.
7. Cáceres I, Esteban-Nadal M, Bennàsar M, Fernández-Jalvo Y. Was it the Deer or the Fox? J. Archaeol. Sci. 2011; 38, 2767–2774.
8. Hutson JM, Burke CC, Haynes G. Osteophagia and bone modifications by giraffe and other large ungulates. J. Archaeol. Sci. 2013; 40, 4139–4149.
9. Domínguez-Solera SD, Domínguez-Rodrigo M. A Taphonomic study of bone modification and of tooth-mark patterns on long limb bone portions by Suids. Internat. J. Osteoarchaeol. 2009; 19, 345–363.
10. Saladié P, Huguet R, Díez C, Rodríguez-Hidalgo A, Carbonell E. Taphonomic modifications produced by modern brown bears (Ursus arctos). Internat. J. Osteoarchaeol. 2011; 23(1), 13–33.
11. Backwell LR, Parkinson AH, Roberts EM, d'Errico F, Huchet JB. Criteria for identifying bone modification by termites in the fossil record. Palaeogeog., Palaeoclimatol., Palaeoecol. 2012; 337, 72–87.
12. Sanders WJ, Trapani J, Mitani JC. Taphonomic aspects of crow hawk-eagle predation on monkeys, J. Human Evol. 2003; 44, 87–105
13. Lloveras L, Cosso A, Solé J, Claramunt-López B, Nadal J. Taphonomic signature of Golden eagles (Aquila chrysaetos) on bone prey remains, Historical Biology. 2017 https://doi.org/10.1080/08912963.2017.1319830
14. Singer R. The "Bone Tools" from Hopefield. Amer. Anthropol. 1956; 58, 1127–1134.
15. Peterhans JCK, Singer R. Taphonomy of a lair near the peers (or skildegat) cave in Fish Hoek, Western Cape Province, South Africa. S. Afr. Archaeol. Bull. 2006; 61(183), 2–18.
16. O'Regan HJ, Kuman K, Clarke RJ. The likely accumulators of bones: Five cape porcupine den assemblages and the role of porcupines in the Post-Member 6 Infill at Sterkfontein, S. Afr. J. Taphonomy. 2011; 9(2), 69–87
17. Ames JA, Morejohn GV. Evidence of white shark, Carcharodon carcharius, attacks on sea otters, Enhydra lutris. California Fish and Game. 1980; 66, 196–209.
18. Deméré TA, Cerutti RA. A Pliocene shark attack on a cetotheriid wale. J. Palaeontol. 1982; 56, 1480–1482.

19. Njau JK, Blumenschine RJ. A diagnosis of crocodile feeding traces on larger mammal bone, with fossil examples from the Plio-Pleistocene Olduvai Basin, Tanzania. J. Hum. Evol. 2006; 50(2), 142–162 https://doi.org/10.1016/j.jhevol.2005.08.008 PMID: 16263152

20. D'Amore DC, Blumenschine RJ. Komodo monitor (Varanus komodensis) feeding behaviour and dental function reflected through tooth marks on bone surfaces as the application of ziphodont paleobiology. Paleobiol. 2009; 35, 525–552.

21. Erickson GM, Olson KH. Bite marks attributable to Tyrannosaurus rex: Preliminary description and implications. J. Vertebr. Paleontol. 1996; 16(1), 175–178.

22. Rogers RR, Krause DW, Rogers KC. Cannibalism in the Madagascan dinosaur *Majungatholus atopus*. Nature. 2003; 422, 515–518. https://doi.org/10.1038/nature01532 PMID: 12673249

23. Longrich NR, Horner JR, Erickson GM, Currie PJ. Cannibalism in Tyrannosaurus rex. PLoS ONE. 2010; 5(10).

24. Fernández-Jalvo Y, Andrews P. When humans chew bones. J. Hum. Evol. 2011; 60, 117–123. https://doi.org/10.1016/j.jhevol.2010.08.003 PMID: 20951407

25. Saladié P, Rodríguez-Hidalgo A, Díez C, Martín-Rodríguez P, Carbonell E. Range of bone modifications by human chewing. J. Archaeol. Sci. 2012; 40(1), 380–397.

26. Domínguez-Rodrigo M, Piqueras A. The use of tooth pits to identify carnivore taxa in tooth-marked archaeofaunas and their relevance to reconstruct hominid carcass processing behaviours. J. Archaeol. Sci. 2003; 30, 1385–1391.

27. Delaney-Rivera C, Plummer TW, Hodgson JA, Forrest F, Hertel F, Oliver JS. Pits and pitfalls: taxonomic variability and patterning in tooth mark dimensions J. Archaeol. Sci. 2009; 36(11), 2597–2608.

28. Yravedra J, Lagos L, Bárcena F. A taphonomic study of wild wolf (Canis lupus) modification of horse bones in Northwestern Spain. J. Taphonomy. 2011; 9(1), 37–65.

29. Andrés M, Gidna AO, Yravedra J, Domínguez-Rodrigo M. A study of dimensional differences of tooth marks (pits and scores) on bones modified by small and large carnivores. Archaeol. Anthropol Sci. 2012; 4(3), 209–219.

30. Yravedra J. A taphonomic perspective on the origins of the faunal remains from Amalda Cave (Spain). J. Taphonomy. 2011; 8(4), 301–334.

31. Domínguez-Rodrigo M, Barba R, Egeland CP. Deconstructing Olduvai: A taphonomic study of the Bed I sites. Netherlands: Springer; 2007.

32. Saladié P, Fernández P, Rodríguez-Hidalgo A, Huguet R, Pineda A, Cáceres I, et al. The TD6.3 faunal assemblage of the Gran Dolina site (Atapuerca, Spain): a late Early Pleistocene hyena den. Hist. Blo. 2019; 31(6), 665–683.

33. Rodríguez-Hidalgo A, Saladié P, Ollé A, Arsuaga JL, Bermúdez de Castro JM, Carbonell E. Human predatory behavior and the social implications of communal hunting based on evidence from the TD10.2 bison bone bed at Gran Dolina (Atapuerca, Spain). J. Hum. Evol. 2017; 105, 89–122. https://doi.org/10.1016/j.jhevol.2017.01.007 PMID: 28366202

34. Pineda A, Saladié P. The Middle Pleistocene site of Torralba (Soria, Spain): a taphonomic view of the Marquis of Cerralbo and Howell faunal collections. Archaeol Anthropol Sci. 2019; 11, 2539–2556.

35. Bookstein FL. Morphometric tools for landmark data. Cambridge: Cambridge University Press; 1991.

36. Rohlf FJ, Marcus LR. A revolution in morphometrics. Trends in Ecol. Evol. 1993; 8, 129–132.

37. O'Higgins P. The study of morphological variation in the hominid fossil record: Biology, landmarks and geometry. J. Anat. 2000; 197, 103–120. https://doi.org/10.1046/j.1469-7580.2000.19710103.x PMID: 10999273

38. Singleton M. Patterns of cranial shape variation in the Papionini (Primates: Cercophithecinae). J. Hum. Evol. 2002; 42, 547–578. https://doi.org/10.1006/jhev.2001.0539 PMID: 11969297

39. Guy F, Brunet M, Schmittbuhl M, Viriot L. New approaches in hominoid taxonomy: morphometrics. Amer. J. Phy. Anthropol. 2003; 121, 198–218.

40. Baylac M, Villemant C, Simbolotti G. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. Biol. J. Linnean Soc. 2003; 80(1), 89–98.

41. Lockwood CA, Lynch JM, Kimbel WH. Quantifying temporal bone morphology of great apes and humans: an approach using geometric morphometrics. J. Anat. 2002; 201, 447–464. https://doi.org/10.1046/j.1469-7580.2002.00122.x PMID: 12489757

42. Bastir M, Rosas A. Hierarchical nature of morphological integration and modularity in the human posterior face, Amer. J. Phy. Anthropol. 2005; 128, 26–34.

43. Galland M, Friess M. Three-Dimensional Geometric Morphometrics view of the cranial shape variation and population history in the New World. Amer. J. Phy. Anthropol. 2016; 28, 646–661.

**44.** Gunz P, Neubauer S, Falk D, Tafforeau P, Le Cabec A, Smith TM, et al. *Australopithecus afarensis* endocasts suggest ape-like brain organization and prolonged brain growth. Sci. Adv. 2020; 6(14), eaaz4729.

**45.** García-Medrano P, Ollé A, Ashton N, Roberts MB. The Mental Template in Handaxe Manufacture: New Insights into Acheulean Lithic Technological Behavior at Boxgrove, Sussex, UK. J. Archaeol. Meth. Theor. 2018; 26(1), 396–422.

**46.** Lycett SJ, Cramon-Taubadel N, Foley RA. A Crossbeam Co-Ordinate Caliper for the Morphometric Analysis of Lithic Nuclei: a Description, Test and Empirical Examples of Application. J. Archaeol. Sci. 2006; 33, 847–861.

**47.** Dryden IL, Mardia KV. Statistical Shape Analysis. New York: John Wiley and Sons; 1998.

**48.** Bookstein FL. Principal Warps: Thin Plate Spline and the Decomposition of Deformations. Transactions on Pattern Anal. Mach. Intell. 1989; 11(6), 567–585.

**49.** Richtsmeier JT, Lele SR, Cole TM. Landmark morphometrics and the analysis of variation. In: Hallgrimsson B. and Hall B.K. (Eds.) Variation. Boston: Elsevier Academic Press. 2005; 49–68.

**50.** Maté-González MÁ, Yravedra J, González-Aguilera D, Palomeque-González JF, Domínguez-Rodrigo M. Micro-photogrammetric characterization of cut marks on bones. J. Archaeol. Sci. 2015; 62, 128–142.

**51.** Courtenay LA, Yravedra J, Maté-González MÁ, Aramendi J, González-Aguilera D. 3D Analysis of Cut Marks using a New Geometric Morphometric Methodological Approach. J. Archaeol. Anthropol. Sci. 2019; 11, 651–665.

**52.** Yravedra J, Diez-Martín F, Egeland CP, Maté-González MÁ, Palomeque-González JF, Arriaza MC, et al. FLK-West (Lower Bed II, Olduvai Gorge, Tanzania): a new early Acheulean site with evidence for human exploitation of fauna. Boreas. 2017; 46(3), 486–502.

**53.** Komo L, Grassberger M. Experimental Sharp Force injuries to ribs: multimodal morphological and geometric morphometric analyses using micro-CT, macro photography and SEM, Forensic Sci. Int. 2018; 288, 189–200 https://doi.org/10.1016/j.forsciint.2018.04.048 PMID: 29758447

**54.** Kieser J, Bernal V, Gonzalez P, Birch W, Turmaine M, Ichim I. Analysis of experimental cranial skin wounding from screwdriver trauma, Int. J. Legal Med. 2008; 122, 179–187 https://doi.org/10.1007/s00414-007-0187-1 PMID: 17701196

**55.** Aramendi J, Maté-González MÁ, Yravedra J, Ortega MC, Arriaza MC, González-Aguilera D, et al. Discerning carnivore agency through the three-dimensional study of tooth pits: Revisiting crocodile feeding behaviour at FLK- Zinj and FLK NN3 (Olduvai Gorge, Tanzania). Palaeogeog., Palaeoclimatol., Palaeoecol. 2017; 488, 93–102.

**56.** Yravedra J, García-Vargas H, Maté-González MÁ, Aramendi J, Palomeque-González JF, Vallés-Iriso J, et al. The use of Micro-Photogrammetry and Geometric Morphometrics for identifying carnivore agency in bone assemblages. J. Archaeol. Sci. Rep. 2017; 14, 106–115.

**57.** Yravedra J, Aramendi J, Maté-González M.Á, Courtenay LA, González-Aguilera D. Differentiating Percussion Pits and Carnivore Tooth Pits using 3D Reconstructions and Geometric Morphometrics, PLoS ONE. 2018; 13(3), e0194324. https://doi.org/10.1371/journal.pone.0194324 PMID: 29590164

**58.** Yravedra J, Maté-González MÁ, Courtenay LA, González-Aguilera D, Fernández-Fernández M. The use of canid tooth marks on bone for the identification of livestock predation. Sci. Rprts. 2019; 9, 16301.

**59.** Courtenay LA, Yravedra J, Huguet R, Aramendi J, Maté-González MÁ, González-Aguilera D et al. Combining Machine Learning Algorithms and Geometric Morphometrics: a Study of Carnivore Tooth Marks. Palaeogeog., Palaeoclimatol., Palaeoecol. 2019; 522, 28–29

**60.** Valeri CJ, Cole TM, Lele S, Richtsmeier JT. Capturing data from three-dimensional surfaces using fuzzy landmarks. Amer. J. Phy. Antrhopol. 1998; 107, 113–124.

**61.** Sholts SB, Flores L, Walker PL, Wärmländer SKTS. Comparison of Coordinate Measurement Precision of Different Landmark Types on Human Crania using a 3D Laser Scanner and a 3D Digitiser: Implications for Applications of Digital Morphometrics. Internat. J. Osteoarchaeology. 2011; 21, 535–543.

**62.** Yezerniac SM, Lougheed SC, Handford P. Measurement error and morphometric studies: statistical power and observer experience. Systemat. Bol. 1992; 41(4), 471–482.

**63.** Arnqvist G, Martensson T. Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape. Acta Zoologica Academiae Scientiarum Hungaricae. 1998; 44(1–2), 73–96.

**64.** Blackwell GL, Bassett SM, Dickman CR. Measurement error associated with external measurements commonly used in small-mammal studies. J. Mammology. 2006; 87(2), 216–223.

**65.** Cramon-Taubadel N, Frazier BC, Lahr MM. The problem of assessing landmark error in geometric morphometrics: theory methods and modifications. Amer. J. Phy. Anthropol. 2017; 134, 24–35.

**66.** Muñoz-Muñoz F, Perpiñán D. Measurement error in morphometric studies: comparison between manual and computerized methods. Annales Zoologici Fennici. 2010; 47(1), 46–56.

**67.** Robinson C, Terhune CE. Error in geometric morphometric data collection: Combining data from multiple sources. Amer. J. Phy. Anthropol. 2017; 164(1), 62–75.

**68.** Shearer BM, Cooke SB, Halenar LB, Reber SL, Plummer JE, Delson E, et al. Evaluating causes of error in landmark-based data collection using scanners. PLoS ONE. 2017; 12(11), e0187452. https://doi.org/10.1371/journal.pone.0187452 PMID: 29099867

**69.** Daboul A, Ivanovska T, Bülow R, Biffar R.; Cardini A. (2018) Procrustes-based geometric morphometrics on MRI images: An example of inter-operator bias in 3D landmarks and its impact on big datasets, PLoS ONE. 13(5):e0197675 https://doi.org/10.1371/journal.pone.0197675 PMID: 29787586

**70.** Jamison PL, Ward RE. Brief Communication: Measurement Size Precision and Reliability in Craniofacial Anthropometry: Bigger is Better. Amer. J. Phy. Anthropol. 1993; 90, 495–500.

**71.** Gunz P, Mitteroecker P, Bookstein FL, Weber GW. Computer aided reconstruction of incomplete human crania using statistical and geometrical estimation methods. Enter the past: computer applications and quantitative methods in archeology. BAR Internat. Series. 2004; 1227, 96e98.

**72.** Gunz P, Mitteroecker P, Bookstein FL. Semilandmarks in three dimensions in: Slice DE (Ed). Modern Morphometrics in Physical Antrhopology. New York: Plenum Publishers. 2005; 73–98.

**73.** Gunz P, Mitteroecker P, Neubauer S, Weber GW, Bookstein FL. Principles for the Virtual Reconstruction of Hominin Crania. J. Hum. Evol. 2009; 57, 48–62. https://doi.org/10.1016/j.jhevol.2009.04.004 PMID: 19482335

**74.** Raina, R.; Ng, A.Y.; Killer, D. (2006) Constructing Informative Priors using Transfer Learning, Twentythird International Conference on Machine Learning. DOI: 10.1145/1143844.1143934

**75.** Gidna A.; Yravedra J.; Domínguez-Rodrigo M. (2013) A cautionary note on the use of captive carnivores to model wild predator behavior: a comparison of bone modification patterns on long bones by captive and wild lions, Journal of Archaeological Science. 40:1903–1910

**76.** Mencha J.A.; Kreger M.D. (1996) Ethical and welfare issues associated with keeping wild mammals in captivity. In: Kleiman D.G.; Allen M.E.; Thompson K.V.; Lumpkin S. (Eds.) Wild Mammals in Captivity: Principles and Techniques. Chicago: The University of Chicago Press. 5–15

**77.** Maté-González MA, Aramendi J, Yravedra J, González-Aguilera D. Statistical Comparison between Low-Cost Methods for 3D Characterization of Cut-Marks on Bones. Remot. Sens. 2017; 9(9), 873.

**78.** Wiley DF, Amenta N, Alcantara DA, Ghosh D, Kil YJ, Delson E, et al. Evolutionary Morphing. Proceedings of the IEEE Visualization 2005 (VIS'05). 2005; 431–438.

**79.** O'Higgins P, Jones N. Facial Growth in Cercocebus torquatus: an application of three-dimensional geometric morphometric techniques to the study of morphological variation. J. Anatomy. 1998; 193, 251–272.

**80.** Viðarsdóttir US, O'Higgins P, Stringer C. A geometric morphometric study of regional differences in the ontogeny of the modern human facial skeleton. J. Anatomy. 2002; 201, 211–229.

**81.** Corner BD, Lele S, Richtsmeier JT. Measuring precision of three-dimensional landmark data. J. Quant. Anthropol. 1992; 3, 347–359.

**82.** Chapman RE. Conventional Procrustes Approaches In: FJ Rohlf and FL Bookstein (Eds). Proceedings of the Michigan Morphometrics Workshop. Ann Arbor: University of Michigan Museum of Zoology. 1990; 2, 251–267.

**83.** Höhle J, Höhle M. Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS J. Photogramm. Remot. Sens. 2009; 64, 398–406.

**84.** Cichosz P. (2015) Data Mining Algorithms: Explained using R. Chichester: John Wiley & Sons

**85.** Rodríguez-Gonzálvez P, Garcia-Gago J, Gomez-Lahoz J, González-Aguilera D. Confronting passive and active sensors with Non-Gaussian statistics. Sens. 2014; 14, 13759–13777.

**86.** Hasan A, Pilesjö P, Persson A. The use of LIDAR as a data source for digital elevation models–a study of the relationship between the accuracy of digital elevation models and topographical attributes in northern peatlands. Hydrol. E. Syst. Sci. Discuss. 2011; 8(3), 5497–5522.

**87.** Herrero-Huerta M, Lindenbergh R, Rodríguez-Gonzálvez P. Automatic tree parameter extraction by a mobile LiDAR System in an urban context. PLoS ONE. 2018; 13(4), e0196004. https://doi.org/10.1371/journal.pone.0196004 PMID: 29689076

**88.** Ariza-López FJ, Rodríguez-Avi J, González-Aguilera D, Rodríguez-Gonzálvez P. A new method for positional accuracy control for non-normal errors applied to airborne laser scanning data. Appl. Sci. 2019; 9, 1–18.

89. Rodríguez-Martín M, Rodríguez-Gonzálvez P, Ruiz de Oña Crespo E, González-Aguilera D. Validation of portable mobile mapping system for inspection tasks in thermal and fluid-mechanical facilities. Remot. Sens. 2019; 11, 1–19.

90. Heathcote GM. The magnitude and consequences of measurement error in human craniometry. Canad. Rev. Phy. Antrhopol. 1981; 3, 18–40.

91. Hanihara T, Dodo Y, Kondo O, Nara T, Doi N, Sensui N. Intra- and Interobserver errors in facial flatness measurements. Anthropol. Sci. 1999; 107(1), 25–39.

92. Vargha A, Delaney HD. The Kruskal-Wallis test and stochastic homogeneity. J. Educat. Behavioral Stats. 1998; 23, 170–192.

93. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. San Diego: Elsevier; 2005.

94. Tomarken AJ, Serlin RC. Comparison of ANOVA Alternatives under Variance Heterogeneity and Specific Noncentrality Structures. Quant. Meth. Psychol. 1986; 99(1), 90–99.

95. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J, Am. Stat. Assoc. 1952; 47 (260), 583–621

96. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd Internat. Conference on Knowledge Discovery and Data Mining München Germany. 1996. pp. 226–231.

97. Satopa V, Albrecht J, Irwin D, Raghavan B. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. IEEE 31st Conference on Distributed Computing Systems Workshops. 2011. pp. 166–171.

98. Fukunaga K, Hostetler L. The estimation of the gradient of a density function with applications in pattern recognition. IEEE Trans. Info. Theory. 1975; 21(1), 32–40.

99. Cheng Y. Mean shift mode seeking and clustering. IEEE Transactions on Pattern Anal. Mach. Intell. 1995; 17(8), 790–799.

100. Comaniciu D, Meer P. Mean Shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Anal. Mach. Intell. 2002; 24(5), 603–619.

101. Bookstein FL. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med. Img. Anal. 1996; 1, 225–243.

102. Albrecht GH. Assessing the affinities of fossils using canonical variates and generalized distances. 1992; 7(4), 46–69.

103. Kovarovic K, Aiello LC, Cardini A, Lockwood CA. Discriminant function analyses in archaeology: are classification rates too good to be true? 2011; 38(11), 3006–3018. https://doi.org/10.1016/j.jas.2011.06.028

104. Adams JW, Olah A, McCurry MR, Potze S. Surface model and tomographic archive of fossil primate and other mammal holotype and paratype specimens of the Ditsong National Museum of Natural History Pretoria South Africa. PLoS ONE. 2015; 10, e0139800. https://doi.org/10.1371/journal.pone.0139800 PMID: 26441324

105. Copes LE, Lucas LM, Thostenson JO, Hoekstra HE, Boyer DM. A collection of non-human primate computed tomography scans housed in MorphoSource a repository for 3D data. Sci. Data. 2016; 3(1), 1–8.

106. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016.

107. Pratt LY. Discriminability-Based Transfer between Neural Networks. Neural Info. Process. Syst. 1993; 5, 204–211.

108. Pan SJ, Yang Q. A survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(10), 1345–1359.

109. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Adv. Neural Info. Process. Syst. 2014. pp. 33220–3328.

110. Fei-Fei L, Fergus R, Perona P. One-Shot Learning for Object Categories. IEEE Transactions on Pattern Anal. Mach. Intell. 2006; 28(4), 594–611.

111. Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines. Proceedings of the 25th Internat. Conference on Mach. Learn. 2008. pp. 536–543.

112. Palatucci M, Pomerleau D, Hinton G, Mitchell TM. Zero-Shot learning with semantic output codes. Neural Inf. Process. Syst. 2009; 22, 1410–1418.

113. Socher R, Ganjoo M, Sridhar H, Bastani O, Manning CD, Ng AY. Zero-Shot Learning through Cross-Modal Transfer. Neural Inf. Process. Syst. 2013. pp. 935–943.

114. Srivastava N, Salakhutdinov R. Multimodal Learning with Deep Bolzmann Machines. J. Mach. Learn. Res. 2014; 15, 2949–2980.

**115.** Konečny J, McMahan HB, Ramage D. Federated Optimization: Distributed Optimization Beyond the Datacenter. Neural Info. Process. Sys. arXiv: 151103575v1. 2015.

**116.** Konečny J, McMahan HB, Yu FX, Suresh AT, Bacon D, Richtárik P. Federated Learning: Strategies for Improving Communication Efficiency. Neural Info. Process. Sys. arXiv: 161005492v2. 2017.

**117.** McMahan B, Moore E, Ramage D, Hampson S, Agüera B. Communication-efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th Internat. Conference on Art. Intell. Statistics. arXiv: 160205629v3. 2017.

**118.** AlBadawy EA, Saha A, Mazurowski MA. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. Med. Phy. 2018; 45(3), 1150–1158.

**119.** Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, et al. Distributed Deep Learning networks among institutions for medical imaging. J. the Amer. Med. Infor. Association. 2018; 25(8), 945–954.

**120.** Balachandar N, Chang K, Kalpathy-Cramer J, Rubin DL. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. J. Amer. Infor. Association. 2020. pp. 1–9.

**121.** Aramendi J, Arriaza MC, Yravedra J, Maté-González MÁ, Ortega MC, Courtenay LA, et al. Who ate OH80 (Olduvai Gorge, Tanzania)? A geometric-morphometric analysis of Surface bone modifications of a Paranthropus boisei skeleton. Quat. Internat. 2019; 517, 118–130.

**122.** Arriaza MC, Aramendi J, Maté-González MÁ, Yravedra J, Stratford D. Characterising leopard as taphonomic agent through the use of micro-photogrammetric reconstruction of tooth marks and pit to score ration. Hist. Blo. 2019. pp. 1–10.

**123.** Lombao D, Guardiola M, Mosquera M. Teaching to make stone tools: new experimental evidence supporting a technological hypothesis for the origins of language. Sci. Rep. 2017; 7(1), 1–14. https://doi.org/10.1038/s41598-016-0028-x PMID: 28127051

**124.** Petö R, Elekes F, Oláh K, Király I. Learning how to use a tool: mutually exclusive tool-function mappings are selectively acquired from linguistic in-group models. J. Exp. Child Psychol. 2018; 171, 99–112. https://doi.org/10.1016/j.jecp.2018.02.007 PMID: 29567562

**125.** Crosilla F, Beinat A. Use of Generalised Procrustes Analysis for the photogrammetric block adjustment by independent models. ISPRS J. Photogramm. Remot. Sens. 2002; 56(3), 195–209.

**126.** González-Aguilera D, Gómez-Lahoz J, Muñoz-Nieto Á, Herrero-Pascual J. Monitoring the health of an emblematic monument from terrestrial laser scanner. Nondestructive Test. Eval. 2008; 23(4), 301–315.

**127.** Mitteroecker P, Gunz P, Windhager S, Schaefer K. A brief review of shape form and allometry in geometric morphometrics with application to human facial morphology. Hystrix Italian J. Mammalogy. 2013; 24(1), 59–66.

*Spanish Translation of Title and Abstract*

# Análisis 3D de los efectos de la cautividad sobre la masticación y las marcas de diente del lobo; implicaciones en los estudios ecológicos tanto del presente como del pasado.

Las poblaciones humanas han desarrollado relaciones complejas con las especies de grandes carnívoros a lo largo del tiempo, con evidencias tanto de competencia como de colaboración para obtener recursos a lo largo del Pleistoceno. Desde esta perspectiva, muchos yacimientos arqueológicos y paleontológicos presentan pruebas de modificaciones en los huesos por parte de los carnívoros. Ante ello, los especialistas en el estudio de las modificaciones microscópicas de la superficie ósea han recurrido tanto al uso de técnicas de modelado 3D como a la ciencia de datos para la inspección de estos elementos, alcanzando resultados novedosos para la diferenciación entre la actividad de diferentes carnívoros. El presente estudio analiza la variabilidad de las marcas de dientes producidas por múltiples grupos de lobo ibérico, con el objetivo de estudiar cómo el cautiverio puede afectar a las marcas de dientes que dejan en el hueso. Aquí se comparan cuatro poblaciones de lobo distintas, tanto salvajes como en cautividad, para profundizar en la variabilidad intraespecífica. Esta investigación muestra estadísticamente que las marcas tipo depresión de los cánidos son las menos afectadas por el cautiverio, mientras que las marcas tipo surco son más superficiales cuando son producidas por lobos cautivos. La naturaleza superficial de los surcos producidos por lobos cautivos se ve además correlacionada con otras características métricas, influyendo así en la morfología general de las marcas. Desde esta perspectiva, el presente estudio abre un nuevo diálogo sobre las razones que subyacen a los resultados obtenidos, aconsejando precaución a la hora de utilizar las marcas de diente tipo surco para la identificación de carnívoros, y contemplando elementos como el estrés como una de las principales variables que influyen en la morfología de las marcas dejadas por lobos.

*Supplementary Information and Links*

**Supplementary Information available from:**
`https://www.mdpi.com/2076-2615/11/8/2323#supplementary`

*animals*

MDPI

*Article*

# 3D Insights into the Effects of Captivity on Wolf Mastication and Their Tooth Marks; Implications in Ecological Studies of Both the Past and Present

**Lloyd A. Courtenay [1,*], Darío Herranz-Rodrigo [2,3], José Yravedra [2,3], José Mª Vázquez-Rodríguez [4], Rosa Huguet [5,6,7], Isabel Barja [8,9], Miguel Ángel Maté-González [1,10], Maximiliano Fernández Fernández [11,12], Ángel-Luis Muñoz-Nieto [1] and Diego González-Aguilera [1,11]**

[1] Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003 Ávila, Spain; mategonzalez@usal.es (M.Á.M.-G.); almuni@usal.es (Á.-L.M.-N.); daguilera@usal.es (D.G.-A.)
[2] Department of Prehistory, Complutense University, Prof. Aranguren s/n, 28040 Madrid, Spain; dario.herranz.rodrigo@gmail.com (D.H.-R.); joyravedra@hotmail.com (J.Y.)
[3] C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren 2/n, 28040 Madrid, Spain
[4] Department of Prehistory and Archaeology, Humanities Faculty, UNED University, C/Senda del Rey, 7, 28040 Madrid, Spain; jmvr.preh@gmail.com
[5] Institut Català de Paleoecologia Humana I Evolució Social (IPHES), Zona Educacional 4, Campus Sescelades URV (Edifici W3), 43700 Tarragona, Spain; rhuguet@iphes.cat
[6] Department d'Historia i Historia de l'Art, Universitat Rovira i Virgili (URV), Avinguda de Catalunya 35, 43002 Tarragona, Spain
[7] Unit Associated to CSIC, Departamento de Paleobiologia, Museo de Ciencias Naturales, Calle José Gutiérrez Abascal, s/n, 28006 Madrid, Spain
[8] Zoology Unit, Department of Biology, Autónoma University of Madrid, C/Darwin 2, Campus Universitario de Cantoblanco, 28049 Madrid, Spain; isabel.barja@uam.es
[9] Center of Investigation in Biodiversity and Global Change (CIBC-UAM), Universidad Autónoma de Madrid, 28049 Madrid, Spain
[10] Department of Topographic and Cartography Engineering, Higher Technical School of Engineers in Topography, Geodesy and Cartography, Universidad Politécnica de Madrid, Mercator 2, 28031 Madrid, Spain
[11] Gran Duque de Alba Institution, Dibutación Provincial de Ávila, Paseo Dos de Mayo, 8, 05001 Ávila, Spain; maxifernandezav@hotmail.com
[12] Department of Sciences of Communication and Sociology, Faculty of Communication Sciences, University Rey Juan Carlos, Camino del Molino, s/n, 28943 Madrid, Spain
* Correspondence: ladc1995@gmail.com; Tel.: +34-633-647-825

check for updates

**Citation:** Courtenay, L.A.; Herranz-Rodrigo, D.; Yravedra, J.; Vázquez-Rodríguez, J.M.; Huguet, R.; Barja, I.; Maté-González, M.Á.; Fernández, M.F.; Muñoz-Nieto, Á.-L.; González-Aguilera, D. 3D Insights into the Effects of Captivity on Wolf Mastication and Their Tooth Marks; Implications in Ecological Studies of Both the Past and Present. *Animals* **2021**, *11*, 2323. https://doi.org/10.3390/ani11082323

Academic Editor: Bertram Ostendorf

Received: 9 July 2021
Accepted: 3 August 2021
Published: 6 August 2021

**Simple Summary:** Recent years have seen major advances in the analysis of carnivore modifications to bone during feeding, based on the integration of 3D modeling and data science techniques, and with special attention being paid to tooth marks. From this perspective, carnivore tooth scores and pits have slowly converted into a protagonist in the identification of the carnivores producing them. The present study confronts the intra-species variability of tooth mark morphologies produced by Iberian wolves, taking into account not only different populations but also whether wild and captive wolves produce different shaped tooth marks. Here we show how, in the case of tooth scores, differences are notable and should thus be treated with caution. Further conclusions reveal that carnivore tooth pits are currently the most diagnostic elements for the study of carnivore feeding traces on bone, pending future studies that compare closely related taxa with sufficient intraspecific variability. In light of this, further investigation into the possible stress captivity may cause on these animals could be of great importance for both the study of past and present. If differences were to exist, these results could implicate a larger margin of error than previously perceived for some experimental samples, affecting both prehistoric and modern-day ecological studies.

**Abstract:** Human populations have been known to develop complex relationships with large carnivore species throughout time, with evidence of both competition and collaboration to obtain resources throughout the Pleistocene. From this perspective, many archaeological and palaeontological sites present evidence of carnivore modifications to bone. In response to this, specialists in the study of microscopic bone surface modifications have resorted to the use of 3D modeling and data science techniques for the inspection of these elements, reaching novel limits for the discerning of carnivore agencies. The present research analyzes the tooth mark variability produced by multiple Iberian wolf individuals, with the aim of studying how captivity may affect the nature of tooth marks left on bone. In addition to this, four different populations of both wild and captive Iberian wolves are also compared for a more in-depth comparison of intra-species variability. This research statistically shows that large canid tooth pits are the least affected by captivity, while tooth scores appear more superficial when produced by captive wolves. The superficial nature of captive wolf tooth scores is additionally seen to correlate with other metric features, thus influencing overall mark morphologies. In light of this, the present study opens a new dialogue on the reasons behind this, advising caution when using tooth scores for carnivore identification and contemplating how elements such as stress may be affecting the wolves under study.

**Keywords:** wild wolves; captive wolves; tooth marks; geometric morphometrics; 3D modeling; advanced statistics; taphonomy

## 1. Introduction

### *1.1. The Impact of Wolves Both Past and Present*

Wolves are one of the most prominent carnivore taxa of the Northern Hemisphere, appearing in a number of ancient sites as a competitor with humans [1,2]. The genus *Canis* [3] has been most recently estimated to have emerged towards the Messinian stage of the Miocene, dated at approximately 5.7 Ma, with Bayesian inferred Highest Posterior Density (HPD) intervals between 8.5 and 4.0 Ma [4]. The evolution of *Canis* species variants has since had a long history, with the emergence of a wide array of different sized carnivores, from the smaller jackals (*Canis aureus*) to the Dire wolf (*Aenocyon dirus*) of North America. The species *Canis lupus* is considered to have emerged during the Middle Pleistocene, strongly linked with the evolution of species such as *C. mosbachensis, C. etruscus* [5,6], and the recently discovered *Canis borjgali* [7]. Thereafter, the *Canis* genus has had a notable presence in numerous sites throughout the Pleistocene epoch, closely coinciding with the movement of hominin populations across the globe, and in many cases coexisting with other competitors of the Canidae family, such as *Lycaon lycaonoides, Cuon alpinus* and *Canis orcensis* [5,8–11].

Canids have been an important protagonist within the European carnivore guild. To this extent, some authors hypothesize the trophic pressure and competition for resources among these species [10,12–19]. From a similar perspective, being a close competitor with large felids at the time, other studies hypothesize these canids to have had a dynamic role in complex food chains throughout the Pleistocene [1,2]. As of the Upper Palaeolithic, interactions between canid and hominin species begin to change, with possible evidence of domestication and collaboration as early as >30 Kya [20–24]. These interactions become more complex as the increase in human populations induces significant pressure on carnivores. From this perspective, animals such as wolves are frequently targeted by farmers and landowners as a means of protecting their livestock [25–31]. While wolves are typically the first suspects as culprits for livestock predation, formal clues to differentiate between potential predators are lacking (e.g., foxes, dogs), especially in cases where flesh evidence is scarce.

A recent study by Yravedra and colleagues [32] proposes the use of diagnostic criteria typically used in archaeology and paleontology for the study of carnivore activity. From this perspective, these authors propose methods on bone as a means of studying the agents responsible for livestock predation. From this perspective, these authors attempt to unify ecological studies of both past and present with the use of high-resolution technologies for 3D modeling and advanced statistics, as well as artificially intelligent tools. Nevertheless, while this line of research has made important advances in the differentiation between different species, little is known on the intra-species variability that may be conditioning these results. Examples of this could include changes due to sexual dimorphism, age differences, natural inter-individual variability, genetic differences between populations, or sufficient sample sizes that can capture these differences.

In archaeology and paleontology, the discerning of precise carnivore agents is a highly informative piece of information, essentially providing a fundamental means of interpreting carnivore-hominin interactions as well as the general ecology of many sites. In modern-day ecology, the conservation of carnivores is under threat from human activity, with livestock predation being a particular source of tension. Research into the types of damage carnivores can produce on bones can thus be considered an important means of finding diagnostic criteria and can therefore answer a number of different ecological questions (e.g., paleoecological modeling, discerning agents involved in contemporary livestock predation). Nevertheless, prior to any widespread applications, experiments must be performed to ensure that our comparative samples are truly analogous.

In light of this, the present study attempts to analyze the variability of tooth marks in wolf populations, testing for the effect of cofounding variables that had not been considered prior to the present study.

### *1.2. State of the Art in Wolf Tooth Mark Analyses*

Since the beginning of research into the faunal remains of archaeological sites, authors have noted the presence of large carnivore activity [33]. Among the different tracks and traces frequently left by carnivores, tooth marks found on bone have been considered of particular interest [34], providing an indirect trace of carnivore activity in sites where carnivore bones may not necessarily be present. The field of taphonomy is a particular protagonist in this type of research, focusing on the multiple agents and processes that may have modified bones since the life of the organism to the present date. At present, four main types of carnivore bite damage are typically considered [16,34–36]; (1) circular depressions or imprints of the tooth's cusps on bone (tooth pits, Figure 1A); (2) elongated depressions with a rounded base produced by the dragging of teeth across the surface (tooth scores, Figure 1); (3) circular holes which are a product of the tooth penetrating the cortical walls (punctures); and (4) the progressive deletion of large portions of bone by continuous chewing (furrowing).

Some of the first notable efforts in the study of tooth marks were performed by Selvaggio and Wilder [37], consisting of the metric analyses of the length and width of tooth scores and pits. While innovative, these methods present important limitations, discerning the most likely size of the chewing carnivore rather than inferring precise agents. More developed attempts were able to augment the size of available datasets, yet with similar limitations to their predecessors [38,39] (*inter alia*).

With the integration of advanced 3D modeling, novel efforts were able to improve this resolution, making closer approximations to the precise agent involved [40–43]. Despite the great improvement in statistical data processing techniques, these studies still presented margins of error that were hard to overcome. Likewise, while the inclusion of advanced data science techniques such as Computational Learning strategies was able to present a more efficient means of processing this data [44], these studies are still limited in terms of sample size and can only be considered the first step in a promising direction. From this point, more consequent efforts have been made to increase sample sizes with equally promising results [32,45] while also improving the means of extracting this information [46].

Other notable advances in literature have shown how volumetric and micro-topographical information derived from confocal microscopy can provide valuable information [47,48]. Likewise, deep learning-based computer vision techniques have also been proposed for answering a number of these questions [49,50].

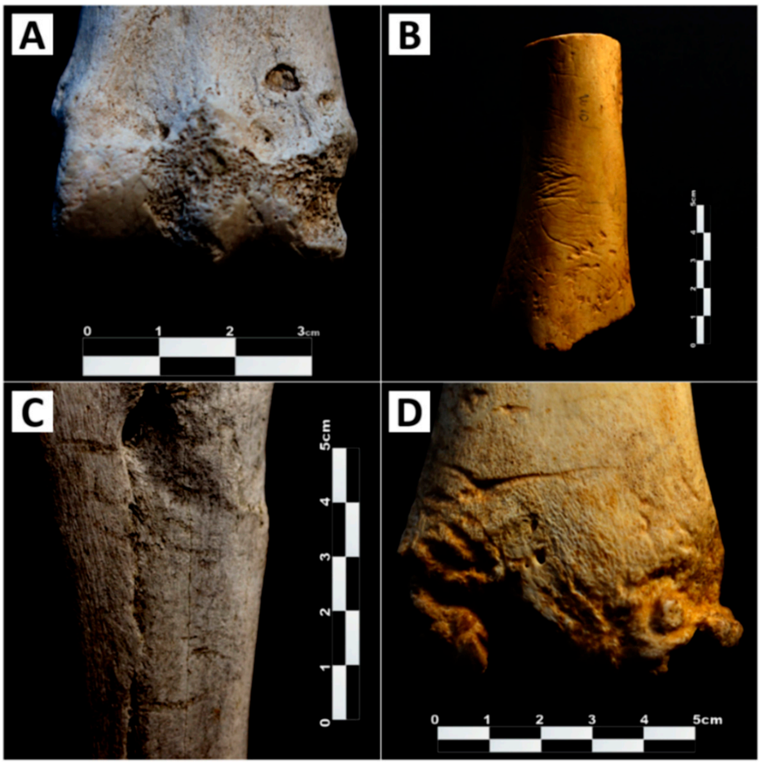


**Figure 1.** Photographic documentation of different types and the extent of bite damage. (**A**) Single isolated tooth pit observed on the distal metadiaphysis of a horse metatarsal from Villardeciervos. (**B**) Tooth-marked bone presenting multiple parallel scores across the diaphysis and pits towards the distal end of a horse humeral shaft from Cabárceno. (**C**) Multiple scores along the shaft of a horse radius-ulna from Villardeciervos. (**D**) Multiple scores and an occasional pit on the distal metadiaphysis of a horse tibia from Cabárceno.

While these studies are hopeful for the future of inter-species analyses, the intra-species perspective has been touched on less [51,52], especially from a morphological perspective [53]. In light of this, important questions must be raised prior to the large-scale application of these methods, such as the possible changes to tooth mark morphologies that may be a product of additional confounding variables. If these variables are found to be important conditioning factors, this may raise important doubts on the applicability of these techniques in real-world applications, on an inter-species as well as an intra-species level. This could essentially alter results in analyses such as the identification of carnivore activities in archaeological sites or the classification of animals responsible for livestock predation.

The present study considers these questions from the standpoint of modifications produced by multiple individuals of the same species; the Iberian wolf. Similarly, considering how obtaining bones modified by wild carnivores can be notably difficult, this study additionally tests to see whether using captive animals in parks is the best analogy for these types of modifications. The present hypotheses to be tested do not consider captivity to be a conditioning factor in tooth mark morphologies, while the difference between wolf populations should also be minimal. Under this premise, four large samples of bones modified by different populations of *Canis lupus signatus* individuals have been used, two originating from parks of varying sizes and the remainder of samples originating from wild wolf packs across the north-western Iberian peninsula.

## 2. Materials and Methods

### *2.1. Samples*

For the purpose of understanding intra-species variability in carnivore populations, four samples of tooth-marked bones produced by *Canis lupus signatus* individuals were included in the present study (Figure 2).

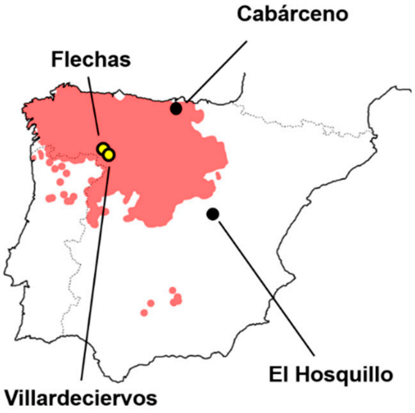


**Figure 2.** Graphic representation showing the distribution of wolves in the Iberian peninsula as well as the location of the wolf tooth mark sample used in the present study, including wild wolf packs (yellow dots) and captive wolves (black dots).

*Canis lupus signatus* [54], commonly known as the Iberian wolf, is a subspecies of wolves populating the northwest of the Iberian peninsula. Wolves are social hunters with a diet predominantly consisting of large and medium-sized game. Wolves additionally present one of the highest potential bite forces for their size when compared with other large carnivores, with an estimated bite force of ≈1200 Newtons using the Carnassial teeth [55]. When chewing, most canids are known to show a preference for the use of teeth furthest back in the mouth, namely the posterior-most inferior molars and upper premolars/molars. A recent study has reported the larger cusp of these teeth to have a mean breadth of 13mm for female wolves and 14mm for male wolves in the case of lower molars, and 16mm for female wolves and 17mm for male wolves in the case of upper pre-molars [56].

Two of the present tooth mark samples originated from wild wolf packs residing in the province of Zamora in north-western Spain, sharing borders with Portugal and located north of the River Duero (Castilla y León, Figure 2). The first of these samples originated from the area of Flechas, while the second from Villardeciervos. Both samples were collected between the months of May and September of 2010, consisting of carcasses from a wide range of different animals, including equids (*Equus ferus*), red deer (*Cervus elaphus*), wild boars (*Sus scrofa*), and roe deer (*Capreolus capreolus*). Control of whether only wolves had intervened in samples was based on the current knowledge about the ecology of the area. In both samples, all anatomical elements were present (including the cranial, axial, and appendicular skeleton); however, only tooth marks originating from the diaphyses and meta diaphyses of appendicular long bones were considered. From Flechas, a total of 55 appendicular bone remains were recovered, from which a total of 63 tooth scores and 49 pits were used. From Villardeciervos, a total of 124 appendicular elements were recovered, from which a total of 56 scores and 79 pits were used. This results in a total sample size of 128 wild wolf tooth pits and 119 scores.

Tooth mark samples from captive wolves were obtained from two separate parks, including Cabárceno (Obregón, Cantabria, Figure 2) and Hosquillo (Cuenca, Castilla-La Mancha, Figure 2). Samples were collected during the winter months of 2010 and 2011. The first of these samples have been previously included in publications including [32,41,44–46,50,53,57], while the latter by [45,53,58]. Both samples were produced by groups of adult individuals, seven individuals in the case of Cabárceno and five in the case of Hosquillo. Animals were fed disarticulated limb elements with meat attached, while some occasional distal epiphyses of humeri and femora have been noted as still articulated with zygopodials. Bones were exposed to animals for varying periods of time, as established by the rules and conditions of each park, with Cabárceno wolves having access to the remains for a single week and Hosquillo individuals for three months.

The Hosquillo sample consists of mainly medium and small-sized animals, including mouflon (*Ovis musimon*), Iberian ibex (*Capra pyrenaica*), roe deer (*C. capreolus*), and wild boar (*S. scrofa*). The Cabárceno sample consists exclusively of large-sized animals; namely, equid (*E. ferus*) and some bovid (*Bos taurus*) remains. Once again, only tooth marks observed on diaphyses or meta diaphyses of appendicular long bone elements were considered. From Hosquillo, a total of 420 remains of highly fractured appendicular elements (femora, humeri, tibiae, and radii) were recovered, from which a total of 113 tooth scores and 113 pits were used. From Cabárceno, a total of 28 appendicular elements were recovered, from which a total of 56 scores and 42 pits were used. This resulted in a total of 169 scores and 155 pits from captive animals.

Both parks are dedicated to the conservation and investigation of different species, some of which are endangered, and keeping animals in open-air enclosures while also being open to the public for educational purposes. Nevertheless, in the interest of transparency it is important to point out that, while previous publications have referred to these samples as a product of semi-captive wolves, after careful evaluation and consideration of how wolves in the wild typically inhabit a large and variable territory with high mobility (≈10 km per day, with great variability among some reports [59–66], *inter alia*), we considered that the enclosures presented in each of these parks were too small to consider these animals as anything other than captive. Personal communications by ground keepers have reported the wolf enclosure in Cabárceno to have an extension of 2700m$^2$, approximately 0.04% of the total area (740 ha) of the Cabárceno natural park, while the Hosquillo individuals have a slightly larger enclosure measuring 10,000 m$^2$, approximately 0.1% of the 910 ha of the Hosquillo natural park. The Cabárceno natural park is a popular year-round destination housing a multitude of species, with approximately 600,000 visitors per year. The Hosquillo natural park, on the other hand, is considerably smaller, with a reported 14,426 visitors during the year 2015 and 20,000 in the year 2018, making it 97 to 98% smaller than that of Cabárceno. In the interest of understanding the behavior of these animals, we also consider it important to note that the Cabárceno natural park is accessible almost exclusively by motor vehicles, implicating that animals are surrounded by additional noise produced by cars on a daily basis. Hosquillo, on the other hand, while also being accessible by motor vehicles, provides a larger distance between the asphalt tracks and each enclosure in some areas while only being open to the public on weekends or festive days, with week days generally being reserved for scholarly activities, thus ensuring less exposure of the wolves to the public.

Finally, with regards to the Hosquillo and Cabárceno samples, wolves from both parks have been estimated to be five years old, with little difference in age between wolves from these parks.

While many more tooth marks of varying types were observed on these samples, the selected tooth marks were chosen on the basis of being found on diaphyses; elements more likely to survive extensive carnivore damage, as well as any other taphonomic processes frequently encountered in the fossil register. Marks were chosen based on the clarity of their micro-topography, while only isolated marks with no overlapping of traces were selected. Inconspicuous marks that would be difficult to model via remote-sensing techniques were also excluded.

All experiments involving carnivores were performed in accordance with the relevant guidelines as set forth by park keepers and general park regulations. No animals were sacrificed specifically for the purpose of these experiments. Likewise, carnivores were not manipulated or handled at any point throughout the collection of samples. In each of the parks, collection of chewed bones was performed directly by park staff, assisted by one of the authors (J.Y.). No licenses or permits were required to perform these experiments. In the case of Cabárceno and Hosquillo, bone samples were provided directly by the park in accordance with their standardized feeding protocols. For wild wolves, carcasses were collected by one of the authors (J.Y.), additionally aided by forest rangers where necessary. Once collected, all bone samples were cleaned in boiling water without the use of additional chemical agents.

Photographic examples of tooth pits and tooth scores from different samples have been included in Figure 1.

### *2.2. Data Collection*

Tooth marks were digitized using the DAVID SLS-1 Structured-Light Surface Scanner located at the TIDOP Research Group of the Polytechnic School of Ávila (University of Salamanca, Spain), identical to the methods described in [46,53,67]. Once 3D models had been obtained for each of the tooth marks, marks were treated differently according to their type (Figure 3), following the methodological approaches by Yravedra et al. [40] with slight adaptations by Courtenay et al. [53] in the case of tooth scores, and Courtenay et al. [46] in the case of tooth pits.

Tooth scores were analyzed according to the morphology of their cross-sections (Figure 3), extracted at the mid-point of each score. For this, 2D data were derived from 3D models via the calculation of digital elevation models for micro-topographies. Cross-sections were thus extracted using the Global Mapper v.18 Geographic Information System (GIS) software, obtained at mid-length of each tooth score between ≈30% and ≈70% of the mark's total length. Cross-section profiles were then exported as images for further processing in the tpsDig2 (v.2.1.7) software [68].

For each cross-section, 7-landmark 2D coordinates (x, y) are extracted, following the model proposed by Yravedra et al. [40]. Landmarks 1 and 7 (LM1 & LM7) mark the left and right shoulder of the scores' cross-section, respectively, while LM4 marks the deepest-most point. LM2, LM3, LM5, and LM6 traditionally mark varying points along the marks' walls. Nevertheless, these have recently been replaced by computational landmarks so as to avoid analyst subjectivity [53]. These computational points are therefore calculated at equidistant intervals by the tpsDig2 software: between LM1/LM4 for the left wall and LM4/LM7 for the right wall. These landmark coordinates can then be used to calculate the metric dimensions of each profile, inspired by the methods of Bello and Soligo [69], later adapted by Maté-González et al. [70] and Yravedra et al. [40]. From this perspective, the distance between LM1 and LM7 is used to calculate the Width of the Incision at the Surface (WIS) of the bone, while distances between LM2/LM6 and LM3/LM5 can be used to define the Width of the Incision Midway (WIM) and the Width of the Incision in proximity with the Base (WIB), respectively. The Depth of the mark (D) can then be defined taking the perpendicular distance between LM4 and the plane between LM1/LM7, while the distance between LM1/LM4 and LM7/LM4 is used to calculate the Left and Right Depth of the incision at Convergent (LDC and RDC). All the aforementioned measurements were recorded in millimeters. Finally, the Opening Angle (OA) of the mark is calculated in degrees (°) by triangulating LM1, LM4, and LM7, followed by calculating the interior angle of the LM4 corner.

For further analysis, each of the 2D landmark coordinates was also exported and formatted into morphologika files for the purpose of Geometric Morphometric statistical studies.

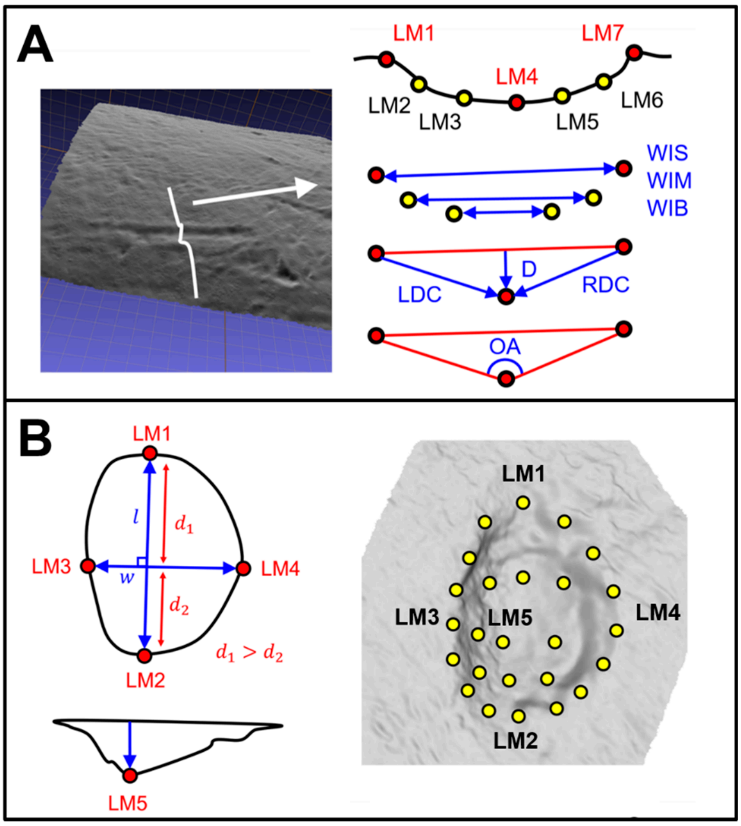


**Figure 3.** Visual description of landmark coordinate positions and the derived measurements. (**A**): 2D 7-Landmark model proposed by Yravedra et al. [40] with adaptations for the inclusion of equidistant computational landmarks by Courtenay et al. [53]. From the 7 landmarks, an additional 7 measurements can be derived, including the Width of Incision at Surface (WIS), Midway (WIM) and in proximity with the Base (WIB), alongside Depth (D), Left (LDC) and Right (RDC) Depth at Convergent and finally Opening Angle (OA). (**B**) 3D 30-Landmark model proposed by Courtenay et al. [46] with 5 fixed landmarks (Red) and a 5 × 5 computed landmark patch (yellow). The positioning of the fixed landmarks is dependent on the perpendicular axes that mark the maximum length ($l$) and width ($w$) of the pit, with Landmark 1 (LM1) being positioning the furthest away from $w$ (distance 1, $d_1$, must be greater than distance 2, $d_2$).

For tooth pits, the entire 3D morphology of marks was analyzed (Figure 3), using the 3D 30-Landmark model proposed by Courtenay et al. [46]. This landmark configuration consists of five fixed Type II landmarks and a 5 × 5 patch of computational landmarks. Of the five fixed landmarks, LM1 and LM2 mark the maximal length of each pit, while LM3 and LM4 are used to define the pit's width. For correct orientation of the pit, LM1 can be considered to be the point along the maximum length furthest away from the perpendicular axis marking the maximum width. LM2 is thus the point marking the other extremity of the pit's length. LM3 and LM4 then mark the extremities of the perpendicular axis, with LM3 marking the left-most point of the maximum width and LM4 marking the right-most point. LM5 is finally defined as the deepest point of the pit. The 5 × 5 computational landmark patch is then positioned over the entirety of the pit so as to capture the internal morphology of the mark. For tooth pits, landmark data collection was performed using the free IDAV Landmark Editor software (v.3.0.0.6), carried out by a single experienced analyst in accordance with the instructional video and Supplementary Appendix provided by the original publication of the landmark model [46].

Finally, attempts to adapt and extrapolate the 3D-based methods to the analysis of tooth score morphologies were tried and tested to see whether the resolution could be improved. Nevertheless, these experiments yielded limited results (Supplementary File

S2; Table S2.1) due to the wide variability of tooth score micro-topographies. Under this premise, only the 2D-based method was used for tooth score analysis (See Supplementary File S2).

### *2.3. Statistical Analyses*

#### 2.3.1. Metric Analyses

Metric data were subject to both univariate and multivariate analyses. Tests additionally were performed, separating the variable OA from other linear metrics (WIS, WIM, WIB, LDC, RDC & D).

Prior to any comparative hypotheses tests, homogeneity of data distributions were checked via multiple Shapiro–Wilks tests [71]. The hypotheses tests employed were then dependent on Shapiro–Wilk results, employing parametric tests when Gaussian distributions were detected and non-parametric tests upon rejecting this assumption. In each of the univariate cases, robust and traditional descriptive statistics were then employed [46,72–74], using the mean and median to report Gaussian and non-Gaussian central tendencies, followed by the first standard deviation or the square root of the Biweight Midvariance (BWMV; Supplementary File S3, Equations (S3.1)–(S3.3)). Two-One Sided equivalency Tests (TOST) were then employed to test for the magnitude of equivalency between samples according to Cohen's *d* [75]. For parametric versions of TOST, Welch's *t*-statistic was used [76]. In cases where non-parametric approaches were required, Yuen's trimmed robust *t*-statistic was used [77,78]. For ease of differentiating between the two, from this point onward, non-parametric robust TOST is referred to as rTOST. It is important to note that, contrary to many other analyses of variance, both variants of TOST consider the Null Hypothesis ($H_0$) to indicate differences between samples [79]. Equivalency testing was performed using the equivalence (v.0.7.2) R library.

In the case of circular metrics, a different approach was used, taking into consideration the trigonometric properties of angular data. Firstly, it is important to point out that on all accounts, the variable OA fits a *Von Mises* Distribution [80–82], thus conditioning the selection of descriptive and hypothesis tests employed (see Supplementary File S1). Under this premise, descriptive statistics included standardized kurtosis and skewness values [83], alongside sample circular variance as an evaluation of the symmetry of circular distributions. For the assessment of circular "normality", two additional tests were considered. These included analyses of uniformity, using the Rayleigh test [84], and symmetry, using a robust reflective symmetry test [85]. Uniformity and symmetry consider the distribution around the entire circle; it is not the same to have a slight amount of skew as having an overall symmetrical distribution across the entire circle. From this perspective, the combination of these tests checks to see whether the distortion for each distribution is of importance.

Depending on the relative symmetry of the circular distribution, the central tendency $H_0$ was robustly calculated via either the mean (*θ-bar*) or median (*θ-tilde*) angle. Prior to any of these calculations, angles were converted into radians, and converted back to degrees only for the purpose of reporting the final results.

For two-sample hypothesis testing, three different approaches were used to assess differences and similarities between samples. These tests considered; (1) differences in mean, using a bootstrapped alternative to the Watson test [86]; (2) differences in median, using the randomized variant of Fisher's non-parametric test [87]; and (3) differences in distributions, using the randomized variant of the Mardia–Whatson–Wheeler test [83,88]. In-depth descriptions of these analyses and the reasons behind their adoption can be consulted in the Supplementary File S1.

For multivariate analyses of these metric variables, dimensionality reduction via Principal Components Analyses (PCA) was performed. For PCA, the variable OA was combined with linear variables WIS, WIM, WIB, LDC, RDC, and D by prior transformations of OA into a new *xy* variable. This was carried out by first converting OA into radians, followed by the sum of linear transformations for each of these radian angles using cosine and sine calculations ($cos\theta + sin\theta$, see Supplementary File S1). This approach was chosen in light of the considerable decrease in dimensionality reduction error produced when including this transformed version of OA in PCA (Supplementary File S1, Figure S1.9). TOST and rTOST tests were then performed across the PC Scores to test for multivariate equivalence, using the top ranking PC scores explaining up to 95% of total sample variance.

Considering the frequent non-linear relationships between variables, PCA was complemented by non-linear dimensionality reduction technique to observe for possible differences in conclusions. For this purpose, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm was chosen, a popular technique used in Machine Learning [89]. t-SNE was performed using random initialization directly on each of the extracted measurements. This algorithm was trained for 500 iterations, using a perplexity parameter calculated as the ceiling of the square root of the sample size (n, i.e., $\lceil\sqrt{n}\rceil$). t-SNE transformations projected data into a new $\mathbb{R}^3$ feature space. t-SNE was performed using the Rtsne (v.0.15) R library.

Where considered complementary, two additional tests were performed. The first of these considered correlations between variables to better understand relationships metrics. For homogenous data, the parametric Pearson test was used [90], whereas inhomogeneous data were tested using the non-parametric Kendall τ rank-based test [91]. Finally, Bayesian Inferred effect sizes were also calculated according to Cohen's δ [75], with the use of 95% High Posterior Density calculations on the posterior distribution. These were additionally accompanied by Probability of Superiority (PS) metrics. For Bayesian calculations, a Student's t-distribution was used to infer the data's distribution [92] robustly. For this analysis, the No-U-Turn (NUTS) extension of the Hamiltonian Markov Chain Monte Carlo (MCMC) algorithm was employed for sampling [93], as implemented in the PyMC3 (v.3.11) library of the Python (v.3.7.4) programming language. NUTS was fit using 4 MCMC chains, sampling using 1000 tuning steps and 5000 draw iterations. To avoid severe divergences of the NUTS sampling algorithm, a sampling step size of 0.9 was employed. This parameterization resulted in only three divergences throughout all samples (observed in the case of the variable D for wild wolves), an Effective Sample Size of >4000, good mixing across all trace plots, and an R-hat statistic of 1.0 for all parameters. Further details on the Bayes models are included in Supplementary File S4, including equations (Equations (S4.1)–(S4.4)) and a visual representation of the models in the form of a Kruschke diagram (Figure S4.1) [94].

Unless stated otherwise, all statistical applications were performed using the R programming language (v.3.5.3) and multiple packages.

### 2.3.2. Geometric Morphometric Analyses

Geometric Morphometric analyses were performed first, including an orthogonal tangent projection and full Procrustes fit of landmark data [95]. This is a common technique in morphological analyses for data preparation and standardization. This process, frequently referred to as Generalized Procrustes Analysis (GPA), consists of multiple superimposition procedures (translation, rotation, and scaling) that aid in the quantification and visualization of minute displacements of individual landmarks in space. Nevertheless, considering observations made by Courtenay et al. [44–46,53], which reveal tooth mark size to be an important conditioning factor in morphological variance, GPA was then subjected to multiple allometric analyses to test for shape-size relationships [96]. For allometric analyses, multiple regressions were performed using the logarithm of centroid size to estimate shape, testing for the goodness-of-fit to conclude whether (1) shape is notably affected by size but also (2) to see if differences in shape-size relationships for different samples are present.

In cases where allometric relationships were concluded to be important, further geometric morphometric analyses were performed, excluding the scaling process of GPA. When allometry proved to be unimportant, a full GPA, including the scaling process, was used. This exclusion of the scaling process is often referred to as the analysis of *form*, while the inclusion of scaling during GPA is referred to as analysis of *shape* [97–99].

Following GPA, dimensionality reduction in the form of PCA was performed to convert landmark coordinates into a more manageable format. The most important Principal Component Scores (PC Scores) were then extracted for multivariate testing. Across these PC Scores, TOST and rTOST tests were carried out, alongside the calculation of transformation grids and Thin-Plate Splines (TPS) for the visualization of morphological changes [100].

Finally, considering the frequently non-linear relationships that have frequently been observed to emerge among taphonomic geometric morphometric data [46], PCA was also complemented by the use of t-SNE. In this case, t-SNE was trained directly on the superimposed landmark coordinates.

For all Geometric Morphometric analyses, the R programming language (v.3.5.3) was used, primarily employing the geomorph (v.3.3.1) and shapes (v.1.2.4) R libraries.

#### 2.3.3. Hypothesis Testing

In accordance with the recommendations set forth by the editors and contributors of the *American Statistician*, $p$-values were not evaluated using $p < 0.05$ as a threshold for defining statistical significance [101,102]. Likewise, the term "significant" has been avoided throughout the present study. In its place, all hypotheses testing was performed using $p$-value to Bayes Factor Bound (BFB) calibrations in accordance with the recommendations of Benjamin and Berger [103], as well as the calculation of False Positive Risk (FPR) values as suggested by Colquhoun [104]. BFB values (Supplementary File S5; Equations (S5.1) and (S5.2)) represent "the strongest case for the alternative hypothesis [$H_a$] relative to the null hypothesis [$H_0$]" [103], in other words, the odds at most of $H_a$ being true. These values can then be used to derive the final posterior odds by multiplying BFB with the prior odds. FPR (Equation (S5.3)), on the other hand, is the probability that an observed $p$-value is a false positive, otherwise known as a Type I error.

To avoid the inclusion of large strings of numbers in the main text reported values have been mostly limited to the inclusion of a single probability metric in support of the Null Hypothesis ($H_0$), referred to here as the Probability of $H_0$, or $p(H_0)$, of the probability of error if $H_a$ were to be true (Equation (S5.5)). This can be calculated using a combination of FPR and an inverse function (Equation (S5.4)). Considering how Courtenay et al. [45] found the point of maximum curvature of calibration curves to be at $p = 0.3681$, $p(H_0)$ was defined here using this value as the optimal limit in Equation (S5.4).

All formulae used for these calculations have been reported in Supplementary File S5 (Equations (S5.1)–(S5.3)), alongside additional calibration tables (Tables S5.1–S5.4) and curves (Figure S5.1). Precise BFB and FPR values for a selection of $p$-values can be consulted in Tables S5.1–S5.4. Noting the poor quality of $p < 0.05$ as a boundary for strong evidence of $H_a$, the present study chose to adapt Fisher's [105] definition of the second standard deviation from the mean (i.e., 0.05), extending this to the third standard deviation from the mean ($100\% - 99.7\% = 0.3\%$ or 0.003). Finally, unless specified otherwise, all frequentist to Bayesian calibrations have been performed using a prior probability indicative of complete randomness (0.5), as suggested by Colquhoun [104].

For more details on the present use of BFB and FPR values, consult [103], ref [104], or the summary reported in Supplementary Appendix 3 of Courtenay et al. [45]. For more details on $p(H_0)$ consult Courtenay et al. [106]. The code used to perform these calculations can be obtained from https://github.com/LACourtenay/HyperSkinCare_Statistics (accessed on 05/08/2021) [106].

## 3. Results

### *3.1. Morphometric Analyses*

#### 3.1.1. Analyses of Opening Angles (OA)

Opening Angles presented a mixture of "normally" and "abnormally" angular distributions (Table 1); furthermore, wide variability was observed, with captive wolves presenting more obtuse ($\theta$-*tilde* = 151°) and variable ($v = 0.03$) tooth scores than wild wolves.

**Table 1.** Descriptive data for tooth score opening angles. *k-hat* = standardized kurtosis, *s-hat* = standardized skew, $v$ = Sample circular variance, t = test statistic, $p$ = $p$-value. Bold typeface indicates samples where robust statistical measurements have been used. More details about reported values can be found in Supplementary Materials. $p(H_0)$ values have been excluded from the present table due to space restrictions. General calibrations for these values can thus be consulted in Supplementary File S5.

| Sample | Min [1] | *k-Hat* | *s-Hat* | $v$ | Uniformity | | Symmetry | | Central [1,2] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | t | $p$ | t | $p$ | | |
| Cabárceno | 120.73 | 2.28 | 2.28 | 0.02 | 0.98 | $1.1 \times 10^{-23}$ | 1.63 | 0.110 | 158.67 | 176.39 |
| Hosquillo | 112.00 | 0.25 | 1.06 | 0.02 | 0.98 | $1.2 \times 10^{-46}$ | 1.77 | 0.079 | 147.09 | 175.12 |
| Flechas [3] | 101.65 | −1.31 | −0.14 | 0.03 | 0.97 | $3.1 \times 10^{-26}$ | 0.24 | 0.810 | 132.51 | 160.28 |
| Villardeciervos [3] | 110.80 | −1.49 | 0.41 | 0.01 | 0.99 | $6.3 \times 10^{-24}$ | 0.75 | 0.447 | 132.19 | 149.96 |
| Captive | 112.00 | 0.27 | 1.32 | 0.03 | 0.97 | $3.2 \times 10^{-69}$ | 2.60 | 0.007 | 151.40 | 176.39 |
| Wild | 101.65 | −0.96 | −0.03 | 0.02 | 0.98 | $2.0 \times 10^{-49}$ | 0.06 | 0.953 | 132.36 | 160.28 |

[1] Values reported in degrees (°). [2] Central tendencies are measured as a mean (θ-bar) or a median (θ-tilde) depending on whether non-robust or robust statistical measurements were used. [3] Wild animal samples.

General analyses of differences and similarities among OA results show relatively clear differences between most samples (Table 2), with the only exception being comparisons between both wild wolf samples. The greatest magnitude of differences, however, when considering both calculations for mean, median and overall distribution, are found between wild and captive wolves (test-statistics > 69.7, $p$ < 0.0001, $p(H_0)$ < 0.25%), especially in the case of the Cabárceno captive and Villardeciervos wild wolf samples (test-statistics > 55.3, $p$ < 0.0001, $p(H_0)$ < 0.25%).

**Table 2.** Statistical comparisons of Opening Angles between samples testing for common mean, median, and distribution. $Y_g$, $P_g$, and $W_g$ are the test statistics for the corresponding $p$-values ($p$). Probability in favor of the Null-Hypothesis $p(H_0)$ has also been included. Samples marked with * were produced by wild wolves. Bold typeface indicates samples where measurements do not follow a Gaussian distribution and have thus been described using robust statistical measurements.

| | | Mean | | | Median | | | Distribution | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Sample 1** | **Sample 2** | $Y_g$ | $p$ | $p(H_0)$ | $P_g$ | $p$ | $p(H_0)$ | $W_g$ | $p$ | $p(H_0)$ |
| Wild * | Captive | 138.80 | 0.0001 | 0.0025 | 78.53 | 0.0001 | 0.0025 | 69.97 | $7.3 \times 10^{-16}$ | $6.9 \times 10^{-14}$ |
| Cabárceno | Flechas * | 105.98 | 0.0001 | 0.0025 | 50.94 | 0.0001 | 0.0025 | 47.06 | $6.0 \times 10^{-11}$ | $3.8 \times 10^{-09}$ |
| **Cabárceno** | **Hosquillo** | 20.98 | 0.0001 | 0.0025 | 16.99 | 0.0001 | 0.0025 | 12.77 | $1.7 \times 10^{-03}$ | $2.9 \times 10^{-02}$ |
| Cabárceno | Villardeciervos * | 143.00 | 0.0001 | 0.0025 | 55.31 | 0.0001 | 0.0025 | 64.27 | $1.1 \times 10^{-14}$ | $9.6 \times 10^{-13}$ |
| **Flechas *** | **Hosquillo** | 50.69 | 0.0001 | 0.0025 | 29.68 | 0.0001 | 0.0025 | 21.69 | $2.0 \times 10^{-05}$ | $5.9 \times 10^{-04}$ |
| Flechas * | Villardeciervos * | 0.022 | 0.8864 | 0.7749 | 0.22 | 0.6978 | 0.5943 | 4.650 | 0.1000 | 0.3850 |
| Hosquillo | Villardeciervos * | 59.03 | 0.0001 | 0.0025 | 41.36 | 0.0001 | 0.0025 | 34.08 | $4.0 \times 10^{-08}$ | $5.0 \times 10^{-07}$ |

### 3.1.2. Analyses of Measurements

Linear measurement values show highly inhomogeneous distributions on all accounts (W > 0.87, $p < 3.7 \times 10^{-06}$, $p(H_0)$ < 0.01%). Univariate tests reveal important similarities for all metric variables obtained from different wolf populations (rTOST $|d|$ < 0.136, $p$ < 0.003, $p(H_0)$ < 4.5%), with the exception of tooth score depth ($|d| \approx 0.042$, $p \approx 0.658$, $p(H_0) \approx 57.2\%$). The magnitude of these differences in D increases when considering samples according to captivity ($|d|$ = 0.028, $p$ = 0.856, $p(H_0)$ = 73.4%), indicating overall differences in tooth pits to be more likely conditioned by captivity than differences in wolf populations. Moreover, when considering tooth score widths (WIS), similarities are minimal ($|d|$ = 0.127, $p$ = 0.145, $p(H_0)$ = 43.2%).

In light of these observations, and considering each of the descriptive statistics for the samples in Table 3, it can be seen that the largest differences between samples are found in the depth of tooth scores, with captive wolves producing marks approximately 0.4 +/− 0.2 mm shallower than wolves found in the wild. Furthermore, clear relationships can be established between depth and width (Kendall's τ = 0.59, $p < 2.2 \times 10^{-16}$, $p(H_0) < 2.2 \times$

$10^{-12}$%), as well as OA (Kendall's $\tau = -0.5$, $p < 2.2 \times 10^{-16}$, $p(H_0) < 2.2 \times 10^{-12}$%). From this perspective, the marked increase in wild wolf tooth score D values is likely to condition the increased variability and change in tooth score widths (Table 3; Captive $\sqrt{\text{BWMV}}$ = 0.33, Wild $\sqrt{\text{BWMV}}$ = 0.27). Nevertheless, the magnitude of these differences according to Cohen's δ are relatively small for WIS (δ = 0.26, PS = 0.57), with Bayesian inferred differences of 0.08mm in central tendencies (95% HDI = [0.015, 0.017]), while differences are much larger for D (δ = 0.77, PS = 0.71), with differences of 0.048mm in central tendencies (95% HDI = [0.019, 0.020]).

**Table 3.** Descriptive data for measurements extracted from tooth score cross-sections. Measurements are all reported in mm. Bold typeface indicates samples where measurements do not follow a Gaussian distribution.

| Measurement | | Cabárceno | Hosquillo | Flechas [1] | Villardeciervos [1] | Captive | Wild [1] |
|---|---|---|---|---|---|---|---|
| **WIS** | Min. | 0.15 | 0.07 | 0.11 | 0.21 | 0.07 | 0.11 |
| | Central [2] | **0.54** | **0.48** | 0.47 | **0.56** | **0.50** | **0.49** |
| | Deviation [3] | **0.40** | **0.27** | 0.22 | **0.31** | **0.33** | **0.27** |
| | Max. | 1.75 | 1.78 | 1.16 | 1.38 | 1.78 | 0.14 |
| **WIM** | Min. | 0.11 | 0.05 | 0.08 | 0.15 | 0.05 | 0.08 |
| | Central [2] | **0.37** | **0.33** | **0.29** | **0.39** | **0.34** | **0.33** |
| | Deviation [3] | **0.27** | **0.19** | **0.15** | **0.21** | **0.23** | **0.18** |
| | Max. | 1.18 | 1.20 | 0.80 | 0.95 | 1.20 | 0.95 |
| **WIB** | Min. | 0.06 | 0.03 | 0.04 | 0.08 | 0.02 | 0.04 |
| | Central [2] | **0.19** | **0.17** | **0.16** | **0.21** | **0.18** | **0.17** |
| | Deviation [3] | **0.14** | **0.10** | **0.08** | **0.11** | **0.12** | **0.10** |
| | Max. | 0.60 | 0.61 | 0.43 | 0.50 | 0.61 | 0.50 |
| **D** | Min. | 0.00 | 0.01 | 0.02 | 0.04 | 0.00 | 0.02 |
| | Central [2] | **0.05** | **0.07** | **0.09** | **0.12** | **0.07** | **0.11** |
| | Deviation [3] | **0.05** | **0.06** | **0.07** | **0.08** | **0.06** | **0.08** |
| | Max. | 0.24 | 0.31 | 0.31 | 0.37 | 0.31 | 0.37 |
| **RDC** | Min. | 0.08 | 0.04 | 0.06 | 0.12 | 0.04 | 0.06 |
| | Central [2] | **0.28** | **0.26** | **0.24** | **0.31** | **0.26** | **0.27** |
| | Deviation [3] | **0.21** | **0.15** | **0.13** | **0.18** | **0.17** | **0.15** |
| | Max. | 0.90 | 0.91 | 0.65 | 0.76 | 0.91 | 0.76 |
| **LDC** | Min. | 0.08 | 0.04 | 0.06 | 0.11 | 0.04 | 0.06 |
| | Central [2] | **0.29** | **0.26** | **0.24** | **0.31** | **0.27** | **0.26** |
| | Deviation [3] | **0.21** | **0.16** | **0.13** | **0.17** | **0.17** | **0.15** |
| | Max. | 0.90 | 0.92 | 0.65 | 0.78 | 0.92 | 0.78 |

[1] Wild animal samples. [2] Central tendencies are measured as mean or median for Gaussian and non-Gaussian distributed data, respectively. [3] Central tendencies are measured as the standard deviation or square root of the biweight mid variance for Gaussian and non-Gaussian distributed data, respectively.

Coupled with previous observations regarding the notably larger opening angles of these scores (Table 1), it can be predicted prior to any multivariate or geometric morphometric tests that the overall morphology of tooth scores will be different between wild and captive wolves.

When combining all 7 variables multivariately, both PCA and t-SNE (Figure 4) show clear separations between wild and captive wolves, while populations intermingle with no clear patterns. t-SNE results show high clustering of groups, with occasional overlap across all three dimensions. When considering PCA, the first PC Score (PC1) representing 81.5% of sample variance and is represented primarily by variables WIS, WIM, WIB, LDC y RDC, while the second component is strongly conditioned by the prior-mentioned patterns between D and OA, representing 17.6% of the variance. PCs 3 through to 7, however, are mostly residual (Cumulative variance = 0.9%). rTOST results strongly confirm previous observations, highlighting the greatest magnitude of differences to occur across PC2 (D & OA; $|d| = 0.85$, $p = 0.99$, $p(H_0) = 97.4$%), while differences in tooth score width measurements across PC1 are much smaller, yet still fairly conclusive ($|d| = 0.15$, $p = 0.146$, $p(H_0) = 43.3$%).

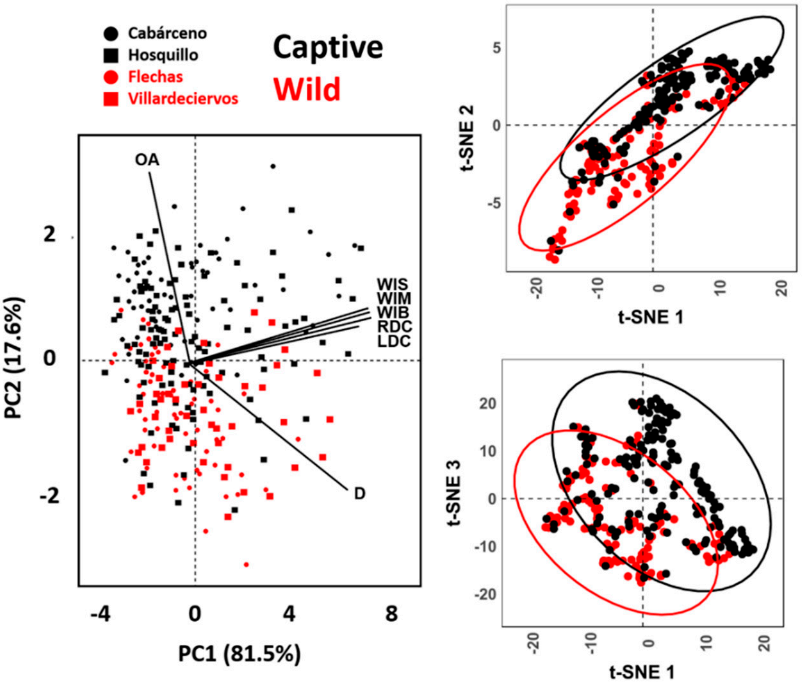


**Figure 4.** PCA scatter-biplots and t-SNE scatterplots performing dimensionality reduction on metric variables obtained from tooth scores.

## 3.2. Geometric Morphometrics

### 3.2.1. Allometric Analyses

Shape-size relationships and simple allometric patterns reveal fairly strong tendencies for size to influence morphology in the case of pits (F = 2.8, $p$ = 0.009, $p(H_0)$ = 10.3%), while scores tend towards the contrary (F = 2.627, $p$ = 0.082, BFB = 1.79, $p(H_0)$ = 35.8%) (Figure 5). Nevertheless, variations in results begin to appear when calculating differences across samples according to groupings.

Scores present inconclusive shape-size relationships when considering variables such as captivity (F = 0.23, $p$ = 0.8, $p(H_0)$ = 67.3%), however studies regarding populations argue otherwise (F = 3.15, $p$ = 0.01, $p(H_0)$ = 11.1%). In-depth analysis of tooth-score residual data, however, presents strong deviations from normality (w > 0.87, $p < 2.0 \times 10^{-12}$, $p(H_0) < 1.5 \times 10^{-08}$%), with a high level of standardized residual disbalance across both Euclidean distances and fitted values. This is especially relevant considering residuals when plotted against the first principal component (w = 0.88, $p = 8.9 \times 10^{-15}$, $p(H_0) = 7.8 \times 10^{-11}$%, ∣skewness∣ = 1.4), and predicted values through regression models (w = 0.87, $p = 5.8 \times 10^{-15}$, $p(H_0) = 5.2 \times 10^{-11}$%, ∣skewness∣ = 1.48). Additionally, considering the irregular residual spread with no concentration around the line of best fit (∣kurtosis∣ = 2.5), regressions estimating linear relationships between logarithmic centroid size and morphological variance according to samples are unlikely to detect true relationships efficiently.

While the $p$ = 0.01 for population differences in scores is indicative of allometry, this $p$-Value corresponds to an upper bound on the Bayes factor of 7.98, which combined with diffuse prior odds of 1:2, would imply posterior odds of 3.99 in favor of the alternative hypothesis and an 11.1% chance of being a false positive. Nevertheless, considering the important levels of residuals produced by these models, it can be assumed that some of these relationships are being exaggerated by the parametric nature of the regression model. In order to adjust for this bias, it could be argued that a more conservative prior probability of 3:10 be adopted. Under this premise, the false-positive risk increases to 22.6%, associated with posterior odds of 2.40 in favor of the alternative hypothesis, and thus presenting a corrected $p(H_a)$ value of 77.4% (1- p($H_0$)). From a critical perspective, it can thus be argued that shape-size relationships according to scores be inconclusive based on the present data, especially when withdrawing finite conclusions according to samples.

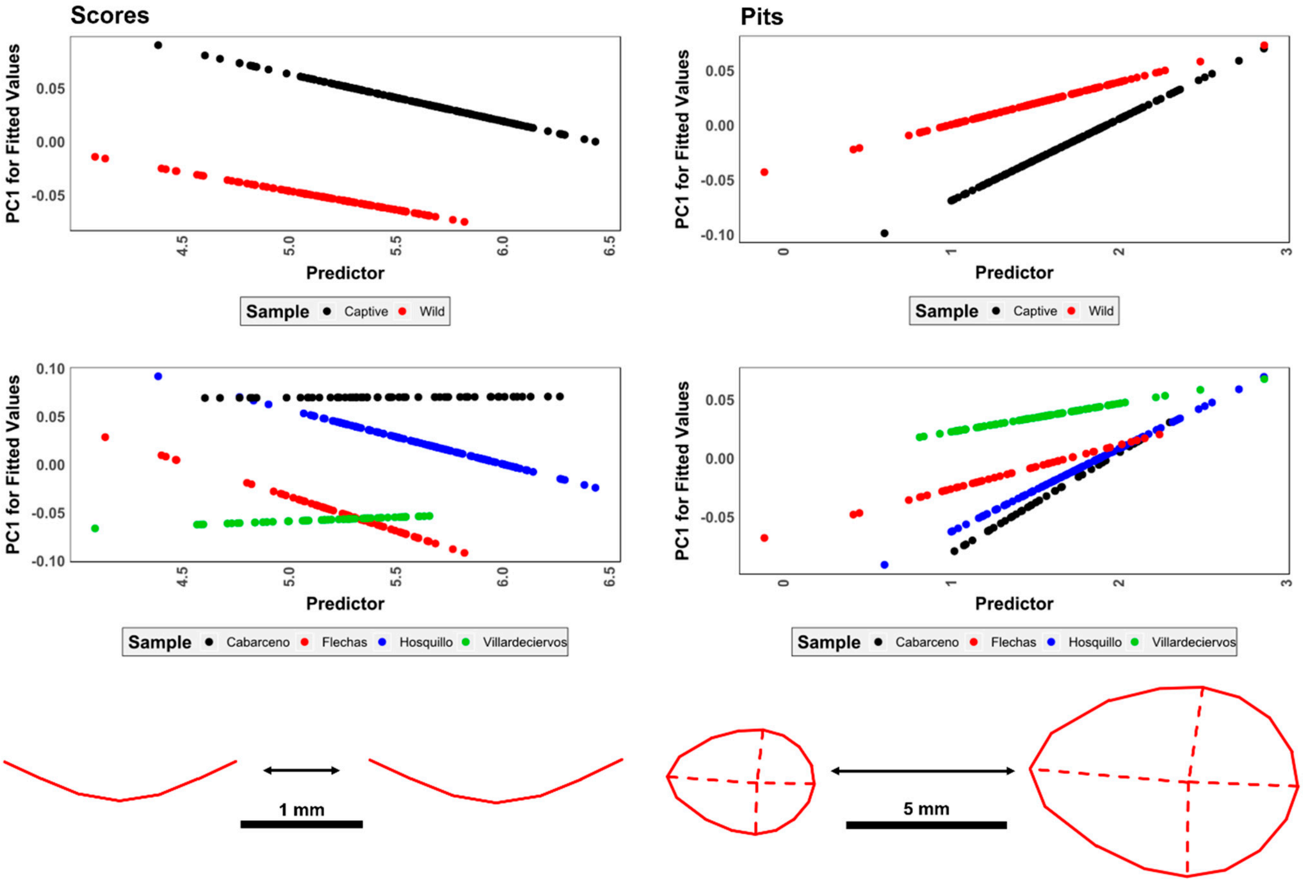


**Figure 5.** Visualization of shape allometry, mapping out tooth mark morphology as a function of size. Predicted values ($\hat{y}$) are sample-specific (Upper panels: captivity vs. wild; lower panels: population origin) in combination with logarithmic centroid size.

When considering pits (Figure 5), unimportant relationships are revealed for both captivity (F = 0.7, $p$ = 0.62, $p(H_0)$ = 55.4%) and population (F = 1.3, $p$ = 0.139, $p(H_0)$ = 42.7%), with highly notable deviations from normality when considering residual data as well (w > 0.83, $p < 2.6 \times 10^{-09}$, $p(H_0) < 1.4 \times 10^{-05}$%). Nevertheless, residual data for pits present a much stronger concentration around the line of best fit (∣kurtosis∣ = {4.55:4.59}), alongside a more even spread across fitted values.

When exploring other possible noise generated within the samples and revisiting the effects observed when studying prey size [53], the present study is able to consolidate the argument that animal size is not the conditioning factor for the morphological variation observed within these samples (Figure 5; F = 1.17, $p$ = 0.25, $p(H_0)$ = 48.5%). Nevertheless, strong allometric correlations do begin to emerge when considering tooth scores from the Cabárceno ($p < 2.2 \times 10^{-16}$, $p(H_0) < 2.2 \times 10^{-12}$%) and Hosquillo ($p < 2.2 \times 10^{-16}$, $p(H_0) < 2.2 \times 10^{-12}$%) samples separately. Under this premise, additional tests were performed using metadata obtained from both parks to analyze the effects of captivity-related stress. For these analyses, stress was modeled considering the number of individuals kept in the space provided by each park as well as the number of individuals in said space. Important correlations were detected ($p < 2.2 \times 10^{-16}$, $p(H_0) < 2.2 \times 10^{-12}$%), resulting in an equally strong effects on morphology (F = 3.1, $p$ = 0.001, $p(H_0)$ = 1.8%). While residuals remain distinguishably high (w = 0.91, $p = 7.8 \times 10^{-09}$, $p(H_0) = 4.0 \times 10^{-05}$%), adopting a rigorous prior probability of 3:10 still reveals only a 4.2% probability of these observations being a false positive, posterior odds of 15.98 in favor of the alternative hypothesis, upper bound Bayes Factor of 53:1 against $H_{0,}$ and thus a corrected $p(H_a)$ of 95.8%.

From this perspective, data from the present study indicate that stress may be an important conditioning factor in tooth score morphology, an observation that warrants further in-depth investigation.

### 3.2.2. Analysis of Variance

When analyzing tooth mark variability in pure morphological shape space, scores continue to appear to present notable differences between wild and captive wolves. This is once again confirmed by very large magnitudes of difference across the first 3 PC scores (96.0% variance, $|d| = 0.15$, $p = 0.99^{\cdot}$, $p(H_0) = 99.99\%$). When considering these differences in accordance with wolf populations, all samples present morphological differences (Table 4).

**Table 4.** Absolute difference ($|d|$), $p$-values ($p$), and probability in favor of the null hypothesis ($p(H_0)$) values obtained from equivalency testing using robust two-one-sided tests on tooth scores.

| | | **Cabárceno** | **Flechas** [1] | **Hosquillo** |
|---|---|---|---|---|
| **Flechas** [1] | $\lvert d \rvert$ | 1.779 | | |
| | $p$ | 1.000 | | |
| | $p(H_0)$ | 0.999 | | |
| **Hosquillo** | $\lvert d \rvert$ | 0.483 | 1.297 | |
| | $p$ | 0.689 | 0.995 | |
| | $p(H_0)$ | 0.589 | 0.986 | |
| **Villardeciervos** [1] | $\lvert d \rvert$ | 2.767 | 0.988 | 2.285 |
| | $p$ | 1.000 | 0.798 | 1.000 |
| | $p(H_0)$ | 1.000 | 0.671 | 1.000 |

[1] Wild animal samples.

Tooth pits, on the other hand, produce PCA graphs of a much different nature, with form space revealing unimportant differences between wild and captive wolves across the first 5 PC scores (86.77% of variance; $|d| = 0.13$, $p = 9.0 \times 10^{-14}$, $p(H_0) = 7.4 \times 10^{-10}\%$). Even if considering a strict prior probability of 3:10, the worst-case scenarios present a probability of $1.7 \times 10^{-11}\%$ of Type I statistical errors. For populations (Table 5), the same can be said throughout comparisons.

**Table 5.** Absolute difference ($|d|$), $p$-values ($p$), and probability in favor of the null hypothesis ($p(H_0)$) values obtained from equivalency testing using robust two-one-sided tests on tooth pits.

| | | **Cabárceno** | **Flechas** [1] | **Hosquillo** |
|---|---|---|---|---|
| **Flechas** [1] | $\lvert d \rvert$ | 0.062 | | |
| | $p$ | $2.5 \times 10^{-11}$ | | |
| | $p(H_0)$ | $1.7 \times 10^{-09}$ | | |
| **Hosquillo** | $\lvert d \rvert$ | 0.013 | 0.050 | |
| | $p$ | $3.2 \times 10^{-18}$ | $4.9 \times 10^{-16}$ | |
| | $p(H_0)$ | $3.5 \times 10^{-16}$ | $4.7 \times 10^{-14}$ | |
| **Villardeciervos** [1] | $\lvert d \rvert$ | 0.034 | 0.028 | 0.021 |
| | $p$ | $6.8 \times 10^{-16}$ | $9.4 \times 10^{-19}$ | $4.4 \times 10^{-26}$ |
| | $p(H_0)$ | $6.5 \times 10^{-14}$ | $1.0 \times 10^{-15}$ | $7.0 \times 10^{-24}$ |

[1] Wild animal samples.

Upon analyzing projections and morphological variations (Figure 6), Thin-Plate Splines (TPS) confirm that tooth scores are greatly represented by a shift from deep grooves in the case of wild wolves (83% of variance) while tooth pits appear much more complex. Tooth pits show high degrees of overlap throughout while revealing general tendencies for both wild and captive wolves to produce between circular and ovular pits on numerous occasions.

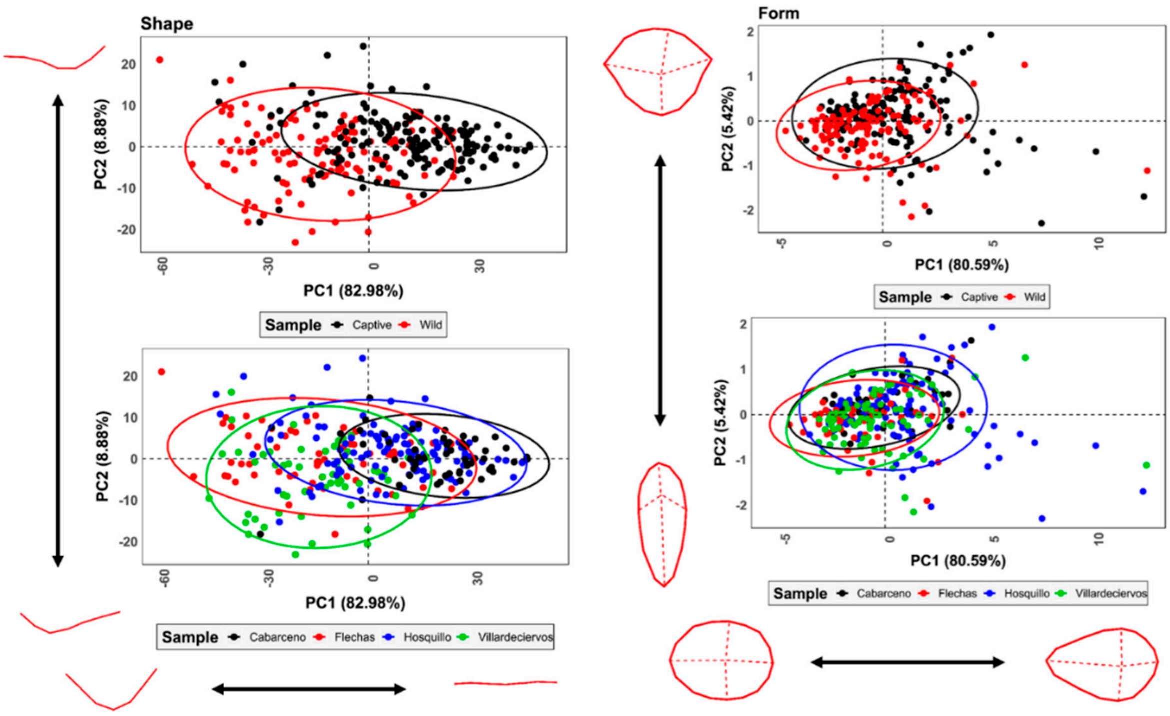


**Figure 6.** PCA scatter plots with 95% confidence intervals presenting variance in tooth score and tooth pit morphology, as represented in shape and form space, respectively. Morphological variance calculated through grid warpings is presented at the extremity of each PC score.

t-SNE results (Figure 7) for both cases similarly show very high clustering of tooth scores according to group labels, while tooth pits show no clear trends.

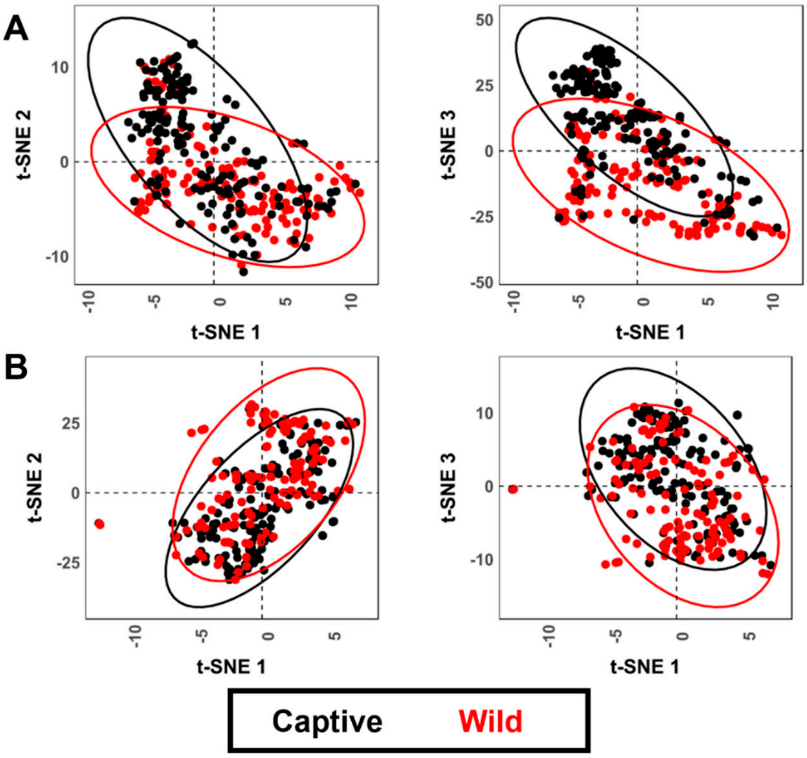


**Figure 7.** t-SNE scatter plots for both tooth (**A**) scores and (**B**) pits when performed analyzing shape and form variables, respectively.

Finally, and in light of observations regarding wolf scores, a detailed analysis of morphological variations was performed across the first 10 PC scores (Figure 8). As can be seen, tooth pits present high morphological variability described by a number of features. Depth of tooth pits does not appear to be a conditioning factor until at least PC8, while only representing 1.44% of morphological variation. While it is true that wild wolves can be seen here to produce deeper pits than their captive relatives, geometric morphometric data highlight tooth pits to be better represented by a number of other morphological factors, as opposed to solely their depth.

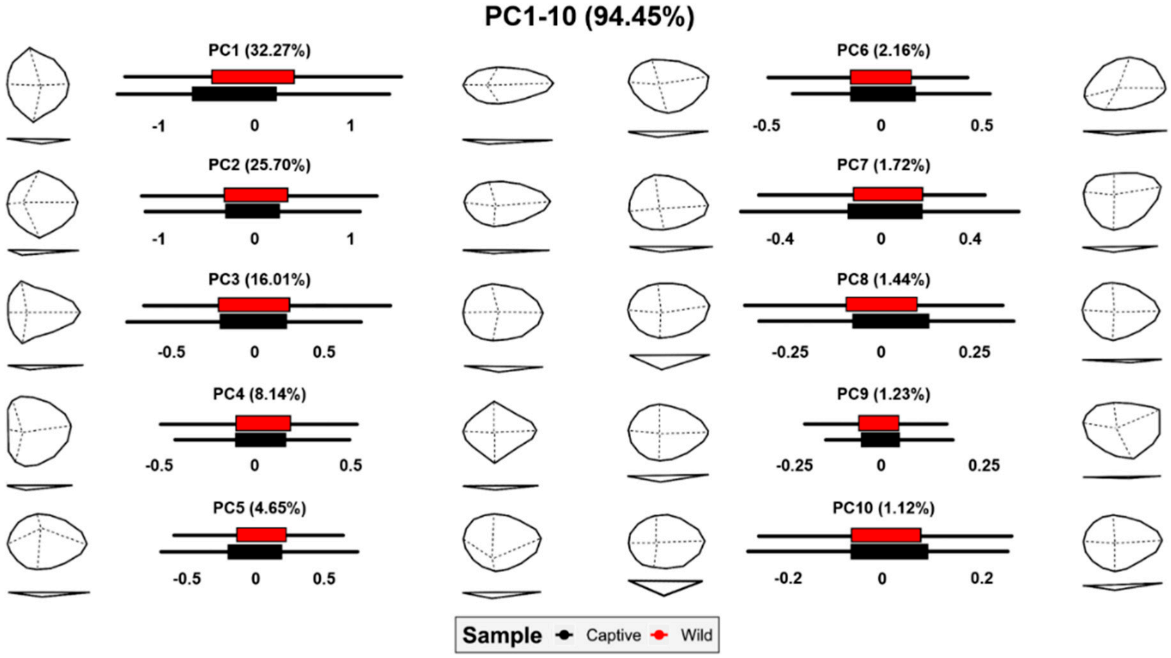


**Figure 8.** Box plots and morphological variance calculated through grid warpings for each of the dimensions in tooth pit PCA.

## 4. Discussion

The conservation of any animal species requires an in-depth understanding of their behavior and ecology. The techniques, therefore, required to analyze animal activities are fundamental. Over the years, archaeologists and paleontologists have developed tools and techniques for the analysis of carnivore activities. While the objectives and research questions from these fields greatly differ from those of modern-day ecological studies, many parallels exist that could benefit from a more transdisciplinary approach to carnivore research.

The present study has shown through both metric and geometric morphometric approaches how wolf tooth scores show important differences when samples are obtained from captive animals. From this perspective, wolf tooth pits have been seen to be more diagnostic elements, less affected by inter-species variability, and still a valuable tool for intra-species analyses [45,46]. The discovery that wolf tooth marks are more superficial among captive wolves raises a number of questions, especially regarding their interpretation. Under this premise, the following sections attempt to describe and contemplate the possible conditioning factors that are behind these observed patterns in variance.

### *4.1. Interpretations behind Morphological Variability*

#### 4.1.1. Physiological Stress of Captivity

Imposing captivity on wild animals has a number of functions. From one perspective, many institutions offer shelter, protection, and health care to species endangered by numerous external environmental factors (most of which are human-induced) and can also serve an educational purpose. In the latter case, this may be for the purpose of investigation while also serving the purpose of providing basic information to the general non-specialized public. Needless to say, captivity can have a major impact on an animal's behavioral, social, or even physiological attributes. One of the key components involved in these processes is known as physiological stress [107].

Stressors may include but are not exclusive to the presence or absence of sensory stimuli (e.g., sound, smell, light, temperature), restrictions of movement (e.g., difficulties to retreat, forced proximity), abnormal social contexts (lack of possibilities for migration, forced husbandry), and the forced imposition of routine (feeding times, type of food, predictable presence or absence of stimuli). Nevertheless, despite the possible combinations of stressors, what is most likely to produce stress among captive animals is their inability to control them [107,108]. In the case of sensory stimuli, most animals, when exposed to uncomfortable situations, have the ability to flee, thus establishing control over the situation. While many zoos and parks housing captive animals pay particular attention to the design of enclosures, the impact of these factors is inevitable [109], especially over prolonged periods of time.

In response to these stressors, captive animals are well known to develop Abnormal Repetitive Behaviours (ARBs). While, in some cases, ARBs are interpreted as a coping mechanism, many specialists argue that this is not necessarily the case [110]. The severity and extent of these ARBs can be dependent on multiple factors; however, they are present in most, if not all, captive species. From one perspective, Mason [111] and Mason et al. [108] state that highly migratory animals are more likely to develop negative ARBs. From a similar perspective, these authors describe how a species's plasticity and socio-cognitive attributes are likely to determine how well these animals cope under stress. Nevertheless, ARBs can not only be seen as an indicator of welfare quality and animal adaptability but, if too extreme, can also put into question the educational objective of these institutions, as these behaviors are not a true reflection of the wild animals' original behavior.

A study by Wells [112] described how correlations exist between the aggressive actions of captive primates and the number of visitors to the zoo. While the social and cognitive attributes of primates may be a poor analogy to that of wolves, wild canids are still openly sociable animals with particular tendencies to interact with, or at least be acutely susceptible to, human-induced stimuli [113]. Likewise, studies carried out on smaller canids show an important change in fox behavior related to the higher peaks in visitors [114]. In each of these cases, the forced proximity with humans is likely to trigger different reactions among animals, which can be perceived as either a threat or a simple annoyance depending on the species [107].

Wild wolves typically modify the extension of their territories in order to avoid other wolf packs but are also frequently found to make particular efforts to avoid humans or human-made structures [66]. A study by Clubb and Mason [115], with the subtitle "Animals that roam over a large territory in the wild do not take kindly to being confined", found correlations between natural territoriality and most ARBs in a number of different animals. From this perspective, and given the reduced amount of "territory size" provided by both the Cabárceno and Hosquillo centers, the size of enclosures and number of visitors are likely to have a joint impact on wolf behavior.

From another perspective, animals tend to eat when they need to, while the added removal of the thrill of the hunt is likely to agitate carnivores [116]. Likewise, the added security of being fed regularly is likely to affect their physiological stress levels.

Finally, chewing is known to calm domestic dogs when they get agitated [117]. Domestic dogs suffering from confinement distress have also been known to chew and destroy items, the latter more common when combined with noise aversion [118,119]. From this perspective, it is a well-known phenomenon that canine species are more prone to noise-related stress, with sounds such as gun-shots, fireworks, people yelling, and heavy traffic causing many dogs to react negatively. Destructive behavior, for example, is one of the most common reactions of dogs in these situations [118]. While many dog species are more accustomed to gun sounds, desensitized to loud noises, and are thus generally less prone to destructive behavior, this is a product of gradual introduction to these experiences, and in some cases, have become a genetic characteristic developed over time [120].

Closely related to some of these points, the lack of new stimuli in most captive environments is a major component in animal temperament. In more general terms, this can be summarized in the simple term "boredom", and is especially relevant in animals of naturally high mobility confined in small spaces [115]. While canids are frequently known to chew on objects to relieve stress, this can sometimes be referred to also as a pass-time and is common, especially in bones found from wild wolf dens or domestic dog yards/kennels [16,35,121]. Analogies of ARBs in taphonomic analyses are commonly known, as in the case study presented by Gidna et al. [51], who found that the number of tooth marks left by captive lions (from Cabárceno, Spain) was almost twice the amount than those left by wild lions (Tarangire, Tanzania). A comparable study by the same authors showed leopards to produce even more extreme differences in the case of captive (Bahari Zoo, Dar es Salaam, Tanzania) and wild (Tarangire, Tanzania) leopards [52]. Likewise, studies by Saladié et al. [122] with captive bears originating from the Barcelona Zoo and Hosquillo Park present much higher tooth mark frequencies than data presented by Sala and Arsuaga [123] and Arilla et al. [124], with wild bears from Cantabria (Spain) and areas of the Spanish Pyrenees. While a full taphonomic analysis of the samples at hand is beyond the scope of the present study, similar observations have been made here (Supplementary Tables S6.1–S6.3).

In each of these examples, evidence points towards captivity being a stressful situation for animals; whether this physiological stress is enhanced by the presence of noise, crowds, lack of space, lack of external stimuli, or a combination of all of the above. In the case of Cabárceno and Hosquillo, both wolf groups are limited to a considerably smaller amount of space than they would experience in the wild. Moreover, this small space is shared between five to seven wolf individuals. Both samples also originate from parks frequently visited by large numbers of people, all of whom carry out their visit by motorized vehicles. Both of these factors are accompanied by the associated noise of both the crowds and the engines. Finally, alongside the established routines and lack of external stimuli, wolves are likely to present signs of boredom, which is commonly associated with excessive chewing.

While it could be hypothesized that the combination of these factors might cause wolves to bite harder, leaving deeper pits, if these continuous chewing behaviors persisted over long periods, this ARB might cause cusps to wear down through excessive use. This would explain the shallower marks. Nevertheless, further research would be required in order to prove this point.

#### 4.1.2. The Biomechanics of Mastication & Additional Reflections on Courtenay et al. "The Effects of Prey Size on Carnivore Tooth Mark Morphologies"

Multiple factors are involved behind the mechanics of mastication, many of which are highly complex, having evolved over millions of years. A previous study by the present authors performed analyses on how tooth mark morphology may be dependent on prey size [53]. Said study statistically concluded that evidence was "insignificant", especially in the case of tooth pits. The discussion of the said study made reference to the influence of skull morphology, tooth morphology, and muscular functions, as well as attributes concerning lifestyle (see discussion of Courtenay et al., [53] and citations therein). While some of these conditioning factors are still to be experimented with, the present study provides a much larger sample size, providing a more empirical means of responding to some of these questions.

First, considering how the aforementioned behavioral attributes behind "playing" and "feeding" are likely to produce different biomechanical movements, it is likely that this is a notable change in the way force is exerted. From another perspective, comments by Toledo–González et al. [56] note that sexual dimorphism is an important component in wolf dental attributes. While controlling the intervention of different sexes in tooth marked samples is difficult, especially in the case of wild wolf samples, the lack of intra-group clusters and detectable patterns in each of the projected feature spaces (Figure 6) is likely to imply that this is not a significant conditioning factor in the case of tooth pits and scores.

In hindsight, and in light of the present study's more rigorous use of $p$-values [103,104], it can be argued that in the case of tooth scores, the probability of these observations being a Type I statistical error falls to 30% for *Large* vs. *Small*-sized prey, 11% for *Large* vs. *Medium*-sized prey, and 8.7% for *Medium* vs. *Small*-sized prey. For tooth pits, Type I statistical error probabilities are between 25 and 4%, with the former calculation being relevant to tooth marks on large-sized animals. From this perspective, the difference between tooth scores on *Large* and *Small*-sized prey as originally reported [53] is likely to be an overestimate. Likewise, tooth pits show larger animals to be more problematic. When paired with a much larger sample here, comparing different populations, these patterns notably decrease. Once again, tooth marks in both wild wolf samples present a lack of separation between different sized animals, with an even smaller probability of being a False Positive for tooth pits (new $p = 3.2 \times 10^{-18}$, FPR = $3.5 \times 10^{-16}$%). Likewise, when compared in a broader context across an additional 144 tooth marks, allometric relationships are revealed to be unimportant when considering prey size (Figure 5; F = 1.17, $p = 0.25$, $p(H_0) = 48.5\%$). From this perspective, the present study can thus consolidate previous hypotheses and conclude with more certainty that an animal's prey size is not a powerful conditioning factor.

### 4.2. Implications for Carnivore Based Research

The present study has provided new insights into the effects captivity may have on typical carnivore modifications to bone, supporting and contributing to the discussions proposed by Gidna et al. [51] and Sala et al. [125]. Here we have found that tooth scores are notably different, while tooth pits are more dependent on other morphological features, leading captivity to be an unimportant conditioning variable. Nevertheless, to what extent are these observations likely to affect other research using carnivore tooth scores?

At present, multiple different techniques exist for the study of carnivore tooth marks, from more traditional metric variables based on simple measurements [37–39], to more complex means of extracting these measurements via 3D models and micro-topographies [40,41,46–48], as well as visual elements extracted from images [49,50]. The present study paid particular attention to how captivity affects variables such as WIS, which can be considered an analogy with any of the metric variables described by authors for the metric study of scores [37–39]. Likewise, the evident changes to score depth, and in turn, the correlated effect this has on WIS, WIM, and WIB variables, surely affects the volumetric properties of these marks [47,48]. From a multivariate approach and considering overall morphology, this has an evident impact on how tooth scores can be studied in geometric morphometric approaches as well [40]. While the present study has not directly considered qualitative components, the aforementioned effects of metric variability are also likely to condition the appearance of these traces in some way or another [49,50].

As noted by Clubb and Mason [115], the ability of a carnivore to adapt to captivity is greatly conditioned by their natural mobility and territoriality in the wild, as well as other factors, including behavioral and physiological attributes [107,108]. Wolves are well known for their high mobility [59–66], and therefore, are more likely to be susceptible to stressors. Nevertheless, the way a wolf deals with this stress may be different from that of other animals. Felids and ursids, for example, have been observed to develop ARBs in the form of extensive pacing [110]. Likewise, these animals are much less durophagous, therefore can be assumed to be less likely to modify bones to the same extent, while truly durophagous animals such as hyaenids have dentition specifically evolved for these types of repetitive forces [126]. This hypothesis, however, is yet to be supported by empirical evidence.

Mindful of all these elements, the present results advise caution when working with tooth scores, especially in the case of wolves. Until further experimentation has been carried out on felids, ursids, and hyaenids, as well as other canids, care must also be taken. Finally, it is important to encourage caution for taphonomists and analysts when publishing samples using the terms "semi-captive", as this term can be considered misleading. The term "semi" should thus consider a number of different factors, including the general size, physiology, and natural behavior of the animal in the wild while factoring in the conditions under which the animals are enclosed.

## 5. Conclusions

The present study started with the hypothesis that, regardless of the population or state of captivity, tooth mark morphologies among wolf populations would not vary. While this hypothesis has been confirmed in the case of tooth pits, tooth scores have presented a notable deviation from our original theories.

The present study has shown a wide array of different perspectives on the effect captivity may have on the morphology of tooth marks. The importance of these observations has an impact on both past and present ecological studies. In conclusion, tooth scores have been shown to present notable variability and should thus be approached with caution when used as a diagnostic tool in tooth-marked assemblages. Nevertheless, pits are less dependent on these variables and are more likely to be a more reliable diagnostic tool for carnivore identification. Future investigation should take into account further interspecific analyses.

**Supplementary Materials:** The following are available online at https://www.mdpi.com/article/10.3390/ani11082323/s1, Supplementary File S1: Circular Statistics and the Processing of Tooth Score Opening Angles, containing; (1) Definitions of Circular Data, (2) Solutions for Multivariate Analyses. Supplementary File S2: Results for 3D Processing of Tooth Scores. Supplementary File S3: Robust Statistical Measures. Supplementary File S4: Bayesian Modelling. Supplementary File S5: *p*-Value Calibrations. Supplementary File S6: Traditional taphonomic data and variables obtained from the tooth marked samples.

**Author Contributions:** Conceptualization, J.Y., R.H., and L.A.C.; methodology, L.A.C. and D.H.-R.; software, L.A.C.; validation, L.A.C. and D.H.-R.; formal analysis, L.A.C. and D.H.-R.; investigation, L.A.C., J.Y., and I.B.; resources, J.Y., I.B., D.G.-A., and J.M.V.-R.; data curation, L.A.C., D.H.-R., J.M.V.-R., and J.Y.; writing—original draft preparation, L.A.C. and J.Y.; writing—review and editing, L.A.C., J.Y., D.H.-R., R.H., and D.G.-A.; visualization, L.A.C.; supervision, D.G.-A. and J.Y.; project administration, D.G.-A. and J.Y.; funding acquisition, D.G.-A., Á.-L.M.-N., M.Á.M.-G. and M.F.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** L.A.C. is funded by the Spanish Ministry of Science, Innovation and Universities by an FPI Predoctoral grant PRE2019-089411 associated to project RTI2018-099850-B-I00. R.H. is supported by MICINN-FEDER PGC2018-093925-B-C32, the AGUAR project number SGR 2017-1040; the Universitat Rovira I Virgili (2014, 2015, and 2016 PFD-URV-B2-147) and CERCA Programme/Generalitat de Catalunya. D.H.-R. is also supported by the Spanish Ministry of Science, Innovation and Universities under the contract PEJ2019-005420-A as part of the i+D+I Garantía Juvenil program. Finally, this work was also supported by the General Foundation of the University of Salamanca under the Plan TCUE 2018-2020; Project "WOLF_FOOTPRINT" (PC-TCUE18-20_2013). Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** Data used for the present study have been included as supplementary files.

**Acknowledgments:** We are greatly appreciative of all the help and support provided by the staff from both the Cabarceno and Hosquillo parks. We would also like to thank the support provided by the TIDOP Group in the Department of Cartographic and Land Engineering of the Higher Polytechnics School of Avila, University of Salamanca, for the use of their tools and facilities. We acknowledge the technical help and support provided by Hasegawa Kaho of the Osaka Technical Institute. Finally, we are extremely grateful for the very constructive comments provided by two anonymous reviewers.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lozano, S.; Mateos, A.; Rodríguez, J. Exploring paleo food-webs in the European Early and Middle Pleistocene: A Network Analysis. *Quat. Int.* **2016**, *413*, 44–54. [CrossRef]
2. Rodríguez-Gómez, G.; Palmqvist, P.; Ros-Montoya, S.; Espigares, M.P.; Martínez-Navarro, B. Resource availability and competition intensity in the carnivore guild of the Early Pleistocene site of Venta Micena (Orce, Baza Basin, SE Spain). *Quat. Sci. Rev.* **2017**, *164*, 154–167. [CrossRef]
3. Linnaeus, C. *Systema Naturae per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species Cum Characteribus, Differentiis. Synonymis, Locis. Tomus, I. Editio Decima, Reformata*; Lautentius Salvius: Stockholm, Sweden, 1758.
4. Perri, A.R.; Mitchell, K.J.; Mouton, A.; Álvarez-Carretero, S.; Hulme-Beaman, A.; Haile, J.; Jamieson, A.; Meachen, J.; Lin, A.T.; Schubert, B.W.; et al. Dire wolves were the last of an ancient New World canid lineage. *Nature* **2021**, *591*, 87–91. [CrossRef]
5. Brugal, J.P.; Boudadi-Maligne, M. Quaternary small to large canids in Europe: Taxonomic status and biochronological contribution. *Quat. Int.* **2011**, *243*, 171–182. [CrossRef]
6. Sardella, R.; Bertè, D.; Iurino, D.A.; Cherin, M.; Tagliacozzo, A. The wolf from Grotta Romanelli (Apulia, Italy) and its implications in the evolutionary history of *Canis lupus* in the Late Pleistocene of Southern Italy. *Quat. Int.* **2014**, *328*, 179–195. [CrossRef]
7. Bartolini-Lucenti, S.; Bukhsianidze, M.; Martínez-Navarro, B.; Lodkipanidze, D. The Wolf from Dmanisi and Augmented Reality: Review, Implications, and Opportunities. *Front. Earth Sci.* **2020**, *8*, 131. [CrossRef]
8. Crégut-Bonnoure, E. Famille des Canidae. In *Les Grands Mammiféres Plio-Pléistocènes d'Europe*; Guérin, C., Patou-Mathis, M., Eds.; Elsevier Masson: Paris, France, 1996; pp. 156–166.
9. Ripoll, M.P.; Morales-Pérez, J.V.; Serra, A.S.; Tortosa, E.A.; Montañana, I.S. Presence of the genus Cuon in the Upper Pleistocene and initial Holocene sites of the Iberian Peninsula: New remains identified in archaeological contexts of the Mediterranean region. *J. Archaeol. Sci.* **2010**, *37*, 437–450. [CrossRef]
10. Mallye, J.B.; Costamagno, S.; Boudadi-Maligne, M.; Prucca, A.; Lautoulandie, V.; Thiébaut, C.; Mourre, V. Dhole (*Cuon alpinus*) as a bone accumulator and new taphonomic agent? The case of Noisetier cave (French Pyrenees). *J. Taphon.* **2012**, *10*, 317–547.
11. Martínez-Navarro, B.; Lucenti, S.B.; Palmqvist, P.; Ros-Montoya, S.; Madurell-Malapeira, J.; Espigares, M.P. A new species of dog from the Early Pleistocene site of Venta Micena (Orce, Baza Basin, Spain). *Comptes Rendus Palevol* **2021**, *20*, 297–314. [CrossRef]
12. Coumont, M.P. Proposition d'un référentiel taphonomique fossile de faunes issues d'avens-pèges. *Ann. De Paléontologie* **2009**, *95*, 1–20. [CrossRef]
13. Castel, J.C.; Coumont, M.P.; Boudadi-Maligne, M.; Prucca, A. Rôle et Origine des Grandes Carnivores dans les Accumulations Naturelles. Le Cas de Loups (*Canis lupus*) de l'Igue du Gral (Sauliac-sur-Célé, Lot, France). *Rev. De Paléobiologie Genève* **2010**, *29*, 411–425.
14. Campmas, E.; Michel, P.; Costamagno, S.; El Hajraoui, M.A.; Nespoulet, R. Which predators are responsable for faunal accumulations at the Late Pleistocene layers of El Harhoura 2 Cave (Témara, Morocco)? *Comptes Rendus Palevol* **2017**, *16*, 333–350. [CrossRef]
15. Mech, L.D. *The Wolf: The Ecology and Behaviour of an Endangered Species*; Natural History Press: New York, NY, USA, 1970.
16. Haynes, G. Prey bones and predators: Potential ecologic information from analysis of bone sites. *Ossa* **1980**, *7*, 75–97.
17. Haynes, G. Bone Modification and Skeletal Disturbances by Natural Agencies: Studies in North America. Ph.D. Thesis, The Catholic University of America, Washington, DC, USA, 1981.
18. Haynes, G. Utilization and skeletal disturbances of North American Prey Carcasses. *Arctic* **1982**, *35*, 266–281. [CrossRef]
19. Yravedra, J.; Lagos, L.; Bárcena, F. The Wild Wolf (*Canis lupus*) as a dispersal agent of animal carcasses in Northwestern Spain. *J. Taphon.* **2012**, *10*, 227–248.
20. Ovodov, N.; Crockford, S.J.; Kuzmin, Y.V.; Higham, T.F.G.; Hodgins, G.W.L.; Plicht, J. A 33,000-Year-Old incipient dog from the Altai Mountains of Siberia: Evidence of the Earliest Domestication Disrupted by the Last Glacial Maximum. *PLoS ONE* **2011**, *6*, e22821. [CrossRef]
21. Germonpré, M.; Fedorov, S.; Danilov, P.; Galeta, P.; Jimenez, E.L.; Sablin, M.; Losey, R.J. Palaeolithic and prehisotric dogs and Pleistocene wolves from Yakutia: Identification of Isolated Skulls. *J. Archaeol. Sci.* **2017**, *78*, 1–19. [CrossRef]
22. Drake, A.G.; Coquerelle, M.; Colombeau, G. 3D morphometric analysis of fossil canid skulls contradicts the suggested domestication of dogs during the Late Palaeolithic. *Sci. Rep.* **2015**, *5*, 8299. [CrossRef]
23. Wilczyński, J.; Haynes, G.; Sobczyk, Ł.; Svoboda, J.; Roblíčková, M.; Wojtal, P. Friend or Foe? Large canid remains from Pavlovian sites and their archaeozoological context. *J. Anthr. Arch.* **2020**, *59*, 101197. [CrossRef]
24. Larson, G.; Karlsson, E.K.; Perri, A.; Webster, M.T.; Ho, S.Y.W.; Peters, J.; Stahl, P.W.; Piper, P.J.; Lingaas, F.; Fredholm, M.; et al. Rethinking dog domestication by integrating genetics, archaeological and biogeography. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 8878–8883. [CrossRef] [PubMed]
25. Cayuela, L. Habitat evaluation for the Iberian wolf *Canis lupus* in Picos de Europa National Park, Spain. *Appl. Geogr.* **2004**, *24*, 199–215. [CrossRef]
26. Blanco, J.C.; Cortes, Y. Ecological and social constraints of wolf recovery in Spain, A New Era for Wolves and People. In *Wolf Recovery, Human Attitudes and Policy*; Musiani, M., Boitani, L., Paquet, P.C., Eds.; University of Calargy Press: Calargy, AB, Canada, 2009; pp. 41–66.
27. Woodroffe, R.; Redpath, S.M. When the hunter becomes the hunted. *Science* **2015**, *348*, 1312–1314. [CrossRef]

28. Pimenta, V.; Barroso, I.; Boitani, L.; Beja, P. Wolf predation on cattle in Portugal: Assessing the effects of husbandry systems. *Biol. Conserv.* **2017**, *207*, 17–26. [CrossRef]
29. Pimenta, V.; Barroso, I.; Boitani, L.; Beja, P. Risks *a la carte*: Modelling the occurrence and intensity of wolf predation on multiple livestock species. *Biol. Conserv.* **2018**, *228*, 331–342. [CrossRef]
30. Ericsson, G.; Heberlein, T.A.; Karlsson, J.; Bjärvall, A.; Lundvall, A. Support for hunting as a means of wolf Canis lupus population control in Sweden. *Wildlife Biol.* **2004**, *10*, 269–276. [CrossRef]
31. Lososová, J.; Kouřilová, J.; Soukupová, N. Controversial approach to wolf management in the Czech Republic. *Agr. Econ.* **2021**, *67*, 1–10. [CrossRef]
32. Yravedra, J.; Maté-González, M.Á.; Courtenay, L.A.; González-Aguilera, D.; Fernández-Fernández, M. The use of canid tooth marks on bone for the identification of livestock predation. *Sci. Rep.* **2019**, *9*, 16301. [CrossRef]
33. Dawkins, W.B. *Cave Hunting, Researches of the Evidence of Caves Respecting the Early Inhabitants of Europe*; Macmillian & Co.: London, UK, 1874.
34. Martin, H. Reserches sur L'evolution du Mousterien dans le Gisement de la Quina Charante. In *Premier Volume—Industrie Osseuse*; Schleicher Freres: Paris, France, 1907–1910.
35. Binford, L.R. *Bones: Ancient Men and Modern Myths*; Academic Press Inc.: New York, NY, USA, 1981.
36. Blumenschine, R.J. Percussion marks, tooth marks and the experimental determinations of the timing of hominid and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania. *J. Hum. Evol.* **1995**, *29*, 21–51. [CrossRef]
37. Selvaggio, M.M.; Wilder, J. Identifying the involvement of multiple carnivore taxa with archaeological bone assemblages. *J. Archaeol. Sci.* **2001**, *28*, 465–470. [CrossRef]
38. Delaney-Rivera, C.; Plummer, T.W.; Hodgson, J.A.; Forrest, F.; Hertel, F.; Oliver, J.S. Pits and pitfalls: Taxonomic variability and patterning in tooth mark dimensions. *J. Archaeol. Sci.* **2009**, *36*, 2597–2608. [CrossRef]
39. Andrés, M.; Gidna, A.O.; Yravedra, J.; Domínguez-Rodrigo, M. A study of dimensional differences of tooth marks (pits and scores) on bones modified by small and large carnivores. *J. Archaeol. Anthropol. Sci.* **2012**, *4*, 209–219. [CrossRef]
40. Yravedra, J.; García-Vargas, E.; Maté-González, M.Á.; Aramendi, J.; Palomeque-González, J.; Vallés-Iriso, J.; Matesanz-Vicente, J.; González-Aguilera, D.; Domínguez-Rodrigo, M. The use of Micro-Photogrammetry and Geometric Morphometrics for identifying carnivore agency in bone assemblage. *J. Archaeol. Sci: Rep.* **2017**, *14*, 106–115. [CrossRef]
41. Aramendi, J.; Maté-González, M.Á.; Yravedra, J.; Cruz-Ortega, M.; Arriaza, M.C.; González-Aguilera, D.; Baquedano, E.; Domínguez-Rodrigo, M. Discerning carnivore agency through the three-dimensional study of tooth pits: Revisiting crocodile feeding at FLK-Zinj and FLK NN3 (Olduvai Gorge, Tanzania). *Palaeogeog. Palaeoclim. Palaeoecol.* **2017**, *488*, 93–102. [CrossRef]
42. Arriaza, M.C.; Yravedra, J.; Domínguez-Rodrigo, M.; Maté-González, M.Á.; García-Vargas, E.; Palomeque-González, J.F.; Aramendi, J.; González-Aguilera, D.; Baquedano, E. On applications of micro-photogrammetry and geometric morphometrics to studies of tooth-mark morphology: The modern Olduvai Carnivore Site (Tanzania). *Palaeogeog. Palaeoclim. Palaeoecol.* **2017**, *488*, 103–112. [CrossRef]
43. Arriaza, M.C.; Aramendi, J.; Maté-González, M.Á.; Yravedra, J.; Stratford, D. Characterising leopard as taphonomic agent through the use of micro-photogrammetric reconstruction of tooth marks and pit to score ratio. *Hist. Biol.* **2019**, *33*, 176–185. [CrossRef]
44. Courtenay, L.A.; Yravedra, J.; Huguet, R.; Aramendi, J.; Maté-González, M.Á.; González-Aguilera, D.; Arriaza, M.C. Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks. *Palaeogeog. Palaeoclim. Palaeoecol.* **2019**, *522*, 28–29. [CrossRef]
45. Courtenay, L.A.; Herranz-Rodrigo, D.; González-Aguilera, D.; Yravedra, J. Developments in Data Science Solutions for Carnivore Tooth Pit Classification. *Sci. Rep.* **2021**, *11*, 10209. [CrossRef]
46. Courtenay, L.A.; Herranz-Rodrigo, D.; Huguet, R.; Maté-González, M.Á.; González-Aguilera, D.; Yravedra, J. Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy. *PLoS ONE* **2020**, *15*, e0240328. [CrossRef]
47. Pante, M.C.; Muttart, M.V.; Keevil, T.L.; Blumenschine, R.J.; Njau, J.K.; Merritt, S.R. A new high-resolution 3-D quantitative method for identifying bone surface modifications with implications for the Early Stone Age archaeological record. *J. Hum. Evol.* **2017**, *102*, 1–11. [CrossRef]
48. Gümrükçu, M.; Pante, M.C. Assessing the Effects of Fluvial Abrasion on Bone Surface Modifications using High-Resolution 3-D Scanning. *J. Archaeol. Sci. Rep.* **2018**, *21*, 208–221. [CrossRef]
49. Jiménez-García, B.; Aznarte, J.; Abellán, N.; Baquedano, E.; Domínguez-Rodrigo, M. Deep learning improves taphonomic resolution: High accuracy in differentiating tooth marks made by lions and jaguars. *J. R Soc. Interface* **2020**, *17*, 2020046. [CrossRef]
50. Abellán, N.; Jiménez-García, B.; Aznarte, J.; Baquedano, E.; Domínguez-Rodrigo, M. Deep learning classification of tooth scores made by different carnivores: Achieving high accuracy when comparing African carnivore taxa and testing the hominin shift in the balance of power. *Archaeol. Anthropol. Sci.* **2021**, *13*, 31. [CrossRef]
51. Gidna, A.; Yravedra, J.; Domínguez-Rodrigo, M. A cautionary note on the use of captive carnivores to model wild predator behavior: A comparison of bone modification patterns on long bones by captive and wild lions. *J. Archaeol. Sci.* **2013**, *40*, 1903–1910. [CrossRef]
52. Gidna, A.; Domínguez-Rodrigo, M.; Pickering, T.R. Patterns of bovid long limb bone modification created by wild and captive leopards and their relevance to the elaboration of referential frameworks for palaeoanthropolgy. *J. Archaeol. Sci. Rep.* **2015**, *2*, 302–309. [CrossRef]

53. Courtenay, L.A.; Yravedra, J.; Maté-González, M.Á.; Vázquez-Rodríguez, J.M.; Fernández-Fernández, M.; González-Aguilera, D. The effects of prey size on carnivore tooth mark morphologies on bone; the case study of *Canis lupus* signatus. *Hist. Biol.* **2020**, 1–13. [CrossRef]
54. Cabrera, A. *Los Lobos de España*; Boletín de la Sociedad Royal Española Historia: Madrid, Spain, 1907.
55. Christainsen, P.; Adolfssen, S. Bite Foreces, canine strength and skull allometry in carnivores (Mammalia, carnivore). *J. Zool.* **2005**, *266*, 133–151. [CrossRef]
56. Toledo-González, V.; Ortega-Ojeda, F.; Fonseca, G.M.; García-Ruiz, C.; Navarro-Cáceres, P.; Pérez-Lloret, P.; Marín-García, M.P. A morphological and morphometric dental analysis as a forensic tool to identify the iberian wolf (*Canis lupus signatus*). *Animals* **2020**, *10*, 975. [CrossRef]
57. Yravedra, J.; Andrés, M.; Domínguez-Rodrigo, M. A taphonomic study of the African Wild Dog (*Lycaon pictus*). *Archaeol. Anthropol. Sci.* **2014**, *6*, 113–124. [CrossRef]
58. Moclán, A.; Domínguez-Rodrigo, M.; Yravedra, J. Classifying agency in bone breakage: An experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using machine learning (ML) algorithms. *Archaeol. Anthropol. Sci.* **2019**, *11*, 4663–4680. [CrossRef]
59. Wabakken, P.; Sand, H.; Kojola, I.; Zimmermann, B.; Arnemo, J.M.; Pedersen, H.C.; Liberg, O. Multistage, long-range natal dispersal by a global positioning system-collared Scandinavian wolf. *J. Wildl. Manag.* **2007**, *71*, 1631–1634. [CrossRef]
60. Kojola, I.; Aspi, J.; Hakala, A.; Heikkinen, S.; Ilmoni, C.; Ronkainen, S. Dispersal in an Expanding Wolf Population in Finland. *J. Mammal.* **2006**, *87*, 281–286. [CrossRef]
61. Kojola, I.; Kaartinen, S.; Hakala, A.; Heikkinen, S.; Voipio, H.M. Dispersal behavior and the connectivity between Wolf populations in Northern Europe. *J. Wildl. Manag.* **2009**, *73*, 309–313. [CrossRef]
62. Ciucci, P.; Reggioni, W.; Maiorano, L.; Boitani, L. Long-distance dispersal of a rescued Wolf from the Northern Apennines to the Western Alps. *J. Wildl. Manag.* **2009**, *73*, 1300–1306. [CrossRef]
63. Ražen, N.; Brugnoli, A.; Castagna, C.; Groff, C.; Kaczensky, P.; Kljun, F.; Kos, I.; Krofel, M.; Luštrik, R.; Majić, A.; et al. Long-distance dispersal connects Dinaric-Balkan and Alpine grey wolf (*Canis lupus*) populations. *Eur. J. Wildl. Res.* **2015**, *62*, 137–142. [CrossRef]
64. Byrne, M.E.; Webster, S.C.; Lance, S.L.; Love, C.N.; Hinton, T.G.; Shamovich, D.; Beasley, J.C. Evidence of long-distance dispersal of a gray wolf from the Chernobyl Exclusion Zone. *Eur. J. Wildl. Research.* **2018**, *64*, 39. [CrossRef]
65. Packila, M.L.; Riley, M.D.; Spence, R.S.; Inman, R.M. Long-distance wolverine dispersal from Wyoming to historic range in Colorad. *Northwest. Sci.* **2017**, *91*, 399–407. [CrossRef]
66. Barry, T.; Gurarie, E.; Cheraghi, F.; Kojola, I.; Fagan, W. Does dispersal make the heart grow bolder? Avoidance of anthropogenic habitat elements across wolf life history. *Anim. Behav.* **2020**, *166*, 219–231. [CrossRef]
67. Maté-González, M.Á.; Aramendi, J.; Yravedra, J.; González-Aguilera, D. Statistical Comparison between Low-Cost Methods for 3D Characterization of Cut-Marks on Bones. *Remote Sens.* **2017**, *9*, 873. [CrossRef]
68. Rohlf, F.K. *tpsDig2 v.2.29*; Ecology & Evolution and Anthropology, Stony Brook University: New York, NY, USA, 2021; Available online: http://life.bio.sunsyb.edu/morph/ (accessed on 30 July 2021).
69. Bello, S.M.; Soligo, C. A new method for the quantitative analysis of cutmark micromorphology. *J. Archaeol. Sci.* **2008**, *35*, 1542–1552. [CrossRef]
70. Maté-González, M.Á.; Yravedra, J.; González-Aguilera, D.; Palomeque-González, J.F.; Domínguez-Rodrigo, M. Micro-photogrammetric characterization of cut marks on bones. *J. Archaeol. Sci.* **2015**, *62*, 128–142. [CrossRef]
71. Razali, N.M.; Wah, Y.B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
72. Höhle, J.; Höhle, M. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 398–406. [CrossRef]
73. Rodríguez-Gonzálvez, P.; Garcia-Gago, J.; Gomez-Lahoz, J.; González-Aguilera, D. Confronting passive and active sensors with non-gaussian statistics. *Sensors* **2014**, *14*, 13759–13777. [CrossRef]
74. Rodríguez-Martín, M.; Rodríguez-Gonzálvez, P.; Ruiz de Oñá Crespo, E.; González-Aguilera, D. Validation of portable mobile mapping system for inspection tasks in termal and fluid-mechanical facilities. *Remote Sens.* **2019**, *11*, 2205. [CrossRef]
75. Cohen, J. *Statistical Power Analysis for Behavioural Sciences*; Routledge: New York, NY, USA, 1988.
76. Schurimann, D.L. A comparison of the two one-sided test procedure and the power approach for assessing the equivalence of average biovariability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680. [CrossRef]
77. Yuen, K.K.; Dixon, W.J. The approximate behaviour and performance of the two-sample trimmed t. *Biometrika* **1973**, *60*, 369–374. [CrossRef]
78. Yuen, K.K. The two-sample trimmed t for unequal population variances. *Biometrika* **1974**, *61*, 165–170. [CrossRef]
79. Dixon, P.M.; Saint-Maurice, P.F.; Kim, Y.; Hibbing, P.; Bai, Y.; Welk, G.J. A primer on the use of equivalence testing for evaluating measurement agreement. *Med. Sci. Sports Exerc.* **2018**, *50*, 837–845. [CrossRef]
80. Mardia, K.V.; Jupp, P.E. *Directional Statistics*; John Wiley: Chichester, UK, 1999.
81. Best, D.; Fisher, N. Efficient simulation of the von mises distribution. *Appl. Stats.* **1979**, *28*, 152–157. [CrossRef]
82. von Mises, R. Über die "ganzzahligkeit" der atomgewichte und verwandte fragen. *Phys. Z.* **1918**, *19*, 490–500.
83. Mardia, K.V. *Statistics of Directional Data*; Academic Press: London, UK, 1972.

84. Watson, G.S.; Williams, E.J. On the construction of significance tests on the circle and on the sphere. *Biometrika* **1956**, *43*, 344–352. [CrossRef]
85. Pewsey, A. Testing circular symmetry. *Can. J. Stats.* **2002**, *30*, 591–600. [CrossRef]
86. Watson, G.S. *Statistics on Spheres*; John Wiley: New York, NY, USA, 1983.
87. Fisher, N.I. *Statistical Analysis of Circular Data*; Cambridge University Press: Camridge, UK, 1993.
88. Wheeler, S.; Watson, G.S. A distribution-free two-sample test on the circle. *Biometrika* **1964**, *51*, 256–257. [CrossRef]
89. Hinton, G.E.; Roweis, S.T. Stochastic neighbor embedding. *In NIPS* **2002**, *15*, 833–840.
90. Pearson, K. Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* **1895**, *58*, 347–352. [CrossRef]
91. Kendall, M.G. *Rank Correlation Methods*; Hafner Publishing Co.: New York, NY, USA, 1955.
92. Martin, O. *Bayesian Analysis with Python*, 2nd ed.; Packt: Birmingham, UK, 2018.
93. Hoffman, M.D.; Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2011**, *15*, 1593–1623.
94. Krushke, J.K. *Doing Bayesian Data Analysis*, 2nd ed.; Academic Press: New York, NY, USA, 2014.
95. Dryden, I.L.; Mardia, K.V. *Statistical Shape Analysis*; John Wiley and Sons: New York, NY, USA, 1998.
96. Adams, D.C.; Rohlf, F.J.; Slice, D.E. A field comes of age: Geometric morphometrics in the 21st century. *Hystrix* **2013**, *24*, 7–14. [CrossRef]
97. Oxnard, C.E. The measurement of form: Beyond biometrics. *Cleft Palate J. Suppl.* **1986**, *23*, 110–128.
98. Goodall, C.R. Procrustes methods in the statistical analysis of shape. *J. R Stat. Soc. B* **1991**, *53*, 285–339. [CrossRef]
99. Jungers, W.L.; Falsetti, A.B.; Wall, C.E. Shape, relative size, and size-adjustments in morphometrics. *Am. J. Phys. Anthropol.* **1995**, *38*, 137–161. [CrossRef]
100. Bookstein, F.L. Principal warps: Thin plate spline and the decomposition of deformations. *Trans. Pattern Anal. Mach. Intel.* **1989**, *11*, 567–585. [CrossRef]
101. Wasserstein, R.L.; Lazar, N.A. The ASA Statement on *p*-Values: Context, process and purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]
102. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a world beyond "$p < 0.05$". *Am. Stat.* **2019**, *73*, 1–19.
103. Benjamin, D.J.; Berger, J.O. Three recommendations for improving the use of *p*-values. *Am. stat.* **2019**, *73*, 186–191. [CrossRef]
104. Colquhoun, D. The False Positive Risk: A proposal concerning what to do about *p*-values. *Am. Stat.* **2019**, *73*, 192–201. [CrossRef]
105. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, Scotland, 1925.
106. Courtenay, L.A.; González-Aguilera, D.; Lagüela, S.; del Pozo, S.; Ruiz Méndez, C.; Barbero-García, I.; Román-Curto, C.; Cañueto, J.; Santos-Durán, C.; Cardeñoso-Álvarez, M.E.; et al. Hyperspectral imaging and robust statistics in non-melanoma skin cancer análisis. *Biomed. Optics. Express.* **2021**, *12*, 5107–5127. [CrossRef]
107. Morgan, K.N.; Tromborg, C.T. Sources of stress in captivity. *Appl. Anim. Behav. Sci.* **2007**, *102*, 262–307. [CrossRef]
108. Mason, G.; Burn, C.C.; Dallaire, J.A.; Kroshko, J.; Kinkaid, H.D.; Jeschke, J.M. Plastic animals in cages: Behavioural flexibility and responses to captivity. *Anim. Behav.* **2013**, *85*, 1113–1126. [CrossRef]
109. Mason, G.; Clubb, R.; Latham, N.; Vickery, S. Why and how should we use environmental enrichment to tackle stereotypic behavior? *Appl. Anim. Behav. Sci.* **2007**, *102*, 163–188. [CrossRef]
110. Rose, P.E.; Nash, S.M.; Riley, L.M. To pace or not to pace? A review of what abnormal repetitive behavior tells us about zoo animal management. *J. Vet. Behav.* **2017**, *20*, 11–21. [CrossRef]
111. Mason, G.J. Species differences in response to captivity: Stress, welfare and the comparative method. *Trends Ecol. Evol.* **2010**, *25*, 713–721. [CrossRef] [PubMed]
112. Wells, D.L. A note on the influence of visitors on the behavior and welfare of zoo-housed gorillas. *Appl. Anim. Behav. Sci.* **2005**, *93*, 13–17. [CrossRef]
113. Wheat, C.H.; Temrin, H. Intrinsic Ball Retrieving in Wolf Puppies Suffests Standing Ancestral Variation for Human-Directed Play Behavior. *Iscience* **2020**, *23*, 100811. [CrossRef]
114. Carlstead, K. Husbandry of the Fennec Fox (*Fennecus zerda*): Environmental conditions influencing stereotypic behavior. *Int. Zoo Yb.* **1991**, *30*, 202–207. [CrossRef]
115. Clubb, R.; Mason, G. Captivity Effects on Wide-Ranging Carnivores. *Nat. Brief. Comm.* **2003**, *425*, 473. [CrossRef] [PubMed]
116. Kawata, K. Zoo Animal Feeding: A Natural History Viewpoint. *Der. Zool. Garten.* **2008**, *78*, 17–42. [CrossRef]
117. Riemer, S. Effectiveness of treatments for firework feats in dogs. *J. Vet. Behav.* **2020**, *37*, 61–70. [CrossRef]
118. Sherman, B.L.; Mills, D.S. Canine anxieties and phobias: An update on separation anxiety and noise adversions. *Vet. Clin. N. Am. Small Anim. Pract.* **2008**, *38*, 1081–1106. [CrossRef] [PubMed]
119. Ballantyne, K.C. Seperation, confinement, or noises: What is scaring that dog? *Vet. Clin. N. Am. Small Anim. Pract.* **2018**, *48*, 367–386. [CrossRef]
120. Storengen, L.M.; Lingaas, F. Noise sensitivity in 17 dog breeds: Prevalence, breed risk and correlation with fear in other situations. *Appl. Anim. Behav. Sci.* **2015**, *171*, 152–160. [CrossRef]
121. Haynes, G. A guide for differentiating mammalian carnivore taxa responsible for gnaw damage to herbivore limb bones. *Paleobiology* **1983**, *9*, 164–172. [CrossRef]
122. Saladié, P.; Huguet, R.; Díez, C.; Rodríguez-Hidalgo, A.; Carbonell, E. Taphonomic Modifications Produced by Modern Brown Bears (*Ursus arctos*). *Int. J. Osteoarchaeol.* **2013**, *23*, 13–33. [CrossRef]

123. Sala, N.; Arsuaga, J.L. Taphonomic studies with wild brown bears (*Ursus arctos*) in the mountains of northern Spain. *J. Archaeol. Sci.* **2013**, *40*, 1389–1396. [CrossRef]

124. Arilla, M.; Rosell, J.; Blasco, R.; Domínguez-Rodrigo, M.; Pickering, T.R. The "Bear" Essentials: Actualistic Research on Ursus arctos arctos in the Spanish Pyrenees and Its Implications for Paleontology and Archaeology. *PLoS ONE* **2014**, *9*, e102457. [CrossRef] [PubMed]

125. Sala, N.; Arsuaga, J.L.; Haynes, G. Taphonomic comparison of bone modifications caused by wild and captive wolves (*Canis lupus*). *Quat. Int.* **2014**, *330*, 126–135. [CrossRef]

126. Ferretti, M.P. Evolution of Bone-Cracking Adaptations in Hyaenids (Mammali, Carnivore). *Swiss J. Geosci.* **2007**, *100*. [CrossRef]

*Publications: Section 2*

# Data Simulation and Augmentation

*Spanish Translation of Title and Abstract*

## Aumento de datos de morfometría geométrica mediante el uso de algoritmos de redes generativas antagónicas

El registro fósil es notorio por ser incompleto y estar distorsionado, lo que frecuentemente condiciona el tipo de conocimiento que se puede extraer a partir de él. En muchos casos, esto suele plantear problemas a la hora de realizar análisis estadísticos complejos empleando la morfometría geométrica, como tareas de clasificación, modelización predictiva, y análisis de varianza. En este artículo, se experimenta con diferentes arquitecturas de redes generativas antagónicas, comprobando cómo influyen variables como el tamaño de la muestra y la dimensionalidad del dominio en la calidad de los resultados. Para evaluar la calidad de los datos aumentados, utilizamos métodos de la estadística robusta. A partir de este estudio, se observa que todos los algoritmos son capaces de producir datos realistas. Además, la red generativa antagónica, empleando diferentes funciones de pérdida, produce datos sintéticos multidimensionales significativamente equivalentes a los datos de entrenamiento originales. No obstante, las redes generativas antagónicas condicionales no tuvieron tanto éxito. Los métodos propuestos pueden reducir el impacto del tamaño de la muestra en futuras aplicaciones de aprendizaje estadístico. Aunque las redes generativas antagónicas no son la solución de todos los problemas relacionados con el tamaño de la muestra, estos pueden superarse si se combinan con otros pasos de preprocesamiento. Por tanto, este acercamiento presenta un medio valioso para aumentar los conjuntos de datos de la morfometría geométrica.

*Supplementary Information and Links*

**Supplementary Information available from:**
https://www.mdpi.com/2076-3417/10/24/9133#supplementary

**Code available from:**
https://github.com/LACourtenay/GMM_Generative_Adversarial_Networks

**Data available from:**
Dataset 1 - https://doi.org/10.6084/m9.figshare.8081105.v1
Dataset 2 (part 1) - https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fjmi.12873&file=jmi12873-sup-0003-SuppMat.txt
Dataset 2 (part 2) - https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fjmi.12873&file=jmi12873-sup-0001-SuppMat.txt
Dataset 3 - https://github.com/LACourtenay/GMM_Measurement_Accuracy_Tools/blob/master/Landmark_Files/All_Landmarks.txt

applied sciences

MDPI

Article

# Geometric Morphometric Data Augmentation Using Generative Computational Learning Algorithms

**Lloyd A. Courtenay * and Diego González-Aguilera**

Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003 Ávila, Spain; daguilera@usal.es

* Correspondence: ladc1995@gmail.com; Tel.: +34-633-647-825

Received: 6 November 2020; Accepted: 19 December 2020; Published: 21 December 2020

**Featured Application: Geometric Morphometrics are a powerful multivariate statistical toolset for the analysis of morphology. While typically used in the study of biological and anatomical variance, modern applications now incorporate these tools into a number of different fields of non-biological origin. Nevertheless, as with many fields of data science, Geometric Morphometric techniques are often impeded by issues concerning sample size. The present study thus evaluates a number of different computational learning algorithms for the augmentation of different datasets. Here we show how generative algorithms from Artificial Intelligence are able to produce highly realistic synthetic data; helping improve the quality of any statistical or predictive modelling applications that may follow.**

**Abstract:** The fossil record is notorious for being incomplete and distorted, frequently conditioning the type of knowledge that can be extracted from it. In many cases, this often leads to issues when performing complex statistical analyses, such as classification tasks, predictive modelling, and variance analyses, such as those used in Geometric Morphometrics. Here different Generative Adversarial Network architectures are experimented with, testing the effects of sample size and domain dimensionality on model performance. For model evaluation, robust statistical methods were used. Each of the algorithms were observed to produce realistic data. Generative Adversarial Networks using different loss functions produced multidimensional synthetic data significantly equivalent to the original training data. Conditional Generative Adversarial Networks were not as successful. The methods proposed are likely to reduce the impact of sample size and bias on a number of statistical learning applications. While Generative Adversarial Networks are not the solution to all sample-size related issues, combined with other pre-processing steps these limitations may be overcome. This presents a valuable means of augmenting geometric morphometric datasets for greater predictive visualization.

**Keywords:** archaeological data science; artificial intelligence; unsupervised learning; generative adversarial networks; robust statistics

## 1. Introduction

### 1.1. Geometric Morphometrics

Geometric Morphometrics (GM) is a powerful multivariate statistical toolset for the analysis of morphology [1]. These methods are of a growing importance in fields such as biology and physical anthropology, with many implications for evolutionary theory and systematics. GM applications employ the use of two or three dimensional homologous points of interest, known as *landmarks*, to quantify geometric variances among individuals [1–4].

GM practices first project landmark configurations onto a common coordinate system. This process is carried out via a series of superimposition procedures, including scaling, rotation and translation, frequently known as Generalized Procrustes Analyses (GPA). GPA is a powerful technique that allows for the direct comparison of landmark configurations, quantifying minute displacements of individual landmarks in space [5,6]. These distortions and deformations can then be used to highlight geometric variations among organisms and can be visualized with ease.

From these superimposed configurations, matrix operations from linear algebra can be performed to project each element under study as a single multidimensional ($\mathbb{R}^n$) point in a newly constructed feature space. This procedure, known as Principal Components Analysis (PCA) is useful for dimensionality reduction and converting landmarks into more manageable data for complex statistical applications [7,8].

A wide array of techniques are known for different pattern recognition and classification tasks in GM. From one perspective, more traditional parametric and non-parametric multivariate statistical analyses can be performed to assess differences and similarities among sample distributions [7]. Likewise, generalized distances and group association probabilities can be used to compare groups of organisms and trends in variation and covariation [9]. Moreover, many popular classification tasks rely on parametric discriminant functions [10,11].

In more recent years, tasks in pattern recognition and classification have received an increase in efficiency and precision with the implementation of Artificially Intelligent Algorithms (AIAs), reporting >90% accuracy in GM applications. In a broad sense, AIAs are algorithms designed to "learn" from data so as to perform a wide array of different tasks. In this context, AIAs can be programmed to automatically learn from subsets of data to adjust their internal parameters, while using other subsets to validate these parameters for performing a certain task [12]. Among the multitude of available algorithms, the most popular AIAs for classification purposes in GM currently include Support Vector Machines (SVM) [13–16], and Artificial Neural Networks (ANN) [17–21]. Both algorithms present distinct advantages, especially in the processing of complex high-dimensional data. As opposed to traditional Linear and/or Partial Least-Squares Discriminant Analyses (LDA and PLSDA), SVMs and ANNs are less susceptible to underlying assumptions within model properties. SVMs, for example, are able to use numerous different kernel functions to overcome issues imposed by linearity [22,23]. ANNs, on the other hand, are highly versatile non-linear algorithms inspired by information processing in the brain, achieving above human performance in a multitude of real-life situations [23–25].

Nevertheless, each of these types of analyses are susceptible to a number of different problems, all of which can affect the reliability of the extracted data. From one perspective, numerous studies have focused on the error produced through data collection procedures, whether this type of error be induced by analyst experience, collection protocols or the definition of the landmark itself [26–29].

Landmarks, for example, can be divided into several different types. While some of the original definitions of landmark types were based on strictly biological features [30], these definitions can be considered somewhat restrictive for morphological analyses outside of anatomy. Under this premise, we prefer to define landmarks in a more general sense (Figure 1), referring to Type I landmarks as anatomical points of biological significance [3,30]; Type II landmarks can be defined as points of mathematical significance (e.g., maximal curvature or length) [3]; and Type III landmarks can be considered constructed points located around outlines or in relation to other landmarks [3]. From another perspective, valuable contributions in the field of GMs have seen the incorporation of computational landmarks into analyses. From this perspective "semi-landmarks" can be computed that "slide" over curves and surfaces in an attempt to reduce bending energy [4]. Finally other promising efforts have been made to develop automated tools for landmark digitization [31].

With both a more generalized definition of landmark types, as well as the inclusion of computational tools for their digitization, GMs have been able to quantify morphological traits across a wide array of different objects (Figure 1), including stone implements and tools [32], as well as microscopic anomalies found on bone [14–16,21,29].

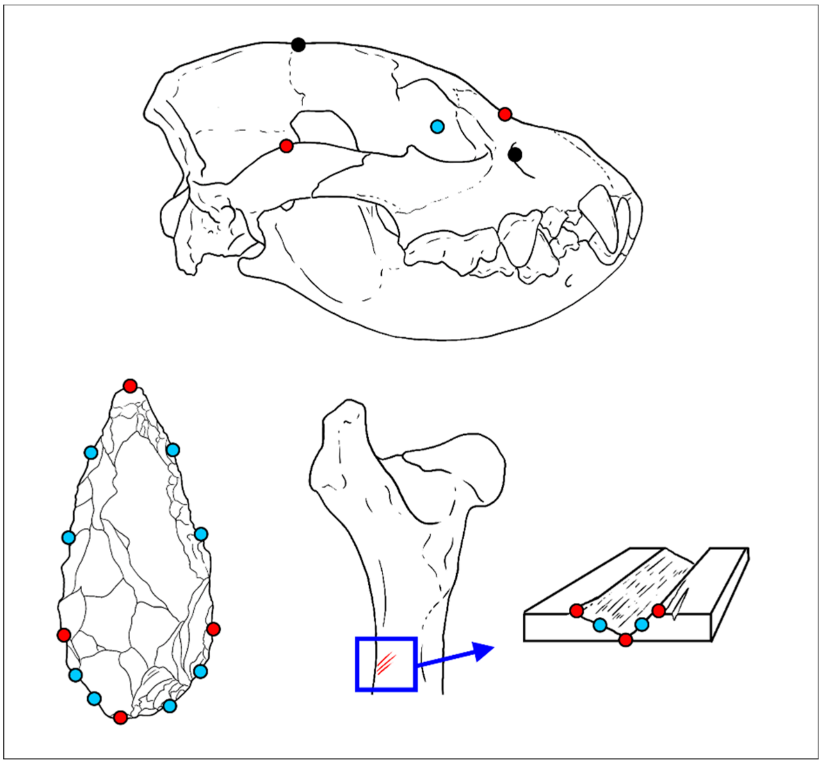


**Figure 1.** Examples of landmark types on different artefacts in the fossil register. Type I landmarks (black) refer to points of biological and anatomical interest, such as the meeting of two sutures or a foramen as represented on the skull above. Type II landmarks (red) are mathematically defined points of interest, such as those points marking the maximal curvature of one, the length of an item, or the deepest point in a microscopic groove. Type III landmarks (blue) are constructed points of interest located in approximation or relation with other elements, such as the centroid of an eye-socket, the general outline of an object, or points in between other landmark types.

More often than not, however, the preservation rate of fossils results in the loss of landmarks, impeding many types of analyses [33,34]. The completeness of the fossil record is thus a major conditioning factor in archaeological and paleontological GM analyses. Considering the number of available fossils for certain species, construction of reliable datasets is difficult, resulting in sample bias. Statistical tests such as Canonical Variant Analyses (CVA), for example, are highly sensitive to small or imbalanced datasets [9]. Moreover, the impact of bias is directly proportional to the number of variables included in multivariate analyses [35]. Even if samples are balanced, in fields such as paleoanthropology obtaining large sample sizes is often difficult, and thus the predictive capacity of discriminant models may fall significantly.

### *1.2. Data Augmentation*

Resampling techniques in traditional statistics have had great success in providing more robust methods to test statistical approximations and *p*-value calculations. Tests requiring permutations as well as more computationally efficient Monte Carlo simulations have been a standard procedure in statistical practices for over half a century [36,37]. Their versatility to both parametric and non-parametric assumptions makes handling imbalanced and skewed data much more reliable, while proving less sensitive to samples of smaller sizes [38]. Nevertheless, a critical issue when considering small sample sizes are an "insufficiency of information density" that is able to correctly provide a general overview of the population's distribution [39]. This issue becomes apparent when trying to classify new individuals. With insufficient knowledge of the true coverage of a domain, the interpretation of new information is much more difficult. In data science this phenomenon is usually known as *overfitting* for classification algorithms [25].

One statistical technique frequently used to overcome this issue is resampling with replacement, known as *bootstrapping*. Bootstrapping duplicates the data multiple times creating a virtual population from a distribution sample [40,41]. As opposed to resampling techniques without replacement (e.g., permutation, cross-validation, jackknife), bootstrap procedures are efficient in inferential tasks helping to simulate the general nature of the population. Nevertheless, neither of these resampling procedures, in truth, simulate new information. While they may be useful for inflating the dataset and providing enough information for a model to adjust its weights, overfitting is likely, as the space between data points can still be considered "uncharted territory".

In response, data scientists and specialists in AIAs propose the use of synthetically produced new data to overcome these problems [42]. While using synthetic "fake" data has drawn some skepticism from scientists, numerous experiments in predictive modelling have empirically shown how these synthetic datasets not only reduce overfitting, but actually produce an increase in accuracy [43]. This is achieved through creating new data that is "meaningful" to the real distribution by adapting the data that is already available [44] (Tanaka and Aranha, 2019). These advances have had a major impact on scientific disciplines dedicated to computational learning, especially in the case of highly complex applications for computer vision [25]. One of the key AIAs responsible for this success is the Generative Adversarial Network (GAN).

GANs were originally presented as an unsupervised AIA capable of creating new data, based on the training data provided [45]. In less than a decade, GANs have been efficiently incorporated into a wide variety of applications, especially in fields of computer vision and image processing. A GAN consists of two neural networks trained simultaneously. The first model, known as the *Generator*, is trained to produce synthetic information which the second model, the *Discriminator*, evaluates for authenticity. The two models are trained in competition (i.e., adversarial), with the generator working to produce data that the discriminator is unable to classify as synthetic. The final product is a generator model capable of producing completely new data that is indistinguishable from the real training set. With the additional advantage of a neural network's non-linear internal configuration, GANs are highly efficient in mapping out any type of probability distribution. From this perspective, GANs have been used for a wide arrange of different applications, including the generation of photo-realistic images, anomaly detection, music generation, and the approximation of a number of different statistical distributions [46,47].

## 2. Materials and Methods

This study presents an experimental protocol used to evaluate and assess different types of GANs for augmenting GM datasets. Through experimenting with different architectures, configurations and training strategies, this study aims to propose an optimal architecture for augmenting data of this type. In order to evaluate these results, both descriptive statistics and equivalency testing have been used. Figure 2 presents a general visualization of the described workflow.

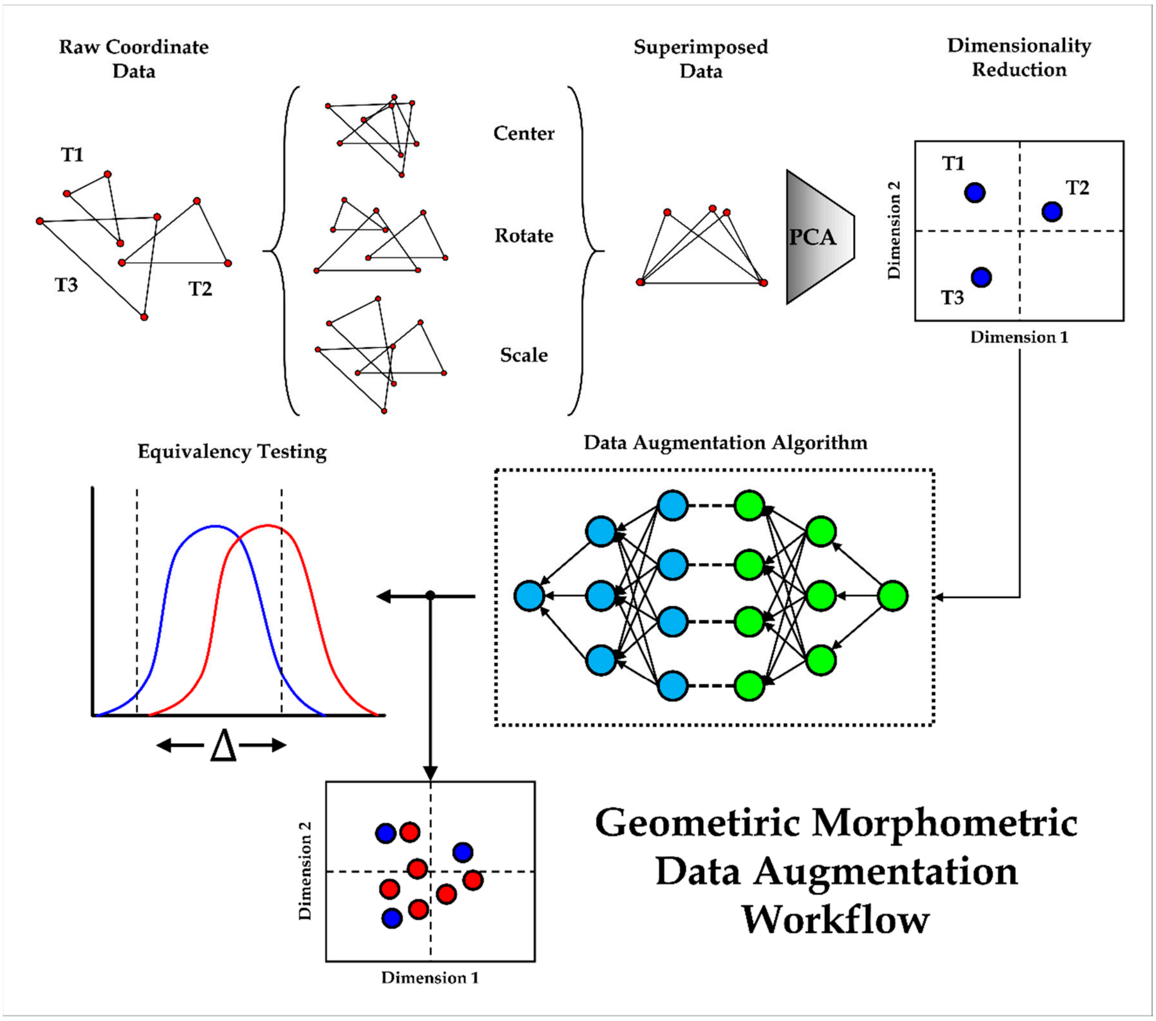


**Figure 2.** Workflow proposed for data augmentation tasks in geometric morphometrics.

### *2.1. Datasets*

Experiments included within this study were performed on a total of three GM datasets. These datasets originated from experimental archaeology samples in taphonomy. Each of these samples thus represent the morphological features of microscopic alterations observed on bone, using GM to quantify these morphologies for diagnostic purposes.

Nevertheless, considering the objective of this study is to observe the effects of generative learning for GM data augmentation, the origin of these datasets was considered unimportant. The reason behind this lies in how, regardless of the element under study, data used for GM analysis consists of superimposed landmark coordinates (Figures 1 and 2). From these coordinates, dimensionality reduction can be used to convert each element into a single vector from which models can learn from (Figure 2). Therefore, irrespective of whether the raw landmark data was obtained from paleoanthropological specimens, lithic tools or carnivore feeding samples (Figure 1), all landmarks are similarly embedded into a single vector that can be used as the input to our computational models. Additional use of these three case studies was based on how each dataset was personally generated by the corresponding author, providing a means of controlling the origin of information.

The 3 datasets used consists of a mixture of manually placed landmarks (Type II or III; Figure 1), as well as some computational semi-landmarks [3,4]. The three datasets include;

- Dataset 1 (DS1); canid tooth score dataset [16]. This dataset consists of 105 individuals from three different experimental carnivore feeding samples (labelled foxes, dogs and wolves). 3D models for data extraction were generated using a low-cost structured light surface scanner (David SLS-2). The topography of each 3D digital model was then used to extract 2D images where landmarks could be placed. Landmark data consist of a mixture of Type II and Type III 2D landmarks.

- Dataset 2 (DS2); scratch and graze trampling dataset [15]. This dataset consists of 60 individuals from two different experimental trampling mark samples (labelled scratches and grazes). Each of the elements under study were digitized employing a 3D Digital Microscope (HIROX KH-8700), using between 100× and 200× magnification. Collection of landmark data was then performed following a series of measurements that established a 3D coordinate system across the model. Landmark data consist of a mixture of Type II and Type III 3D landmarks.
- Dataset 3 (DS3); semi-landmark based tooth pit dataset [29]. This dataset consists of an adaptation of DS1 using 60 individuals from two carnivore feeding samples (labelled dogs and wolves). 3D models for data extraction were generated using a low-cost structured light surface scanner (David SLS-2). Landmark data consist of a mixture of 3D Type II landmarks and a mesh of semi-landmarks.

These three datasets were chosen considering the dimensionality of the corresponding feature-space produced for GM analysis ($\mathbb{R}^{14}$, $\mathbb{R}^{39}$ and $\mathbb{R}^{60}$ respectively). With each of these datasets presenting different dimensionalities, optimal GAN architectures could therefore be proposed so as to establish a standardized protocol, regardless of the target domain's $\mathbb{R}^{n}$ size.

These datasets were also chosen to observe the effect original sample size has on the accuracy of synthetic data. The latter was tested via minimum sample size calculations according to Cohen's *d* (power = 0.8, d = 0.8, $\alpha = 0.05$, ratio = 1:1) [35]. This established a minimum sample size for two-sample statistical comparisons of 26 individuals, rounded up to 30 for simplicity. In accordance with this calculation, experiments were performed by randomly sampling 30 real individuals and comparing them with 30 synthetic individuals. In datasets where larger samples were available, 60 real individuals were sampled and compared with 60 synthetic data points.

### 2.2. Baseline Geometric Morphometric Data Acquisition

Each of the datasets were prepared using traditional GM techniques, first performing a full Procrustes fit of landmark coordinates via GPA (Figure 2), followed by the extraction of multivariate features through PCA [7,8]. Considering how the objectives of this study are to find the optimal algorithm for mapping out multidimensional distributions, differences in *shape-size* relationships were considered irrelevant for this study. GPA was therefore only performed using fully superimposed coordinates in shape feature space.

From here, PC scores were analyzed evaluating their dimensionality and the proportion of variance represented across each of the decomposed eigenvectors and their eigenvalues. Considering how the final eigenvalues begin to represent little or no variance within the landmark configuration, preference was given to those PC scores representing up to 95% of sample variance for statistical evaluations.

For the purpose of this study, GM pre-processing of samples was performed in the free statistical software R (https://www.r-project.org/, v.3.5.1 64-bit).

### 2.3. Generative Adversarial Networks

A GAN is a Deep Learning (DL) architecture used for the synthesis of data via a generator model. GANs are fit to data using an unsupervised approach, where the generator is trained by competing with a discriminator that evaluates the authenticity of the synthetic data produced [24,41]. While the basic concept behind a GAN is relatively straightforward, the theory behind their configuration and training can be incredibly challenging [25,48–51].

To generate new data, the generator samples from a random Gaussian distribution (e.g., $\mu = 0$, $\sigma = 1$), finding the best means of mapping this data out onto the real sample domain. A fixed-length random vector is used as input, triggering the generative process. Once trained, this vector space can essentially be considered a compressed representation of the real data's distribution. This multidimensional vector space is most commonly referred to in DL literature as *latent space* [21,48].

The discriminator model takes as input the output of the generator. This discriminator can then be used to predict a class label (real or fake) for the generated data. In some cases, this model is referred to as a *critic model* [52,53].

For the purpose of this study, multiple experiments were performed to define an optimal GAN architecture. These experiments followed standard DL protocol, finding the optimal neural network configurations by evaluating the effects of each hyperparameter on model performance. Summaries of the hyperparameters tested are included in Table 1.

**Table 1.** List of hyperparameters and settings tested during optimization of GAN model architectures.

| Hyperparameter | Tested Settings |
|---|---|
| Number of Layers | – |
| Node Density | – |
| Activation Functions | ReLU, Leaky ReLU, Tanh, Swish, ELU, Sigmoid, Linear |
| Kernel Initializer | None, Uniform, Normal and their Random, Truncated or Glorot variants. |
| Dropout | None, Present with thresholds between 0.01 and 0.9 |
| Weight Regularizer | None, l2 with thresholds between 0.01 and 0.0001 |
| Weight Constraint | UnitNorm, MaxNorm, MinMaxNorm |
| Batch Normalization | Present, Absent |
| Training Epochs | Between 100 and 2000 |
| Batch Size | 4, 8, 16, 32 |
| Optimizers | Adam, RMSprop, Stochastic Gradient Descent. Adagrad |
| Learning Rate | Between 0.1 and 0.00001 |
| Decay | Between 0.9 and 0.0001 |
| Momentum | Between 0.99 and 0.1 |
| Loss | Binary cross-entropy, Mean Squared Error, Least Squares, Wasserstein Loss, Wasserstein Gradient Penalty Loss. |

In addition to this, the extensive literature on the "best-practices" in GAN research and different heuristics in GAN hyperparameter selection were considered [48,49,51,54]. Among these, common "GAN-Hacks" were evaluated, including:

- Use of the Adam optimization algorithm (α = 0.0002, β1 = 0.5)
- Use of dropout in the generator with a probability threshold of 0.3
- Use of Leaky ReLU (slope = 0.2)
- Stack hidden layers with increasing size in the generator and decreasing size in the discriminator.

For training, trials experimenting with the number of epochs and batch sizes were performed. The final values were chosen in accordance with the requirements of the model in order to reach an acceptable stability.

While binary cross-entropy is typically a recommended loss function for training, this study experimented with alternatives, such as the Least Squares loss function (*LSGAN*) and two versions of Wasserstein loss (*WGAN*). *LSGAN* was originally proposed as a means of overcoming small or vanishing gradients, which are frequently observed when using binary cross-entropy [50,52]. In *LSGAN*, the discriminator ($D$) attempts to minimize the loss ($L$), using the sum squared difference between the predicted and expected values for real and fake data (Equation (1)), while the generator ($G$) attempts to minimize this difference assuming data is real (Equation (2)):

$$L_D^{LSGAN} = -\mathrm{E}_{x\sim p_d}\left[(D(x)-1)^2\right] + \mathrm{E}_{\hat{x}\sim p_g}\left[D(\hat{x})^2\right] \tag{1}$$

$$L_G^{SGAN} = -\mathrm{E}_{\hat{x}\sim p_g}\left[(D(\hat{x}-1))^2\right] \tag{2}$$

This results in a greater penalization of larger errors (E) which forces the model to update weights more frequently, therefore avoiding vanishing gradients [55]. *WGAN*, on the other hand, is based on the theory of *Earth-Mover's distance* [52], calculating the distance between the two probability distributions so that one distribution can be converted into another (Equations (3) and (4)):

$$L_D^{WGAN} = -\mathrm{E}_{x \sim p_d}[D(x)] + \mathrm{E}_{\hat{x} \sim p_g}[D(\hat{x})] \tag{3}$$

$$L_G^{SGAN} = -\mathrm{E}_{\hat{x} \sim p_g}[D(\hat{x})] \tag{4}$$

*WGAN* additionally uses weight constraints (hypercube of [−0.01, 0.01]) to ensure that the discriminator lies within a 1-Lipschitz function. In certain cases, however, this has been reported to produce some undesired effects [53]. As an alternative, a proposed adaptation, in the form of *gradient penalty WGAN* (*WGAN-GP*), includes the same loss for the generator (Equation (4)) but a modified discriminator (eg., 5) with no weight constraints [53,56]:

$$L_D^{WGANGP} = L_D^{WGAN} + \lambda \mathrm{E}_{\hat{x} \sim p_g}\left[(\|\nabla D(\alpha x + (1 - \alpha \hat{x}))\|_2 - 1)^2\right] \tag{5}$$

For both loss functions to work, the output of *D* requires a linear activation function. Finally, optimization tests were performed using Adam ($\alpha$ = 0.0002, $\beta_1$ = 0.5) and RMSprop ($\alpha$ = 0.00005) [50,53,57,58].

More details on the mathematical components of each GAN loss function can be consulted in the corresponding references [25,50,52,53,55].

GANs were trained on scaled PCA feature spaces with 64-bit values ranging between 1 and −1. This scaling procedure was performed to boost neural network performance and optimization by helping reduce the size of weight updates [24]. For these experiments, GANs were trained on all data within the dataset, regardless of label. This approach was chosen to directly observe how GANs handle this type of input data before considering more complex applications, including sample labels. (See Section 2.4)

All experiments were performed in the Python programming language (https://www.python.org/, v.3.7 64-bit) using TensorFlow (https://www.tensorflow.org/, v.2.0). Neural networks were compiled and trained on the CPU of an ASUS X550VX laptop (Intel® Core$^{\mathrm{TM}}$ i5 6300HQ).

### 2.4. Conditional Generative Adversarial Networks

The final GAN trials performed adapted the optimally defined model in Section 3.1 for Conditional GAN tasks (CGAN). A CGAN is an extension of traditional GANs that incorporate class labels into the input, thus conditioning the generation process. Class labels are encoded and used as input alongside both the latent vector and the original vector in order for the GAN to learn targeted distributions within the dataset [59]. This can be done by using an embedding layer and concatenating the embedded information with the original input [60]. It is recommended that the embedding layer is kept as small as possible [51], with some of the original implementations of CGANs using an embedding layer with a size of only ≈5% of the original flattened generator's output. Because 5% of our largest dataset would still have been <1, experiments were performed with different sized embedding layers to find the optimal configuration. The best results came out using a $\lceil 1/4 \bullet n \rceil$ sized embedding layer, where n corresponds to the number of dimensions in $\mathbb{R}^n$ for each of the targeted feature spaces.

For comparison of GAN and CGAN performance, these models were used separately to augment the DS3. This dataset was chosen considering it was the most complex feature space to map, the most balanced (when compared with DS1), and the most difficult to study (seeing how DS2 presents the highest natural separability).

### *2.5. Synthetic Data Evaluation*

Evaluation of GANs is a complex issue with little general agreement on suitable evaluation metrics [49]. Considering how most practitioners in GAN research work with computer vision applications, many papers use manual inspection of images to evaluate synthesized data [61]. Image evaluation, for example, often consists in the visualization of GAN outputs to check whether they are realistic or not, or the use of specified algorithms that are very image-specific [61]. For the synthesis of numeric data, manual inspection is evidently a very subjective means of evaluating information, while calculations such as Inception-score are not applicable to this type of data [49,61,62]. Under this premise, the majority of metrics used in GAN literature is of little value to the present study, as they almost exclusively focus on the evaluation of images [60,62].

Multidimensional numbers are incredibly difficult to visualize, meaning that precise human inspection of this data is impossible. To overcome this, a number of statistical metrics were adopted for GAN evaluation.

Firstly homogeneity of GM data was tested. In most traditional cases, the elimination of size and preservation of allometry in GPA is known to normalize data [63]. Nevertheless, this assumption does not always hold true. The first logical step was to therefore evaluate distribution homogeneity and normality via multiple Shapiro tests. Synthetic distributions were then compared with the real data to assess the magnitude of differences and the significance of overlapping. For this, a "Two One-Sided" equivalency Test (TOST) was performed. TOST evaluates the magnitude of similarities between samples by using upper ($\varepsilon$S) and lower ($\varepsilon$I) equivalence bounds that can be established via Cohen's $d$. This assesses $H_0$ and $H_a$ using an $\alpha$ threshold of $p < 0.05$, with $H_a$ implicating significant similarities among samples [35,64–67]. For TOST the test statistics used to assess these similarities were dependent on distribution normality. These varied between the traditional parametric method using Welch's $t$-statistic [68], or a trimmed non-parametric approach using Yuen's robust $t$-statistic [69,70]. To differentiate between the two, from this point onwards non-parametric robust TOST will be referred to as rTOST.

More traditional univariate descriptive statistics were also employed. For distributions matching Gaussian properties, sample means and standard deviations were calculated. These were accompanied by calculations of sample skewness and kurtosis. For significantly non-Gaussian distributions, robust statistical metrics were used instead. In these cases, measurements of central tendency were established using the sample median ($m$), while deviations were calculated using the square root of the Biweight Midvariance ($BWMV$) (Equations (6)–(9)) [29,71,72].

$$MAD = m(|x_i - m_x|) \tag{6}$$

$$U = \frac{x_i - m}{9MAD} \tag{7}$$

$$a_i = \begin{cases} 1, if|U_i| < 1 \\ 0, if|U_i| \geq 1 \end{cases} \tag{8}$$

$$BWMV = \frac{n\sum_{i=1}^{n} a_i(x_i - m)^2\left(1 - U_i^2\right)^4}{\left(\sum_{i=1}^{n} a_i\left(1 - U_i^2\right)\left(1 - 5U_i^2\right)\right)^2} \tag{9}$$

Robust skewness and kurtosis values were calculated using trimmed distributions. Trims were established using Interquatile Ranges (IR) [71], with confidence intervals of $p$ = [0.05:0.95]. Both the IR range and the trimmed skewness and kurtosis values were reported.

Finally, wherever possible, correlations were calculated to compare the effect of hyperparameters on the quality of synthesized data. For homogeneous data, the parametric Pearson test was used [73], whereas inhomogeneous data was tested using the non-parametric Kendall $\tau$ rank-based test [74].

Considering neural networks are stochastic in nature, these correlations were performed using data obtained from multiple training runs of each GAN to ensure a more robust calculation.

## 3. Results

All three datasets analyzed present highly inhomogeneous multivariate distributions ($p < 2.2 \times 10^{-16}$). Univariate comparisons (Table 2), however, present a mixture of both inhomogeneous and homogeneous distributions across PC1 and PC2, where the majority of variance is represented.

**Table 2.** Summary of each dataset's target domain with univariate calculations of distribution normality in the top two PC scores.

| | Domain Dimensionality | PCs with 95% Cumulative Variance | PC1 | | PC2 | |
|---|---|---|---|---|---|---|
| | | | **Variance (%)** | **Shapiro Test w (*p*)** | **Variance (%)** | **Shapiro Test w (*p*)** |
| DS1 | ℝ14 | 4 | 69.92 | 0.95 (0.02) | 14.37 | 0.97 (0.15) |
| DS2 | ℝ39 | 11 | 32.27 | 0.96 (0.05) | 25.70 | 0.98 (0.31) |
| DS3 | ℝ60 | 13 | 32.83 | 0.99 (0.75) | 19.55 | 0.98 (0.30) |

GAN failure through mode collapse was frequently observed throughout most of the initial trials, characterized by an intense clustering of points with little to no variation in feature space (Figure 3). Qualitatively, this type of failure is easily diagnosed by visual inspection of graphs. Quantitatively, mode failure can be characterized by a dramatic decrease in variance seen through deviation metrics. To provide an example, Figure 1 presents the use of a vanilla GAN trained on DS2. At first, training can be seen to start well with the closest (yet not optimal) approximation to the target domain's median (Figure 3A). Nevertheless, little variation is present (*BWMV* of PC1 = 0.14, target *BWMV* = 0.38). As training continues, the algorithm is unable to find the correct median, and performance is seen to deteriorate (Figure 3B). This presents an exponential decrease in the variance of synthetic data (Figure 3B PC1 *BWMV* = 0.02, Figure 3C PC1 *BWMV* = 0.0002). Likewise, through mode collapse, the generator is unable to map the true normality of the distribution, generating increasingly normal data in PC1 (Shapiro $w > 0.98$, $p > 0.56$).

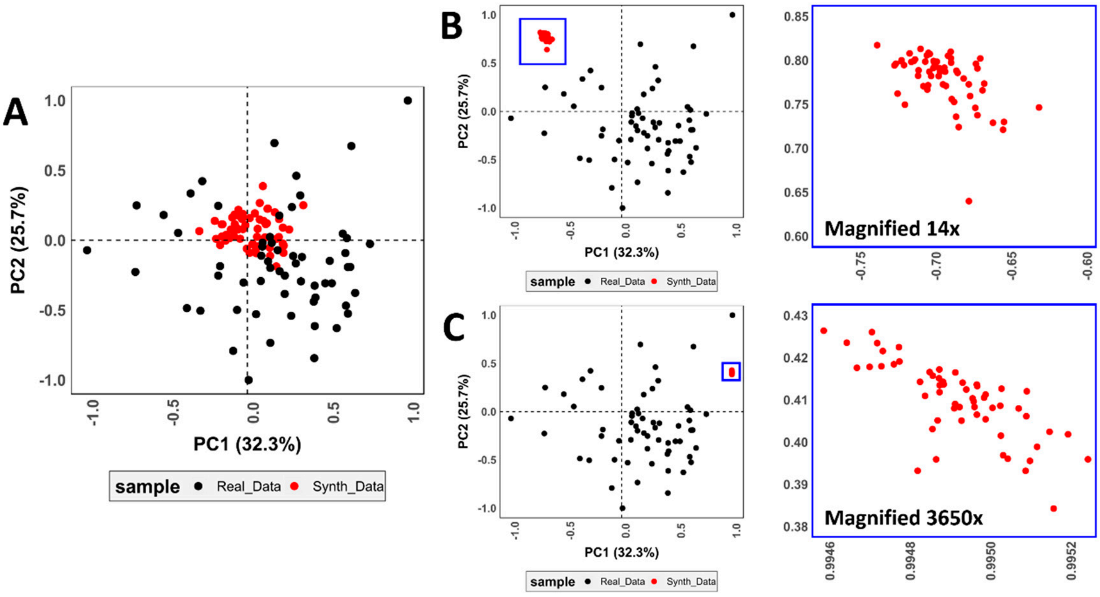


**Figure 3.** Examples of GAN failure in the form of (**A**) slight, (**B**) large and (**C**) extreme mode collapse when used to augment DS2.

Replacing Leaky ReLU with tanh activation functions resulted in a significant improvement of generated sample medians (difference in median for Leaky ReLU = 0.78; tanh = 0.26), yet with little improvement in *BWMV*.

To overcome mode collapse, kernel initializers and batch normalization algorithms were incorporated into both the generator and the discriminator. Batch normalization was included before activation, presenting an increase in *BWMV*. Initializers required careful adjustment, with small standard deviation values resulting in mode collapse. Additional experiments found the discriminator to require a more intense initializer ($\sigma = 0.1$) than the generator ($\sigma = 0.7$), while optimal results were obtained using a random normal distribution. Such a configuration allows the generator more room to adjust its weights, finding the best way of reaching the target domain's median and absolute deviation while preventing the discriminator from learning too quickly.

Experiments adjusting hidden layer densities found symmetry between the generator and the discriminator to be unnecessary. The generator was seen to require more hidden layers in order to learn the distributions efficiently, while a larger density than the output in the last hidden layer also produced an increase in performance. The discriminator worked best with just two hidden layers.

### *3.1. Optimal Architectures*

To optimally adjust these finds with all three datasets, the best GAN architecture that presented no mode collapse was obtained using 3 hidden layers in the generator and two hidden layers in the discriminator. The size of hidden layers are conditioned by the size of the target feature space (Figure 4). To use the example of the largest target domain (DS3 = $\mathbb{R}^{60}$), the generator is programmed so that the first hidden layer ($Gh^1$) is a quarter of the size of the target vector (in this case $Gh^1 = 60 \cdot 1/4 = 15$). If this calculation produces a decimal value (e.g., 59/4 = 14.75), the ceiling of this number is taken ($\lceil 1/4 \cdot 59 \rceil = 15$). This is followed by a layer half the size of the target vector ($Gh^2 = \lceil 1/2 \cdot 60 \rceil = 30$). The generator's final hidden layer is the size of the vector plus one quarter ($Gh^3 = \lceil 1/4 \cdot 60 \rceil + 60 = 75$). The discriminator, on the other hand, is composed of two hidden layers, with the first hidden layer being the same size as $Gh^2$, and the second hidden layer equivalent to $Gh^1$. Each hidden layer is followed by a batch normalization algorithm before being activated using the tanh function. Tanh works best considering the target domain is scaled to values between -1 and 1. Other components of the algorithm include a dropout layer ($p = 0.4$) prior to the discriminator's output and random normal kernel initializers in both models (discriminator $\sigma = 0.1$, generator $\sigma = 0.7$).

Experiments with loss functions and optimization algorithms showed a significant improvement in performance using *LSGAN* and *WGAN* variants when compared with vanilla GAN's binary cross-entropy (rTOST $p < 0.05$, Table 3 with more details explained in Section 3.2). All three GANs were able to generate realistic distributions, with *WGAN-GP* outperforming *WGAN* in some cases (Table 3). *LSGAN* additionally worked best when using Adam optimization ($\alpha = 0.0002$, $\beta_1 = 0.5$), while *WGAN* and *WGAN-GP* excelled using RMSprop ($\alpha = 0.00005$).

**Table 3.** Best obtained absolute difference and p value calculations for robust equivalency testing of each of the synthesized distributions using different GANs.

| | *LSGAN* | | *WGAN* | | *WGAN-GP* | |
|---|---|---|---|---|---|---|
| | $\lvert d\rvert$ | $p$ | $\lvert d\rvert$ | $p$ | $\lvert d\rvert$ | $p$ |
| DS1 | 0.187 | 0.0105 | 0.075 | $2.9 \times 10^{-6}$ | 0.169 | 0.0013 |
| DS2 | 0.019 | $2.2 \times 10^{-16}$ | 0.043 | $1.3 \times 10^{-11}$ | 0.002 | $2.1 \times 10^{-20}$ |
| DS3 | 0.052 | $7.3 \times 10^{-20}$ | 0.040 | $6.7 \times 10^{-23}$ | 0.031 | $3.6 \times 10^{-22}$ |

The optimal batch number was found at 16. This allowed the discriminator enough data to objectively evaluate performance and, thus, resulted in more efficient weight updates for the generator. The number of epochs, however, were highly dependent on the number of individuals used for training.

Finally, the dimensionality of latent space was also found to be conditioned by the size of the target domain, as will be explained in continuation.

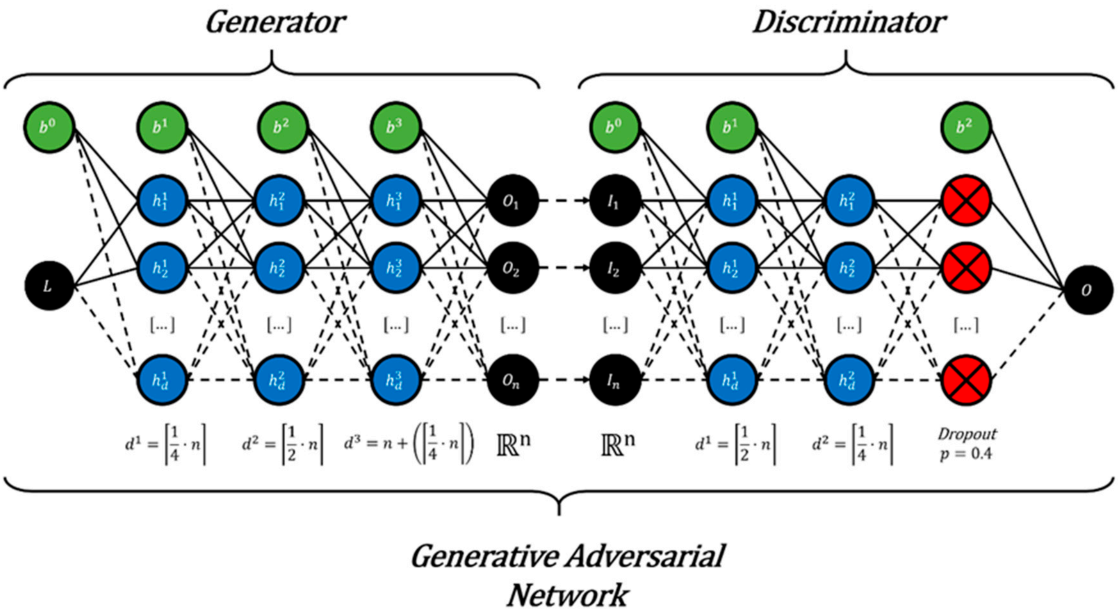

Generator
Discriminator
$b^0$
$b^1$
$b^2$
$b^3$
$L$
$h_1^1$
$h_2^1$
$h_d^1$
$h_1^2$
$h_2^2$
$h_d^2$
$h_1^3$
$h_2^3$
$h_d^3$
$O_1$
$O_2$
$O_n$
$I_1$
$I_2$
$I_n$
$O$
[...]
$d^1 = \left[\frac{1}{4} \cdot n\right]$
$d^2 = \left[\frac{1}{2} \cdot n\right]$
$d^3 = n + \left(\left[\frac{1}{4} \cdot n\right]\right)$
$\mathbb{R}^n$
$\mathbb{R}^n$
$d^1 = \left[\frac{1}{2} \cdot n\right]$
$d^2 = \left[\frac{1}{4} \cdot n\right]$
Dropout
$p = 0.4$
Generative Adversarial Network

**Figure 4.** Descriptive figure presenting the optimal GAN architecture for geometric morphometric data augmentation. Input ($I$) and output ($O$) neurons are represented in black with bias ($b$) in green. The output of the generator and the input of the discriminator is represented by the n number of dimensions in the $\mathbb{R}^n$ dimensional target feature space. The latent vector ($L$) input for the generator must be adjusted according to the dimensionality of the target feature space. Hidden neurons ($h$) in layers $h^n$ have a density ($d$) that is also conditioned by the shape of the target distribution. Finally, the discriminator has an additional dropout layer (red) with a threshold of $p = 0.4$.

### 3.2. Experiments with Dimensionality and Sample Size

Initial trials with latent space found $\mathbb{R}^{50}$ to produce the best results on average, especially in the case of DS2 (results of which have already been reported in Table 2). Nevertheless, interesting patterns emerged when experimenting with larger and smaller latent vector inputs. Starting with the case of the smallest target domain (DS1, Table 4), a significant negative correlation was detected when observing rTOST $p$ values compared with the size of the latent vector over numerous runs (Kendall's $\tau = -0.44$, $p = 0.001$). This is also true when considering rTOST absolute difference values ($\tau = -0.41$, $p = 0.003$). This correlation highlights larger $\mathbb{R}^n$s to work best when working with smaller target domains. When training on larger feature spaces (e.g., DS3, Table 5), correlations proved insignificant for both rTOST $p$ values (Kendall's $\tau = -0.21$, $p = 0.13$) and absolute difference calculations ($\tau = -0.29$, $p = 0.15$). Nevertheless, while correlations remain insignificant, smaller latent vectors were seen to create more predictable and stable data (Figure 5).

**Table 4.** Best obtained absolute difference and $p$ value calculations for robust equivalency testing. Values calculated comparing the original target distribution (DS1) with synthetic data generated by different GANs with different sized latent vectors as generator input.

| | *LSGAN* | | *WGAN* | | *WGAN-GP* | |
|---|---|---|---|---|---|---|
| | **\|*d*\|** | ***p*** | **\|*d*\|** | ***p*** | **\|*d*\|** | ***p*** |
| $\mathbb{R}^{25}$ | 0.337 | 0.726 | 0.179 | 0.014 | 0.199 | 0.017 |
| $\mathbb{R}^{50}$ | 0.116 | $2.0 \times 10^{-4}$ | 0.157 | 0.004 | 0.203 | 0.026 |
| $\mathbb{R}^{75}$ | 0.187 | 0.0105 | 0.075 | $2.9 \times 10^{-6}$ | 0.169 | 0.0013 |

**Table 5.** Best obtained absolute difference and $p$ value calculations for robust equivalency testing. Values calculated comparing the original target distribution (DS3) with synthetic data generated by different GANs with different sized latent vectors as generator input.

| | *LSGAN* | | *WGAN* | | *WGAN-GP* | |
|---|---|---|---|---|---|---|
| | $\|d\|$ | $p$ | $\|d\|$ | $p$ | $\|d\|$ | $p$ |
| $\mathbb{R}^{25}$ | 0.052 | $7.3 \times 10^{-20}$ | 0.039 | $6.7 \times 10^{-23}$ | 0.031 | $3.6 \times 10^{-22}$ |
| $\mathbb{R}^{50}$ | 0.032 | $1.7 \times 10^{-24}$ | 0.025 | $2.6 \times 10^{-25}$ | 0.025 | $1.0 \times 10^{-23}$ |
| $\mathbb{R}^{75}$ | 0.014 | $1.2 \times 10^{-25}$ | 0.043 | $3.9 \times 10^{-22}$ | 0.050 | $2.4 \times 10^{-20}$ |

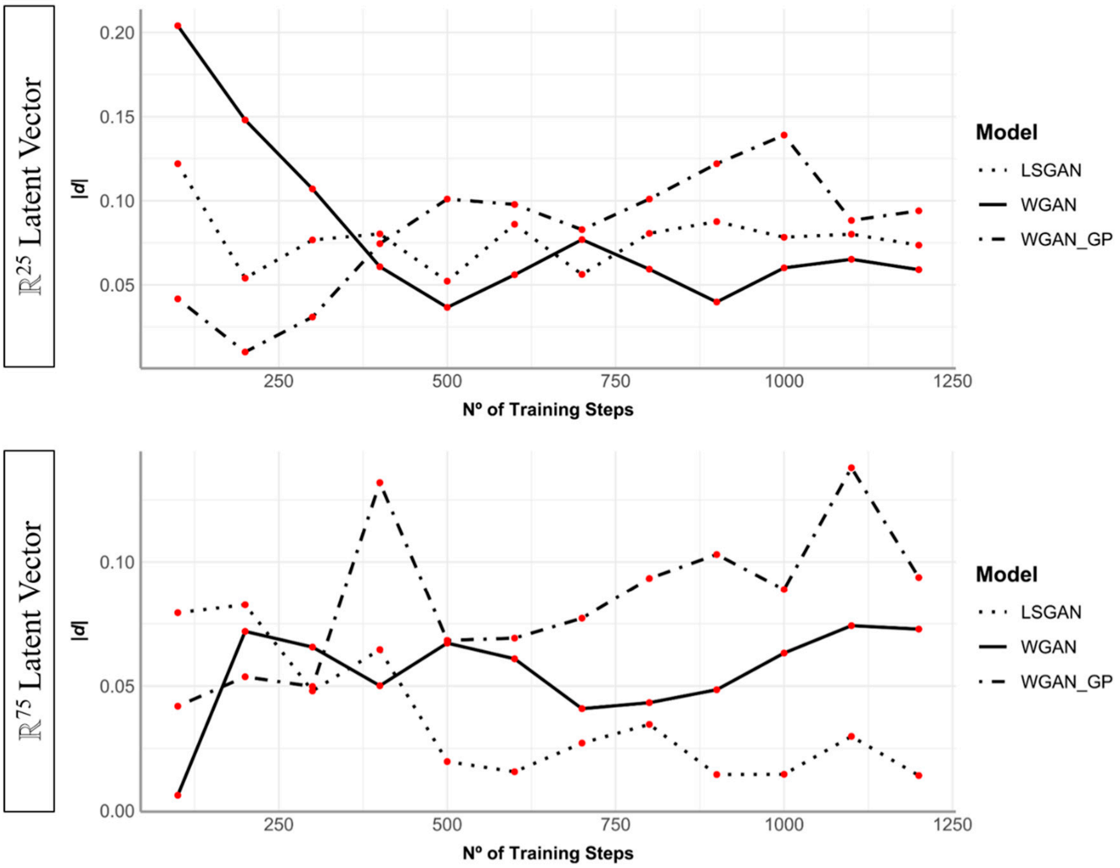


**Figure 5.** rTOST absolute difference values obtained comparing synthetic data generated over multiple training steps with the original distribution (example of DS3). Upper and lower panels compare different GANs using different sized latent vector inputs to the generator. GANs were trained for 400 epochs with batch sizes of 16. Here 3 training steps are considered an epoch.

In most experiments, 400 epochs were considered enough for GANs to produce realistic data. Moreover, the best results of each GAN began appearing after approximately 100-130 epochs. Performance significantly decreased, however, when trained on the same number of epochs using less data. To test this, subsets of each dataset were taken for experimentation (e.g., 30 out of 60 samples from DS2, Table 6). On all accounts, significant correlations were detected, finding smaller datasets to need more training time in order to obtain optimal results (Pearson's $r$ = 0.65, $p$ = 0.0005). Likewise, *LSGAN* appeared to be the model least affected by dataset size, producing the most realistic distributions in each of the cases (Table 6). While training GANs using 400 epochs is still able to produce realistic data on small datasets, when considering the optimal number of epochs, increasing this number to 1000 produces a significant improvement in results.

**Table 6.** Best obtained absolute difference and $p$ value calculations for robust equivalency testing after $x$ number of epochs. Example of GANs trained on a subset of 30 individuals from DS2.

| Results Obtained after $x$ Epochs | *LSGAN* | | *WGAN* | | *WGAN-GP* | |
|---|---|---|---|---|---|---|
| | $\lvert d\rvert$ | $p$ | $\lvert d\rvert$ | $p$ | $\lvert d\rvert$ | $p$ |
| 400 | 0.052 | $6.0 \times 10^{-16}$ | 0.098 | $2.4 \times 10^{-12}$ | 0.095 | $3.8 \times 10^{-12}$ |
| 800 | 0.059 | $3.3 \times 10^{-15}$ | 0.077 | $9.4 \times 10^{-13}$ | 0.075 | $6.02 \times 10^{-13}$ |
| 1000 | 0.013 | $3.8 \times 10^{-22}$ | 0.066 | $1.5 \times 10^{-18}$ | 0.054 | $5.0 \times 10^{-16}$ |

Figure 6 provides a visual summary of the absolute difference results obtained from each of these experiments (Tables 3–6, Figure 6).

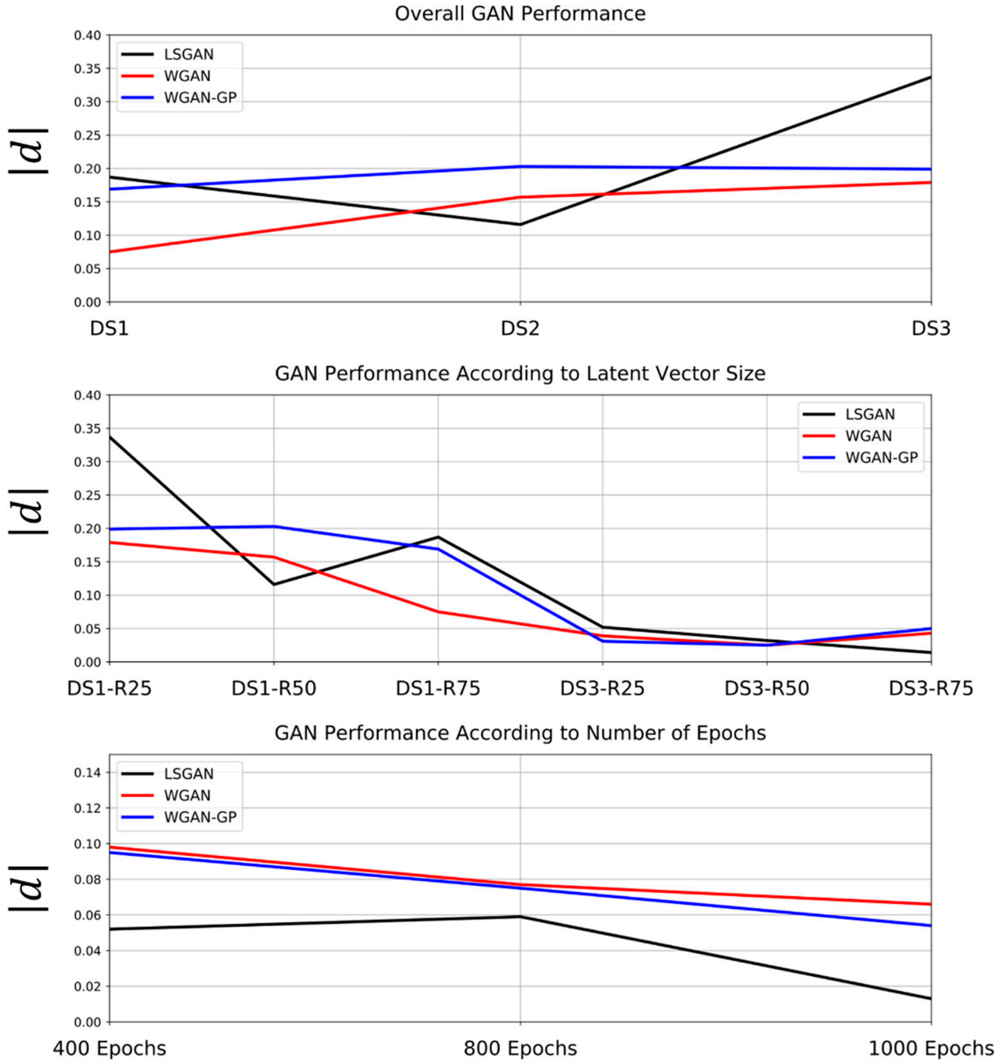


**Figure 6.** Parallel coordinate plots summarizing contents of Tables 3–6. Absolute difference values are obtained from robust equivalency testing.

### *3.3. Data Augmentation Results*

#### 3.3.1. General GAN Performance

All three GANs are highly successful in replicating sample distributions, effectively augmenting each of the distributions without too much distortion (Figure 7, Tables S1–S6). While evaluating standalone synthetic data creates some confusion, seen in some deviations of synthetic central tendency values and IR intervals (Tables S1, S3 and S5), the true value of GANs are observed when considering the augmented sample as a whole (Tables S2, S4 and S6).

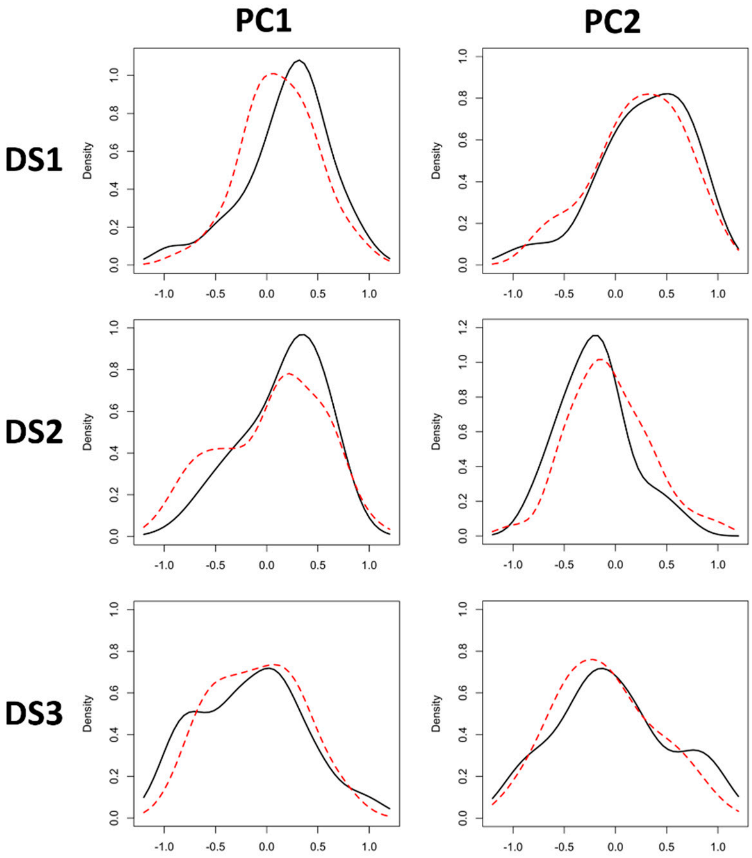


**Figure 7.** Plots of distributions both before (black) and after (red) GAN data augmentation, using the best performing models as described in Table 3. Descriptive statistics for each are included in Tables S2, S4 and S6.

In most cases, it can be seen how even the worst performing algorithms are able to maintain the central tendency of samples while boosting the variance represented. It is also important to highlight that, while in some cases central tendency can be seen to deviate slightly from the original distribution (e.g., *LSGAN* on DS2, Table S2), this is normally only true of one PC score and is still insignificant (rTOST $p < 0.05$). Some algorithms are also seen to affect the normality of sample distributions, creating distortion that is reflected in increased sample skewness. Nevertheless, these distortions are still unable to modify the general magnitude of similarities between synthetic and real data.

The greatest value of GANs can, therefore, be seen in increases in overall sample variance without significant distortion of the real sample's distribution. Deviation values and IR intervals increase, representing more variability without significantly shifting central tendency and without generating outliers. This shows how each algorithm is able to essentially "fill in the gaps" for each distribution while staying true to the original domain.

If GAN performance were to be compared with more traditional augmentation procedures, such as bootstrap, GANs can be seen to smooth out the distribution curve (Figure 8), creating a more general and complete mapping of the target domain. Bootstrapping procedures, on the other hand, tend to exaggerate gaps in the distribution. This can mostly be characterized by noticeable modifications to sample kurtosis while maintaining the general variation (Table 7).

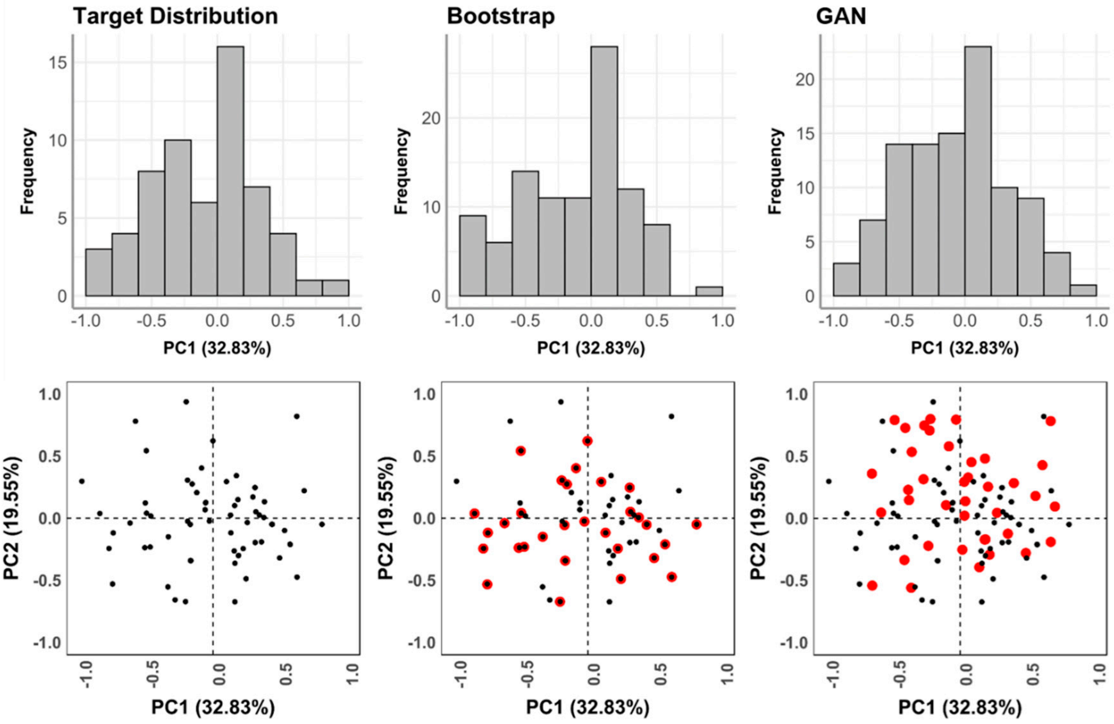


**Figure 8.** Histograms and scatter plots of augmented DS3 using bootstrap and GAN. New points are marked in red.

**Table 7.** Comparison of descriptive statistics obtained when comparing traditional bootstrapping procedures for numeric data augmentation and the best performing GAN on DS3. Dataset was augmented to size 100.

| | Original Data | | Bootstrap | | GAN | |
|---|---|---|---|---|---|---|
| | **PC1** | **PC2** | **PC1** | **PC2** | **PC1** | **PC2** |
| Shapiro $w$ | 0.99 | 0.98 | 0.98 | 0.98 | 0.99 | 0.98 |
| Shapiro $p$ | 0.75 | 0.30 | 0.25 | 0.18 | 0.92 | 0.10 |
| Central Tendency ‡ | −0.10 | −0.16 | −0.14 | −0.21 | −0.06 | −0.04 |
| Deviation § | 0.41 | 0.43 | 0.40 | 0.43 | 0.41 | 0.49 |
| Minimum Value | −1.00 | −1.00 | −1.00 | −1.00 | −1.00 | −1.00 |
| IR 0.05 Limit | −0.81 | −0.98 | −0.81 | −0.99 | −0.78 | −0.85 |
| IR 0.95 Limit | 0.47 | 0.61 | 0.47 | 0.52 | 0.55 | 0.84 |
| Maximum Value | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Kurtosis | −0.33 | 0.19 | −0.43 | −0.24 | −0.49 | −0.65 |
| Skewness | −0.02 | 0.35 | −0.01 | 0.21 | 0.05 | 0.18 |

‡ Metric used = mean; § Metric used = standard deviation.

### 3.3.2. Conditional GAN Performance

CGAN presented limited success when augmenting datasets, with only Wasserstein Gradient Penalty loss succeeding in overcoming mode collapse. Nevertheless, CGAN was still able to generate data with insignificant differences (Table 8), successfully augmenting the targeted datasets (Table 9 and Figure 9).

**Table 8.** Best obtained absolute difference and *p* value calculations for robust equivalency testing when comparing targeted generation of data using CGAN and GAN on DS3. Both CGAN and GAN were trained using Wasserstein Gradient Penalty loss.

| | **Sample 1** | | **Sample 2** | |
|---|---|---|---|---|
| | $\|d\|$ | $p$ | $\|d\|$ | $p$ |
| CGAN | −0.100 | $1.3 \times 10^{-8}$ | −0.035 | $2.0 \times 10^{-10}$ |
| GAN | −0.039 | $2.0 \times 10^{-10}$ | −0.074 | $2.9 \times 10^{-10}$ |

**Table 9.** Descriptive statistics for augmented DS3 targeting label values specifically. Numbers marked in bold indicate the synthetic data that obtained the most significant rTOST equivalency *p*-values. Both CGAN and GAN were trained using Wasserstein Gradient Penalty loss.

| | | **Original Data** | | **CGAN** | | **GAN** | |
|---|---|---|---|---|---|---|---|
| | | **PC1** | **PC2** | **PC1** | **PC2** | **PC1** | **PC2** |
| Sample 1 | Shapiro *w* | 0.98 | 0.97 | 0.97 | 0.97 | **0.98** | **0.98** |
| | Shapiro *p* | 0.75 | 0.59 | 0.25 | 0.25 | **0.31** | **0.31** |
| | Central Tendency ‡ | −0.03 | −0.33 | −0.14 | −0.26 | **0.003** | **−0.21** |
| | Deviation § | 0.43 | 0.37 | 0.41 | 0.42 | **0.47** | **0.34** |
| | Minimum Value | −0.87 | −1.00 | −0.89 | −1.00 | **−0.87** | **−1.00** |
| | IR 0.05 Limit | −0.79 | −0.99 | −0.81 | −0.98 | **−0.79** | **−0.98** |
| | IR 0.95 Limit | 0.64 | 0.13 | 0.47 | 0.52 | **0.69** | **0.22** |
| | Maximum Value | 1.00 | 0.52 | 1.00 | 0.66 | **1.00** | **0.62** |
| | Kurtosis | −0.44 | −0.49 | −0.49 | −0.53 | **−0.97** | **0.20** |
| | Skewness | 0.06 | −0.09 | 0.12 | 0.34 | **−0.02** | **−0.20** |
| Sample 2 | Shapiro *w* | 0.98 | 0.97 | 0.98 | 0.97 | **0.97** | **0.98** |
| | Shapiro *p* | 0.70 | 0.43 | 0.61 | 0.13 | **0.20** | **0.28** |
| | Central Tendency ‡ | −0.17 | 0.01 | −0.08 | 0.19 | **−0.14** | **0.03** |
| | Deviation § | 0.38 | 0.44 | 0.43 | 0.46 | **0.50** | **0.40** |
| | Minimum Value | −1.00 | −0.85 | −1.00 | −0.85 | **−1.00** | **−0.85** |
| | IR 0.05 Limit | −0.81 | −0.59 | −0.78 | −0.59 | **−0.85** | **−0.55** |
| | IR 0.95 Limit | 0.43 | 0.85 | 0.51 | 0.85 | **0.75** | **0.81** |
| | Maximum Value | 0.47 | 1.00 | 0.92 | 1.00 | **0.97** | **1.00** |
| | Kurtosis | −0.78 | −0.23 | −0.58 | −1.00 | **−0.79** | **−0.46** |
| | Skewness | −0.28 | 0.42 | 0.18 | −0.14 | **0.31** | **0.39** |

‡ Metric used = mean; § Metric used = standard deviation.

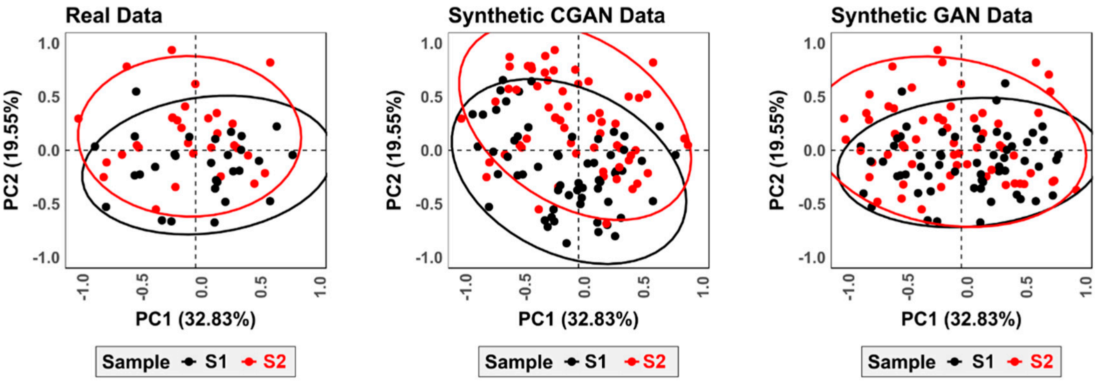


**Figure 9.** PCA scatter plot presenting data augmentation techniques on DS3. Ellipses mark 95% confidence intervals. Both CGAN and GAN were trained using Wasserstein Gradient Penalty loss.

When taking a closer look at the performance of CGAN, however, it is important to note that, while the magnitude of differences between synthetic and real data are insignificant, CGAN distorts the original distribution to a greater degree (Figure 9). In both samples, CGAN deviates greatly from the target central tendency (Table 9) and appears to shift the general skew of the distribution (Figure 9). When using GANs to augment each of the samples separately, however, the generated data is arguably truer to the original domain. This is not to say, however, that CGANs are unable to augment feature spaces successfully. With the right configuration, CGANs are likely to reach similar results to traditional GANs. This, however, goes beyond the scope of the present study.

## 4. Discussion

Many algorithms require large amounts of data in order to efficiently extract information, a task which is particularly difficult when considering data derived from the fossil record. To confront this topic, this study presents a new integration of AIAs into archaeological and paleontological sciences. Here GANs have been shown to be a new and valuable tool for the modelling and augmentation of GM data. Moreover, these algorithms can additionally be employed on a number of different types of datasets and applications; whether this be the handling of paleoanthropological, biological, taphonomic or lithic specimens via GM landmark data.

To demonstrate this latter point, and applying typical geometric morphometric techniques for classification on the O'Higgins and Dryden [75] Hominoid skull dataset, the present study is able to increment balanced accuracy of traditional LDA up to ca. 5% (Figure 10a) with a significant increase in generalization (Figure 10b,c). For this demonstration LDA was trained using a traditional approach [11], as well as an augmented approach based on Machine Teaching [43]. It can be seen how applying Machine Teaching using 100 realistic synthetic points per sample for training, and the original data was used for testing, helps the generalization process (Figure 10b) while providing clearer boundaries for each of the sample domains (Figure 10c).

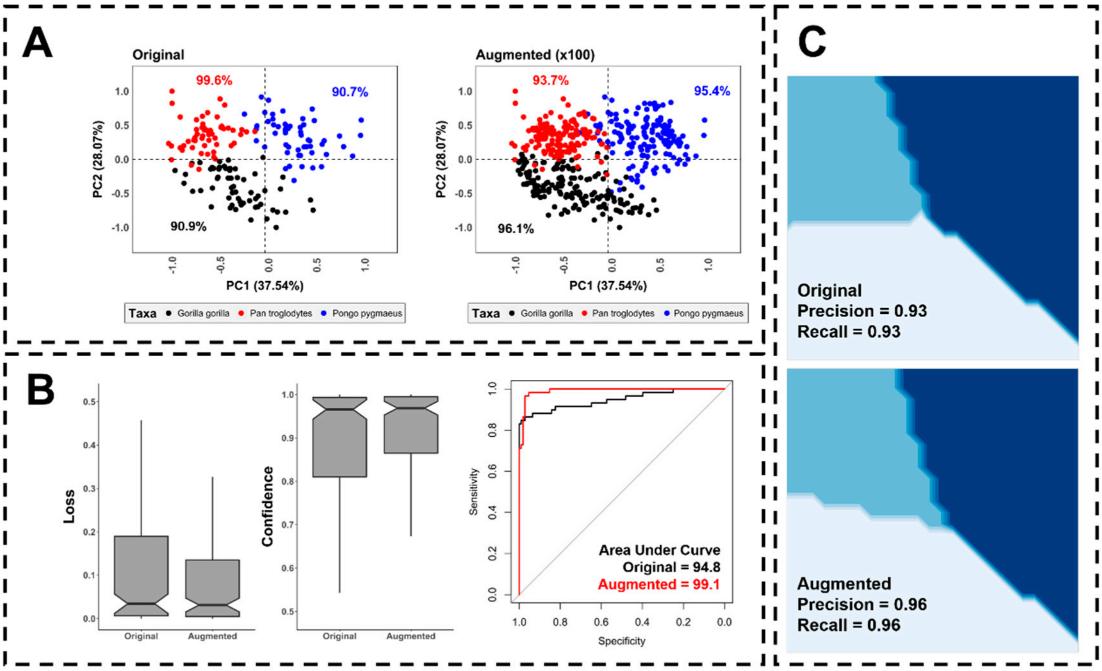


**Figure 10.** Example of the Augmentation (*LSGAN* ×100) of O'Higgins and Dryden's Hominoid dataset comparing gorilla (*Gorilla gorilla*), chimpanzee (*Pan troglodytes*) and orangutan (*Pongo pygmaeus*) skulls [75]. (**A**) Example of both the original and augmented datasets. Percentages next to each group represent the accuracy obtained by a more traditional Linear Discriminant Analysis algorithm. (**B**) Boxplots presenting the loss and confidence of predictions made using both the original and the augmented data set. To the right Receiver Operator Characteristic curves describe the overall performance of models on both datasets. (**C**) Decision boundaries drawn by LDA across PC1 (*x*-axis) and PC2 (*y*-axis). The upper panel and lower panel of C represent the original and the augmented datasets respectively.

It is important to point out, however, that this is not the solution to all sample-size related issues, and a number of components have to be discussed before more advanced applications can be carried out.

Missing data and the availability of fossil finds are a major handicap in prehistoric research. This is increasingly relevant when considering fossils of older ages, such as individuals of the *Australopithecus*, *Paranthropus* and early *Homo* genera. In a number of cases, for example, the representation of *Australopithecine* or *Homo erectus/ergaster* specimens may not even surpass 10 individuals [76–78], while *Homo sapiens* specimens are in abundance. In these cases, and in accordance with the data presented here, GANs would not be able to successfully augment the targeted minority distributions from scratch. Other options, however, could entail the use of algorithms for pre-processing, using variations of the Synthetic Minority Oversampling Technique (SMOTE & Borderline-SMOTE) [79–81], or the adapted version Adaptive Synthetic Sampling (ADASYN) [82].

Both SMOTE and ADASYN are useful, easy to implement algorithms that augment minority samples in imbalanced datasets. SMOTE generates synthetic data in the feature spaces that join data points (e.g., according to $k$ nearest-neighbor theory), thus filling in regions of the target domain [79–81]. ADASYN takes this a step further by modelling on sample distributions based on data-density [82]. Both are valuable algorithms that have become popular in imbalanced learning tasks, generally improving predictive model generalization. Nevertheless, their application should be confronted conservatively.

Preliminary experiments within this study found that resampling via bootstrapping prior to the training of GANs resulted in severe mode collapse. This can be theoretically explained by the manner in which bootstrapping is over-inflating the domain and highlighting very specific regions which the model then learns from. This results in overfitting as the model is repeatedly learning to map out the same value multiple times, boosting the probability of mode collapse through an enhanced lack of variation in the original trainset. Considering how SMOTE and ADASYN produce more "meaningful" data [44,79], these algorithms are more likely to aid the training process rather than produce the adverse effect. Nevertheless, overuse of SMOTE/ADASYN is likely to have a similar effect to bootstrapping, where linear regions of feature space between data points are enhanced while other regions are left empty.

Through this, the current study proposes that a conservative use of SMOTE or ADASYN variants prior to the training of GANs may be able to boost performance on overly-scarce datasets (e.g., the cases of [76–78]). This practice would be able to augment minority samples to a suitable threshold ($n$ = 30), preparing the dataset for more complex generative modelling and enabling an improved generalization of any predictive models used in analyses that follow.

From a similar perspective, the use of Bayesian Inference Algorithms such as Markov Chain Monte Carlo (MCMC) and Metropolis-Hastings algorithms have also been known to effectively model from multiple types of probability distributions [83–85]. In some cases, it may be possible to use these approaches to sample from the probability distribution at hand, and produce simulated information from the target distribution which would essentially be more realistic than simple bootstrap approaches. Further research into how these approaches may be applied could provide a powerful insight into GAN alternatives for different types of numerical data in GM.

In the general context of computational modelling, common criticism of neural network applications in archaeology and paleoanthropology argue that GM datasets are generally insufficient for the training of AIAs. This is based on the fact that most DL algorithms require much more data to avoid overfitting. From this point of view, why would training a GAN on such little data be any different? The present study proves that this is not an issue, considering how, with only 30 individuals, GANs are still able to produce highly realistic synthetic data (*LSGAN* rTOST $p = 3.8 \times 10^{-22}$).

In common DL literature, state-of-the-art models are reported to obtain ≈80% accuracy when trained on thousands to millions of specimens [25]. It is important to consider, however, that in the majority of these cases AIAs are trained on *images* (i.e., computer vision applications). To provide an example, Karras et al. [54] present a GAN capable of producing hyper-realistic fake images of people's faces, building from a subset of the CelebA-HQ dataset using ≈30,000 images. Two main components must be considered in order to understand why such a large dataset is required for their model;

- Karras et al. [54] present a GAN capable of producing high resolution 1024x1024 pixel RGB images. In computer vision applications, each image is conceptualized as a multidimensional numeric matrix (i.e., a tensor). Each of the numbers within the tensor can essentially be considered a variable, resulting in a dataset of approx. three million variables per individual photo.
- In order to efficiently map out these three million numeric values, the featured GAN uses progressively growing convolutional layers (nº layers ≈ 60) with 23.1 million adjustable parameters.
- The present study uses feature spaces that have already undergone dimensionality reduction derived from GM landmark data. In the case of the largest dataset, this results in a target vector of 60 variables that need to be generated. The present study additionally only uses fully connected layers with no convolutional filters, resulting in a model of <11,000 adjustable parameters. A GAN targeting three million values with 23.1 million parameters would thus require a far larger dataset than one targeting 60 values with 11 thousand parameters, explaining why with just 30 specimens, GAN convergence is still possible.

Regardless of the mathematics behind DL theory, the statistical results presented here provide enough empirical evidence to argue the value of the proposed GAN with as little as 30 individuals. Nevertheless, even in cases where datasets are too scarce for GANs to be developed from scratch, pre-trained models can be adjusted to different domains via multiple DL techniques. This arguably opens up new possibilities for the incorporation of Transfer Learning into GMs [25].

Finally, it is important to highlight how no absolute protocol can be established for generative modelling of any type. DL practitioners are usually required to adapt their model according to the dataset at hand, using the best practices established in other studies as a baseline from which to work from. Under this premise, recommendations established for the augmentation of GM datasets using GANs can be listed as follows:

- Best results are obtained when scaling the target domain to values between −1 and 1.
- Hidden layer densities should be adjusted according to the number of dimensions within the target domain (Figure 4). Tanh activation functions in both the generator and the discriminator are recommended.
- Dropout, batch normalization and kernel initializers (discriminator $\sigma = 0.1$, generator $\sigma = 0.7$) are recommended to regulate the learning process and avoid mode collapse.
- The Adam optimization algorithm is recommended when using Least-Square loss, while RMSProp is more efficient when using the Wasserstein (*WGAN* or *WGAN-GP*) function. A minimum batch size of 16 obtains the best results.
- *LSGAN* is recommended when training data is limited, increasing the number of epochs to at least 1000.
- *WGAN* and *WGAN-GP* work best on larger datasets, while approximately 400 epochs are usually enough to produce realistic data.
- The smaller the target domain, the larger the latent vector required for generator input.
- For conditional augmentation, optimal results are obtained by training GANs on each sample separately, rather than using CGANs.

## 5. Prospective and Future Research

While GANs are difficult to develop, the complexity of these AIAs should not be cause for discouragement. There currently exists a wide range of literature and helpful guides dedicated to teaching scientists about AIA development, even for those with no background in mathematics or applied statistics. With platforms such as *ScienceDirect* (https://www.sciencedirect.com/) in 2019 alone reporting over 3000 papers including the term Deep Learning (in keywords, title or abstract), and ca. 6000 for Machine Learning, AI can be considered one of the most popular lines of research in modern science. This presents a promising future for applications in archaeological and paleontological research.

Future lines will thus address the testing of GAN performance in actual archaeological and paleontological studies. Likewise, it would be highly recommendable to test whether these algorithms can effectively model data of a non-GM origin, employing the robust statistical techniques described here for more empirical evaluations of numeric data. Finally, comparisons with other augmentation techniques could help provide a more global view of the tools available for researchers.

## 6. Conclusions

To the authors' knowledge, this is the first comparative study in DL and GM using GANs for high dimensional numeric simulations that further employ advanced descriptive statistical metrics for evaluation. While augmented data is by no means a substitute for real data, real-life DL practices and applications have shown "meaningful" synthetic data to significantly increase the confidence and power of statistical models. In many cases, this has even been seen to exceed human-level precision.

The present paper has shown how GANs can be trained efficiently on numerical data obtained using geometric morphometrics. Under this premise, whether geometric morphometrics be used for the analysis of biological or non-biological individuals, GANs can effectively be used for data augmentation techniques prior to any consequent statistical modelling or learning tasks. The present study has additionally shown that three different loss functions are able to effectively learn from this type of data. Finally, robust statistical metrics have proven a valuable tool for performance evaluation.

**Supplementary Materials:** The following are available online at http://www.mdpi.com/2076-3417/10/24/9133/s1, Table S1: Descriptive statistics for DS1 of both target and synthetic sample distributions when compared separately. Table S2: Descriptive statistics for augmented DS1. Table S3: Descriptive statistics for DS2 of both target and synthetic sample distributions when compared separately. Table S4: Descriptive statistics for augmented DS2. Table S5: Descriptive statistics for DS3 of both target and synthetic sample distributions when compared separately. Table S6: Descriptive statistics for augmented DS3.

**Author Contributions:** Conceptualization, L.A.C.; methodology, L.A.C.; software, L.A.C.; validation, L.A.C. and D.G.-A.; formal analysis, L.A.C.; investigation, L.A.C.; resources, L.A.C.; data curation, L.A.C.; writing—original draft preparation, L.A.C.; writing—review and editing, L.A.C. and D.G.-A.; visualization, L.A.C.; supervision, D.G.-A.; project administration, D.G.-A.; funding acquisition, D.G.-A. All authors have read and agreed to the published version of the manuscript.

**Funding:** L.A.C. is funded by the Spanish Ministry of Science, Innovation and Universities with an FPI Predoctoral Grant (Ref. PRE2019-089411) associated to project RTI2018-099850-B-I00 and the University of Salamanca.

**Acknowledgments:** The corresponding author sincerely appreciates the support and advice of his directors; José Yravedra, Rosa Huguet and D.G.-A. We are also grateful to Jason Brownlee for his excellent work on Artificial Intelligence and his useful advice over multiple email conversations. The present paper was written during the national Spanish lockdown for the COVID-19 pandemic (March–June 2020). With this in mind we would like to thank all those who contributed towards protecting our health and safety in and outside of the health, pharmaceutical and civil sectors. L.A.C. would also like to thank Noé Valtierra, Guillermo (Máster) Blanco, Zaira López-Arias, Miguel Ángel Moreno-Ibañez, Enrique de Casimiro, Juan Hernandez-Rubio, Sara García-Motilla and his family for their support. We also thank Jordan Courtenay for her help proofreading earlier versions of this paper. Finally we thank two anonymous reviewers and the editorial staff of MDPI for their useful comments and suggestions to improve the quality of our paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Data Availability:** All data used in the current study has been fully sited and is available across multiple different data repositories. DS1 is available via a FigShare data repository (doi:10.6084/m9.figshare.c.4494218.v1). DS2 is included within supplementary files 1 and 3 of the associated paper [14]. DS3 is included within the corresponding author's GitHub page within the folder titled Landmark_Files of the https://github.com/LACourtenay/GMM_Measurement_Accuracy_Tools repository. The O'Higgins and Dryden [69] dataset is included within the R "shapes" package and can be called via the command data(apes).

**Code Availability:** All python code for GAN and CGAN applications are available via the author's GitHub page: https://github.com/LACourtenay/GMM_Generative_Adversarial_Networks.

## References

1. Bookstein, F.L. *Morphometric Tools for Landmark Data*; Cambridge University Press: Cambridge, UK, 1991.
2. Bookstein, F.L. Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Med. Image. Anal.* **1996**, *1*, 225–243. [CrossRef]
3. Dryden, I.L.; Mardia, K.V. *Statistical Shape Analysis*; John Wiley and Sons: New York, NY, USA, 1998.
4. Gunz, P.; Mitteroecker, P.; Bookstein, F.L. Semilandmarks in three dimensions. In *Modern Morphometrics in Physical Antrhopology*; Slice, D.E., Ed.; Plenum Publishers: New York, NY, USA, 2005; pp. 73–98.
5. Bookstein, F.L. Principal warps: Thin plate spline and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intel.* **1989**, *11*, 567–585. [CrossRef]
6. Ricthsmeier, J.T.; Lele, S.R.; Cole, T.M. Landmark morphometrics and the analysis of variation. In *Variation*; Hallgrimsson, B., Hall, B.K., Eds.; Elsevier Academic Press: Boston, MA, USA, 2005; pp. 49–68.
7. Rohlf, F.J. Statistical power comparisons among alternative morphometric methods. *Am. J. Phys. Antrhopol.* **2000**, *111*, 463–478. [CrossRef]
8. Klingenberg, C.P.; Monteiro, L.R. Distances and directions in multidimensional shape spaces: Implications for morphometric applications. *Soc. Syst. Biol.* **2005**, *54*, 678–688. [CrossRef] [PubMed]
9. Albrecht, G.H. Assessing the affinities of fossils using canonical variates and generalized distances. *J. Hum. Evol.* **1992**, *7*, 49–69. [CrossRef]
10. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. [CrossRef]
11. Mitteroecker, P.; Bookstein, F. Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evol. Biol.* **2011**, *38*, 100–114. [CrossRef]
12. Bocxlaer, B.V.; Schultheiß, R. Comparison of morphometric techniques for shapes with few homologous landmarks based on machine learning approaches to biological discrimination. *Paleobiology* **2010**, *36*, 497–515. [CrossRef]
13. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow*; O'Reilly: Tokyo, Japan, 2019.
14. Courtenay, L.A.; Yravedra, J.; Huguet, R.; Aramendi, J.; Maté-González, M.Á.; González-Aguilera, D.; Arriaza, M.C. Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth pits. *Palaeogeog. Palaeoclimatol. Palaeoecol.* **2019**, *522*, 28–39. [CrossRef]
15. Courtenay, L.A.; Huguet, R.; Yravedra, J. Scratches and grazes: A detailed microscopic analysis of trampling phenomena. *J. Microsc.* **2020**, *277*, 107–117. [CrossRef]
16. Yravedra, J.; Maté-González, M.Á.; Courtenay, L.A.; González-Aguilera, D.; Fernández-Fernández, M. The use of canid tooth marks on bone for the identification of livestock predation. *Sci. Rep.* **2019**, *9*, 16301. [CrossRef] [PubMed]
17. Dobigny, G.; Baylac, M.; Denys, C. Geometric morphometrics, neural networks and diagnosis of sibling *Taterillus* species (*Rodentia, Gerbillinae*). *Biol. J. Linnean Soc.* **2002**, *77*, 319–327. [CrossRef]
18. Baylac, M.; Villemant, C.; Simbolotti, G. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. *Biol. J. Linnean Soc.* **2003**, *80*, 89–98. [CrossRef]
19. Lorenz, C.; Ferraudo, A.S.; Suesdek, L. Artificial Neural Network applied as a methodology of mosquito species identification. *Acta Trop.* **2015**, *152*, 165–169. [CrossRef]
20. Soda, K.J.; Slice, D.E.; Naylor, G.J.P. Artificial neural networks and geometric morphometric methods as a means for classification: A case-study using teeth from *Carcharhinus* sp. (*Carcharhinidae*). *J. Morphol.* **2017**, *278*, 131–141. [CrossRef]
21. Courtenay, L.A.; Huguet, R.; González-Aguilera, D.; Yravedra, J. A Hybrid Geometric Morphometric Deep Learning approach for cut and trampling mark classification. *Appl. Sci.* **2020**, *10*, 150. [CrossRef]
22. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
23. Bishop, C. *Pattern Recognition and Machine Learning*; Springer: Singapore, 2006.
24. Bishop, C. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, NY, USA, 1995.
25. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
26. Muñoz-Muñoz, F.; Perpiñán, D. Measurement error in morphometric studies: Comparison between manual and computerized methods. *Ann. Zool. Fennici.* **2010**, *47*, 46–56. [CrossRef]
27. Cramon-Taubadel, N.; Frazier, B.C.; Lahr, M.M. The problem of assessing landmark error in geometric morphometrics: Theory methods and modifications. *Am. J. Phys. Anthropol.* **2017**, *134*, 24–35. [CrossRef]

28. Robinson, C.; Terhune, C.E. Error in geometric morphometric data collection: Combining data from multiple sources. *Am. J. Phys. Anthropol.* **2017**, *164*, 62–75. [CrossRef]
29. Courtenay, L.A.; Herranz-Rodrigo, D.; Huguet, R.; Maté-González, M.Á.; González-Aguilera, D.; Yravedra, J. Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy. *PLoS ONE* **2020**. [CrossRef] [PubMed]
30. Bookstein, F.L. Introduction to Methods for Landmark Data. In Proceedings of the Michigan Morphometrics Workshop; Bookstein, F.L., Rohlf, F.J., Eds.; University of Michigan Museum of Zoology: Ann Arbor, MI, USA, 1990; pp. 215–225.
31. Devine, J.; Aponte, J.D.; Katz, D.C.; Liu, W.; Vercio, L.D.L.; Forkert, N.D.; Marcucio, R.; Percival, C.J.; Hallgrímsson, B. A registration and Deep Learning approach to automated landmark detection for geometric morphometrics. *Evol. Biol.* **2020**, *47*, 246–259. [CrossRef]
32. García-Medrano, P.; Ollé, A.; Ashton, N.; Roberts, M.B. The mental template in handaxe manufacture: New insights into Acheulean lithic technological behavior at Boxgrove, Sussex, UK. *J. Archaeol. Meth. Theor.* **2018**, *26*, 396–422. [CrossRef]
33. Gunz, P.; Mitteroecker, P.; Bookstein, F.L.; Weber, G.W. Computer aided reconstruction of incomplete human crania using statistical and geometrical estimation methods. In *Enter the Past: Computer Applications and Quantitative Methods in Archeology*; Stadt Wien, M., Erbe, R.K., Wien, S., Eds.; BAR Internatal Series: Oxford, UK, 2004; Volume 1227, pp. 92–94.
34. Gunz, P.; Mitteroecker, P.; Neubauer, S.; Weber, G.W.; Bookstein, F.L. Principles for the Virtual Reconstruction of Hominin Crania. *J. Hum. Evol.* **2009**, *57*, 48–62. [CrossRef]
35. Cohen, J. *Statistical Power Analysis for Behavioural Sciences*; Routledge: New York, NY, USA, 1988.
36. Fisher, R.A. *The Design of Experiments*; Hafner Pub: New York, NY, USA, 1935.
37. Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341. [CrossRef]
38. Ho Yu, C. Resampling methods: Concepts, applications and justification. *Prac. Assess. Res. Eval.* **2003**, *8*, 1–17. [CrossRef]
39. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer: Heidelberg, Germany, 2018.
40. Efron, B. Bootstrap methods: Another look at the jackknife. *Annals Stat.* **1979**, *7*, 1–26. [CrossRef]
41. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, USA, 1993.
42. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: Gewerbestrasse, Switzerland, 2016.
43. Such, F.P.; Rawal, A.; Lehman, J.; Stanley, K.O.; Clune, J. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. *arXiv* **2019**, arXiv:1912.07768v1.
44. Tanaka, F.H.K.S.; Aranha, C. Data Augmentation using GANs. *arXiv* **2019**, arXiv:1904.09135v1.
45. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *arXiv* **2014**, arXiv:1406.2661v1.
46. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Essen, B.C.V.; Awwal, A.A.S.; Asari, V.K. A state-of-the-art survey on Deep Learning theory and architectures. *Electronics* **2019**, *8*, 292. [CrossRef]
47. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for Deep Learning. *J. Big Data* **2019**, *6*. [CrossRef]
48. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2016**, arXiv:1511.06434v2.
49. Saliman, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. *arXiv* **2016**, arXiv:1606.03498v1.
50. Lucic, M.; Kurach, K.; Michalski, M.; Bousquet, O.; Gelly, S. Are GANs created equal? A large scale study. *arXiv* **2018**, arXiv:1711.10337v4.
51. Goodfellow, I. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv* **2016**, arXiv:1701.00160v4.
52. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875v3.
53. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of Wasserstein GANs. *arXiv* **2017**, arXiv:1704.00028v3.
54. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of GANs for improved quality, stability and variation. *arXiv* **2018**, arXiv:1710.10196v3.

55. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least Square Generative Adversarial Networks. *arXiv* **2017**, arXiv:1611.04076v3.
56. Fedus, W.; Rosca, M.; Lakshminarayanan, B.; Dai, A.M.; Mohamed, S.; Goodfellow, I. Many paths to equilibrium: GANs do not need to decreate a divergence at every step. *arXiv* **2018**, arXiv:1710.08446v3.
57. Hinton, G. Neural Networks for Machine Learning Technical Report. Available online: https://www.cs.toronto.edu/~{}tijmen/csc321/slides/lecture_slides_lec6.pdf (accessed on 6 November 2020).
58. Kingma, D.P.; Lei Ba, J. Adam: A method for stochastic optimization. *arXiv* **2015**, arXiv:1412.6980v9.
59. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784v1.
60. Denton, E.; Chintala, S.; Szlam, A.; Fergus, R. Deep generative image models using a Laplacian pyramid of adversarial networks. *arXiv* **2015**, arXiv:1506.05751v1.
61. Borji, A. Pros and cons of GAN evaluation metrics. *arXiv* **2018**, arXiv:1802.03446v5.
62. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv* **2017**, arXiv:1612.03242v2.
63. Diaconsis, P.; Freedman, D. Asymptotics of Graphical Projection of Pursuit. *Ann. Stat.* **1984**, *12*, 793–815. [CrossRef]
64. Lakens, D. Equivalence tests: A practical primer for t tests, correlations and meta analyses. *Soc. Phychol. Pers. Sci.* **2017**, *8*, 355–362. [CrossRef]
65. Dienes, Z. How bayes factor change scientific practice. *J. Math. Psychol.* **2016**, *72*, 78–89. [CrossRef]
66. Hauk, D.W.W.; Anderson, S. A new statistical procedure for testing equivalence in two-group comparative biovariability trials. *J. Pharm. Biopharm.* **1984**, *12*, 83–91. [CrossRef]
67. Anderson, S.F.; Maxwell, S.E. There's more than one way to conduct a replication study: Beyond statistical significance. *Psychol. Methods* **2016**, *21*, 1–12. [CrossRef] [PubMed]
68. Schurimann, D.L. A comparison of the two one-sided test procedure and the power approach for assessing the equivalence of average biovariability. *J. Pharm. Biopharm.* **1987**, *15*, 657–680. [CrossRef] [PubMed]
69. Yuen, K.K.; Dixon, W.J. The approximate behaviour and performance of the two-sample trimmed t. *Biometrika* **1973**, *60*, 369–374. [CrossRef]
70. Yuen, K.K. The two-sample trimmed t for unequal population variances. *Biometrika* **1974**, *61*, 165–170. [CrossRef]
71. Höhle, J.; Höhle, M. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS J. Photogram. Rem. Sens.* **2009**, *64*, 398–406. [CrossRef]
72. Rodríguez-Martín, M.; Rodríguez-Gonzálvez, P.; Ruiz de Oña Crespo, E.; González-Aguilera, D. Validation of portable mobile mapping system for inspection tasks in thermal and fluid-mechanical facilities. *Remote Sens.* **2019**, *11*, 2205. [CrossRef]
73. Pearson, K. Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* **1895**, *58*, 347–352. [CrossRef]
74. Kendall, M.G. *Rank Correlation Methods*; Hafner Publishing, Co.: New York, NY, USA, 1955.
75. O'Higgins, P.; Dryden, I.L. Sexual dimorphism in hominoids: Further studies of craniofacial shape differences in Pan, Gorilla and Pongo. *J. Hum. Evol.* **1992**, *24*, 183–205. [CrossRef]
76. Wu, L.; Clarke, R.; Song, X. Geometric morphometric analysis of the early Pleistocene hominin teeth from Jianshi, Hubei Province, China. *Sci. China Earth Sci.* **2010**, *53*, 1141–1152. [CrossRef]
77. Freidline, S.E.; Gunz, P.; Janković, I.; Harvati, K.; Hublin, J.J. A comprehensive morphometric analysis of the frontal and zygomatic bone of the Zuttiyeh fossil from Israel. *J. Hum. Evol.* **2012**, *62*, 225–241. [CrossRef]
78. Détroit, F.; Mijares, A.S.; Corny, J.; Daver, G.; Zanolli, C.; Dizon, E.; Robles, E.; Grün, R.; Piper, P.J. A new species of *Homo* from the Late Pleistocene of the Philippines. *Nature* **2019**, *568*, 181–186. [CrossRef] [PubMed]
79. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
80. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing*; Huan, D.S., Xiao-Ping, Z., Huang, G.B., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; Part 1; pp. 878–887.
81. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. *IEEE Int. Workshop Comput. Intell. Appl.* **2009**, *3*, 24–29. [CrossRef]

82. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive Synthetic Sampling approach for Imbalanced Learning. In Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008. [CrossRef]
83. Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092. [CrossRef]
84. Hastings, W. Monte Carlo sampling methods using Markov chains and their application. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
85. Gamerman, D.; Lopes, H.F. *Markov Chain Monte Carlo*; Chapman & Hall: Boca Raton, FL, USA, 2006.

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

*Spanish Translation of Title and Abstract*

# Reclutando a una tripulación de esqueletos – Métodos para la simulación y aumentación de datos paleoantropológicos mediante el uso de algoritmos tipo Monte Carlo.

La recogida de datos es uno de los mayores obstáculos en muchos tipos de análisis dentro del campo de la evolución humana. Aunque el aumento de conjuntos de datos y repositorios de libre acceso ha ayudado a paliar estas limitaciones, algunos investigadores, en un ejemplo de mala praxis científica, siguen siendo reacios a compartir datos. Esta cuestión resulta aún más importante si se tiene en cuenta la escasez y calidad de los datos. Desde esta perspectiva, muchos proyectos de investigación están limitados por la cantidad de datos que pueden tener a la hora de realizar tareas como la clasificación y la modelización predictiva. En este artículo se presenta el uso de métodos basados en algoritmos de tipo *Monte Carlo*, incluyendo el algoritmo *Markov Chain Monte Carlo*, para la simulación de datos paleoantropológicos. Utilizando dos conjuntos de datos, mostramos cómo se pueden simular datos sintéticos, pero realistas, para mejorar cada conjunto de datos y proporcionar nueva información con la que realizar tareas más complejas. Además, presentamos estos algoritmos en una nueva librería de R: *AugmentationMC*. A partir de un conjunto de datos multimodal que contiene datos categóricos y numéricos, mostramos cómo la biblioteca *AugmentationMC* puede mejorar la variabilidad de la muestra sin cambiar significativamente la naturaleza de la distribución de probabilidad. También empleamos un conjunto de datos de morfometría geométrica para simular modelos 3D. Por último, enfatizamos el poder de la enseñanza máquina (*Machine Teaching*), en contraposición al aprendizaje máquina (*Machine Learning*). Aunque los conjuntos de datos sintéticos nunca deberían sustituir a los grandes conjuntos de datos, esto puede considerarse un avance importante en la forma de manejar datos paleantropológicos.

Submitted to **American Journal of Biological Anthropology**

*Supplementary Information and Links*

**Current version of Supplementary Information is available from:**
https://figshare.com/s/e163c3f1232fb48bae84

**R library and code available from:**
https://github.com/LACourtenay/AugmentationMC

# Recruiting a Skeleton Crew – Methods for Simulating and Augmenting Paleoanthropological Data using Monte Carlo based Algorithms.

Lloyd A. Courtenay [a, b, *], Julia Aramendi [c], Diego González-Aguilera [a]

[a] Department of Cartographic and Land Engineering, Higher Polytechnic School of Avila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain
[b] Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain
[c] Department of Geology, Facultad de Ciencia y Tecnología, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa, Spain.

***Corresponding author**
Email: ladc1995@gmail.com
Mobile: +34 633 647 825
ORCID ID: 0000-0002-4810-2001

## Abstract

### Objectives

Data collection is a major hindrance in many types of analyses in human evolutionary studies. This issue is fundamental when considering the scarcity and quality of fossil data to begin with. From this perspective, many research projects are impeded by the amount of data they may have to perform tasks such as classification and predictive modelling.

### Materials and Methods

Here we present the use of Monte Carlo based methods, including the Markov Chain Monte Carlo algorithm, for the simulation of paleoanthropological data. Using two datasets, we show how synthetic, yet realistic, data can be simulated to enhance each dataset, and provide new information from which to perform complex tasks with, in particular classification. We additionally present these algorithms in the form of an R library; *AugmentationMC*. We also use a geometric morphometric dataset to simulate 3D models, and emphasize the power of Machine Teaching, as opposed to Machine Learning.

### Results

Our results show how Markov Chain Monte Carlo algorithms, and Monte Carlo algorithms, are useful for the simulation of this type of data, providing synthetic yet highly realistic data that has been tested statistically to be equivalent to the original data. We additionally provide a critical overview of bootstrapping techniques, showing how Monte Carlo based methods excell over bootstrapping as the data simulated is not an exact copy of the original sample.

### Discussion

While synthetic datasets should never replace large and real datasets, this can be considered an important advance in how paleoanthropological data can be handled.

**Key Words:** Data Augmentation, Geometric Morphometrics, Markov Chain Monte Carlo, Machine Teaching, 3D Model Simulation.

## 1. Introduction

The incomplete nature of the fossil record is one of the greatest obstacles that archaeologists, paleontologists, and paleoanthropologists, have to face. Considering the number of available fossils for certain species, most studies are hindered by a distorted view of the population due to small samples. While this is an inevitable phenomenon, especially for sites of an older geological age, sample size conditions the quality of the analyses and results that can be extracted from this type of data. This is increasingly relevant in contemporary research, where data science and advanced statistics are at the forefront of virtual paleoanthropological analyses.

All statistical tests are conditioned by sample size, regardless of the field of science (Cohen, 1988). This is especially relevant in paleoanthropology, when considering how some species are only known via a single, or at most, a handful, of specimens. Even in cases where multiple individuals are available, the likelihood that the same region or even the same anatomical element will coincide between two specimens is extremely low. This is even before taking into account the possible distortion or damage this fossil may have suffered over time (Gunz, Mitteroecker, Bookstein & Webber, 2004; Gunz, Mitteroecker, Neubaeuer, Weber & Bookstein, 2009). Any type of calculation that considers sample variance or covariance matrices (e.g., Mahalanobis distances, Analyses of Variance), are also fundamentally conditioned by the amount of information available (Greenwood & Sandomire, 1950; Gupta and Gupta, 1987; Cohen, 1988; Takeshita, Nozawa & Kimura, 1993). This increases as the number of dimensions (or essentially the number of variables) increases (Gupta & Gupta, 1987; Bookstein, 2017, 2019, Cardini, O'Higgins & Rohlf, 2019), and is evidently important when estimating parameters of normality or general and descriptive statistics (LaPlace, 1823; Pearson, 1900; Gosset, 1908; Razali, 2011). Building from this, Principal Components Analyses (PCA) are also affected by sample size (Takeshita et al., 1993; Jollife, 2002; Bishop, 2006, Bookstein, 2017), while statistical tests such as Canonical Variance Analyses (CVA), and between-group PCA (bg-PCA), are highly sensitive to data size and balance (Albrecht, 1992; Mitteroecker and Bookstein, 2011; Bookstein, 2019; Cardini et al., 2019). Finally, any type of Discriminant Analyses (e.g., Linear Discriminant Analysis) is strongly hindered by a lack of information density.

In most statistical applications, resampling techniques have proven to be a popular means of estimating population parameters robustly, even when data is limited (Fisher, 1935; Efron, 1979). Resampling consists in randomly drawing observations from a dataset. This can either be done by extracting subsets of data (e.g., Jackknife and cross-validation), or drawing an observation with (bootstrap), or without (permutation), replacing it. These procedures are valid ways of emulating what the general nature of the population may look like. Nevertheless, these techniques are only able to duplicate this data, as opposed to generating new information.

The duplication of information in this way can be problematic for a number of reasons. Firstly, inferences and exploratory analyses are still limited to the information density available (Fernández et al., 2018). This, in turn, leads to the concept of *overfitting*, where we are unable to extract information from the true coverage of a domain, making the processing of new information difficult. From this perspective, duplicating data can be considered insufficient when trying to construct a more generalized view of the population (Courtenay & González-Aguilera, 2020; McPherron, Archer, Otárola-Castillo, Torquato & Keevil, 2022).

Beyond the Frequentist perspective, Bayesian statisticians often use inference engines for the approximation of probability distributions (Gamerman & Lopes, 2006). In this sense, Bayesian statistics uses Bayes' theorem to calculate an updated probability of a set of observations, based on the information provided, and a prior probability typically used to account for uncertainty (Martin, 2018). The final calculated probability distribution is then sampled from and used to infer or explain the original distribution, often based on Monte Carlo methods (Gordon, Salmond & Smith, 1993). Because this approach is based on distributions, as opposed to the observations themselves, Bayesian approaches produce observations that are not exact duplicates of the original dataset, providing a more generalized view of the possible original domain.

While Bayesian based algorithms are powerful approaches to estimating properties of the population, they are often computationally expensive, and can be difficult to implement (Carlin & Louis, 2008; Martin, 2018). Likewise, some criticism towards the Bayesian perspective considers this approach to statistics somewhat subjective, due to the nature of prior probability definitions. Needless to say, and with regards to the latter point, prior probabilities can easily be defined as diffuse, or weakly informative, diminishing the subjectivity of this type of approach (Carlin & Louis, 2008; Martin, 2018).

Building from this, Data Augmentation can be defined as a means of increasing a given sample by producing synthetic (yet realistic) data that is meaningful to the original distribution (Tanaka and Aranha, 2019; Courtenay & González-Aguilera, 2020). Computational tools can be developed to approximate a distribution, and estimate what the population may look like without creating exact copies of the original information.

Courtenay and González-Aguilera (2020) recently proposed a Neural Network (NN) and unsupervised deep learning approach for the simulation of new geometric morphometric data, paying particular attention to the use of such methods for the improvement of classification algorithm performance. In this study, these authors used Generative Adversarial Networks (GAN) with a Gradient Penalty Wasserstein loss function to simulate new information (Goodfellow et al., 2014; Gulrajani, Ahmed, Arjovsky, Dumoulin, Courville, 2017), useful for more efficient classification tasks. The present analysis develops from this, proposing the use of more computationally and statistically efficient Monte Carlo (MC) based methods. Additionally, we expand beyond this by including the simulation of both quantitative and qualitative data, not exclusive to geometric morphometrics, and two important sources of information that are highly relevant in paleoanthropological and archaeological research. To ensure reproducibility, here we present a new R library, named "*AugmentationMC*", for the implementation of MC and Markov Chain Monte Carlo (MCMC) based methods for data augmentation.

## 2. Materials and methods

This study presents the use of MC and MCMC for the augmentation of both qualitative and quantitative data. For this purpose, we build on research presented by Courtenay and González-Aguilera (2020), as well as Courtenay and colleagues (2021), to define a protocol for the simulation and evaluation of synthetic data. Once generated, this data can be used for more complex applications such as those typically used in computational learning.

### 2.1. Datasets

In order to prove the efficiency and power of MC and MCMC based approaches, the present study was performed using 2 separate datasets containing different types of data (Fig. 1). Each of these datasets are paleoanthropological in nature, and can be used to describe the morphological and mechanical properties of great ape femora.

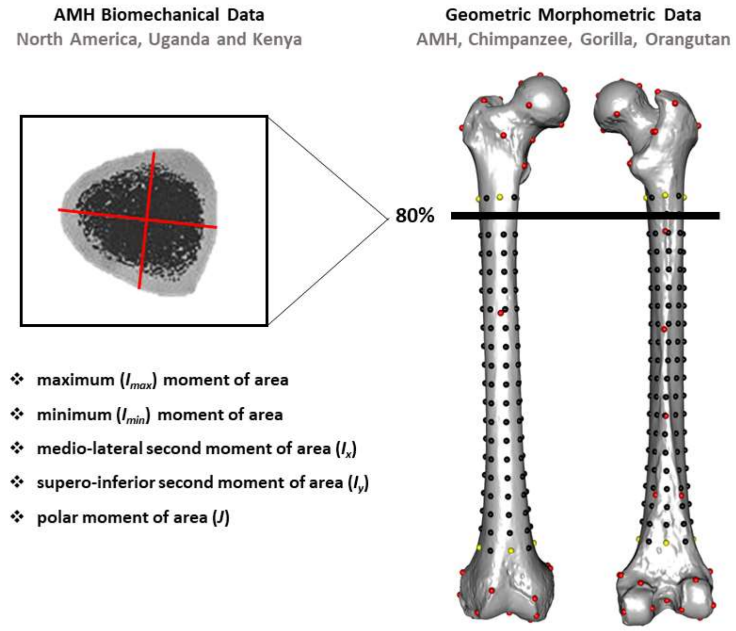

AMH Biomechanical Data
North America, Uganda and Kenya
Geometric Morphometric Data
AMH, Chimpanzee, Gorilla, Orangutan
80%
❖ maximum ($I_{max}$) moment of area
❖ minimum ($I_{min}$) moment of area
❖ medio-lateral second moment of area ($I_x$)
❖ supero-inferior second moment of area ($I_y$)
❖ polar moment of area ($J$)

***Figure 1**. Left: original Anatomically Modern Human (AMH) biomechanical data used in the present study were extracted at the 80% length of the femur from the proximal shaft end by Ruff (1995); Right: original landmark data were obtained for the present study using 40 fixed anatomical landmarks (red) and a dense point net (n = 160) comprised of 8 fixed points (yellow) meant to limit the projection and sliding process of the remaining 152 sliding semilandmarks (black).*

### 2.1.1. Dataset 1: Ruff (1995) femur biomechanical dataset.

This dataset contains information ranging from a number of publications, including data from Ruff & Hayes (1983), Ruff (1991, 1995), and Ruff and colleagues (2018), provided by the corresponding author of each of these papers (C.B. Ruff). The sample used in the present study consists of 100 femora, 60 from the Pecos Pueblo skeletal collections (Ruff & Hayes, 1983; Ruff, 1991), housed in the Harvard Peabody Museum by the time those studies were performed, and 40 East African individuals, 10 of which are from the Makerere University (Kampala, Uganda), and 30 from the Kenyan National Museum (Nairobi, Kenya) (Ruff, 1995). Individuals range in age between 20 and 55 years and include both male and female individuals. North American individuals were recovered from a site in New Mexico (1300-1828 CE), and are believed to have been dedicated to agricultural labors. The occupation of Eastern African individuals is unknown, however considering the time the collection was curated (1960), and the tribes and cultures these individuals were associated with (Acholi, Lango, Etesot, Lugbara, Maasai & Kikuyu), it can be assumed that these individuals were not highly urbanized, and likely dedicated to manual labor. A total of 51 males and 49 females are included within this dataset. Nowadays the use of recent human populations in anthropological studies is questioned as it raises ethical issues. We are aware of the problematic of this type of studies and agree on the need of returning the remains of the ancestors to their descendant communities. That is why we would like to stress that no new study was performed on these materials by the authors on the American and East African populations, neither were their data used to infer any kind of anthropological conclusions on these populations. Only raw data extracted several years ago by C.B. Ruff which has already published was used with the solely purpose of data augmentation.

The data obtained from these individuals was extracted from the 80% cross-section (subtrochanteric) at the proximal end of the femoral head. For the North American sample, these cross-sections were originally obtained in Ruff & Hayes (1983) by directly sectioning femora using a water-cooled diamond-bladed circular saw. For the East African sample, cross-sections were extracted using Computed Tomographic (CT) scans. From these cross-sections, 5 quantitative measurements were included within this portion of the study (*sensu* Ruff, 1987). These include the maximum ($I_{max}$) and minimum ($I_{min}$) moments of area, used to calculate the relative maximum bending strength of the bone, as well as the second moments of area around the medio-lateral breadth ($I_x$), and supero-inferior breadth ($I_y$). Finally, the polar moment of area ($J$) was also provided by this dataset (Fig. 1). For a more detailed description of these variables and how they are extracted see Ruff & Hayes (1983) and Ruff (1995).

### 2.1.2. Dataset 2: femur geometric dataset

The geometric morphometric dataset contains 80 femora belonging to four modern great ape genera; *Pan, Pongo, Gorilla* and *Homo*. Non-human samples were compiled from two online open-source archives; the Digital Morphology Museum KUPRI, as well as MorphoSource. Both sources provided 3D models in the form of CT scans. Samples downloaded from KUPRI include specimens that have already been segmented and cropped, as well as whole bodies or limbs. In the case of MorphoSource, specimens were already cropped, and in some cases rendered as *ply* or *stl* files. The human sample was obtained through the New Mexico Decedent Image Database (NMDID), which complies with the Declaration of Helsinki and thus meets the ethical standards that serve to promote and ensure respect for all human beings and to protect their health and individual rights. NMDID contains complete body CT scans of each individual that had to be segmented in order to isolate the right femora. The final sample thus consists of 30 Anatomically Modern Human (AMH), 36 Chimpanzee (C), 9 Gorilla (G), and 5 Orangutan (O) femora (see Supplementary File 1 for further details on sample processing and sources).

A total of 200 points were used to describe the 3D models of the femora, including a set of 40 fixed anatomical landmarks (red points in Fig. 1), and 152 sliding semilandmarks (black points in Fig. 1) constrained by 8 fixed points (yellow points in Fig. 1) to describe the morphology of the diaphysis. The process of template creation involves the digitization of a semilandmark mesh, as explained in Aramendi (2021). This template consists of a geometric model based on a cylinder, so as to facilitate the positioning of the semilandmark net, which is then used to project and slide points along the surface of the diaphysis. All 160 points are distributed in 20 equally spaced rings along the cylinder, with each horizontal ring containing 8 homogeneously distributed points (see Sup. File 1 for further details). This process was performed using the EVAN toolbox (http://evan-society.org/).

Once digitized, landmark data was processed using traditional geometric morphometric techniques. First coordinates were superimposed by means of a full Procrustes fit (Bookstein, 1991; Dryden & Mardia, 1998), using the *shapes* R library. Superimposed landmarks were then subjected to dimensionality reduction via PCA. Finally, the corresponding feature spaces were analyzed to assess the degree of represented morphological variance, extracting only those principal component (PC) scores representing up to 90% of morphological variance. This resulted in a final representation of each individual as $\mathbb{R}^{12}$ vectors containing morphological shape variables.

### 2.2. Data augmentation algorithms

### 2.2.1. Monte Carlo algorithms

MC computing methods include a group of algorithms that employ the use of random sampling to extract data from a probability distribution. This is a highly popular technique in the world of data simulation, proving particularly efficient at solving mathematical and physical problems, with ample applications in ecology (Giró, Padró, Valls & Wagensberg, 1985), chemistry (Souza, Matos, Oliveira-Neto & Albuquerque, 2020), and engineering (García-Martín, Bautista-De Castro, Sáncehz-Aparicio, Fueto & González-Aguilera, 2018; Pisonero et al., 2021), among others.

MC algorithms work by computing a set of possible observations by modelling from a probability distribution. As MCs work with distribution densities, they can randomly sample a value taking into consideration the most probable outcome.

Let $\vec{x}$ be a set of observations describing a single variable in dataset *X*. The probability of *x* (*p(x)*) being a value higher or lower than a given threshold can be calculated through an estimation of *X*'s Probability Density Function (PDF). For PDF, the frequency of data in the domain is first calculated, and smoothed using a cubic Hermite spline interpolation. The consequent interpolated frequencies thus define the density of information in the given domain. For the multivariate integration of these calculations, precise value of $p(x_i)\ \forall x \in \vec{x}$ can then be used to build a more complex equation using the chain rule from probability theory (eq. 1);

$$p\left(\bigcap_{i=1}^{n} x_i\right) = \prod_{i=1}^{n} p\left(x_i \middle| \bigcap_{j=1}^{n} x_j\right) \qquad \text{(eq. 1),}$$

The chain rule provides a means of computing the conditional probability of a multivariate event or vector, describing categorical or numeric variables respectively.

In the present methodology, when working with both qualitative and quantitative variables in the same dataset, we calculate multivariate conditional probability by using *P(X | A)*, where *P(A)* is typically used to describe the probability of an event (a categorical observation), and *P(X)* is the probability of a value (as described above).

Once multivariate probability distributions have been defined, the MC algorithm randomly samples points that belong to this distribution, evidently favoring the selection of observations that are more likely to occur.

### 2.2.2. Markov Chain Monte Carlo algorithms

MCMC algorithms are a family of statistic algorithms that are frequently employed as inference engines in probabilistic programming, advanced calculus applications, and data simulation purposes. MCMCs inherit from MC based methods by generating random values from a given probability distribution, while employing the use of Markovian Chains to stochastically explore these distributions via "random walks" (Gamerman & Lopes, 2006). The Markovian element of MCMC ensures that chain movement across dimensions is based solely on the chain's position at its present state and the probability of transition to the next. The result of these transitions is a convergence on the target probability distribution, favoring areas of high information density over unknown regions within the domain.

The present study employed the use of the Metropolis-Hastings 1st Order variant of MCMC (Metropolis, Rosebluth, Rosenbulth., Teller & Teller, 1953; Hastings, 1970), for the simulation of numeric data. The Metropolis-Hastings algorithm includes a defined acceptance criterion (eq. 2: $p_a$), that reduces correlation between successive states in the Markov chain (Metropolis et al., 1953; Hastings, 1970). To compute this, first the chain proposes a new state ($x_{k+1}$), by sampling a given step size (δ) from a given theoretical probability distribution (e.g., Gaussian). δ is then added to the present state, thus defining the proposed state ($x_{k+1} = x_k + \delta$). The probability of this new point is then calculated, and used to compute the criterion of acceptance (eq. 2);

$$p_a\left(x_{k+1}\middle|x_k\right) = \min\left(1, \frac{p(x_{k+1})q(x_k|x_{k+1})}{p(x_k)q(x_{k+1}|x_k)}\right) \qquad \text{(eq. 2).}$$

where $p(x)$ is the probability of event or value $x$, and $q(x)$ is the probability of the event or value not occurring. If $p(x_{k+1}|x_k)$ is larger than a random value sampled from a uniform [0, 1] distribution, then $x_{k+1}$ is accepted. If $x_{k+1}$ is rejected, then the chain remains in state $x_k$.

The step size (δ) used for MCMC is a debated issue (Graves, 2011), with many applications requiring the user to define a predetermined value of *δ*. Nevertheless, in the case of highly complex, inhomogeneous, and multidimensional feature spaces, no uniform value of δ can be easily defined. The present study robustly defined δ by using the scale of the distribution, as estimated using the Square Root of the Biweight Midvariance (√BWMV; Rodríguez-Martín, Rodríguez-Gonzálvez, Ruiz de Oña Crespo & González-Aguilera, 2019; Courtenay & González-Aguilera, 2020), which is more robust to limitations imposed by distribution normality (Höhle & Höhle, 2009). Nevertheless, future implementations of MCMCs should consider adaptive step sizes, with the possible integration of a decay parameter for the fine tuning of calculations.

MCMCs are typically run for 10,000 iterations, discarding a percentage of the first iterations as a “burn-in period”. For this study, 2,500 iterations were discarded. From the remaining values, 100 unique $\vec{x}$ values were extracted and appended to the original dataset.

### 2.3. Synthetic data evaluation

MC and MCMC algorithm’s performance was evaluated following the recommendations of Courtenay & González-Aguilera (2020). This approach evaluates augmented data from a purely statistical perspective. For numeric data, homogeneity was first evaluated through multiple Shapiro tests. Synthetic distributions were then compared with the original data to assess the magnitude of similarities according to Cohen’s δ. For this purpose, the Two One-Sided Test (TOST) of equivalence was used (Lakens, 2017). TOST evaluates the degree of overlapping between samples by assessing upper ($\epsilon S$) and lower ($\epsilon I$) equivalence bounds. For homogeneous distributions, parametric methods were used to calculate TOST according to Welch’s *t*-statistic (Schurimann, 1987). For inhomogeneous distributions, a trimmed ($t$ = 0.2) non-parametric approach was used according to Yuen’s robust *t*-statistic (Yuen & Dixon, 1973; Yuen, 1974). In either test, $H_0$ considers both samples to be different.

In the case of categorical variables, evaluations were performed testing the frequency of observations for equal proportions, according to Pearson’s $\chi^2$ test statistic.

So as to assess the effect of sample size on augmentation results, experiments were performed taking subsets of data, augmenting the subset, and comparing the augmented data with the original sample as a whole. For this experiment, the largest sample of Dataset 2 was taken (chimpanzee). For each iteration of the experiment, an individual was removed, creating gradually smaller subsets of the sample, and the remaining data was then augmented. TOST was then used to compare the augmented data and the original dataset. Correlations were then calculated between the corresponding test statistics and the sample size, followed by the use of the Chow test to estimate a structural break in patterns (Chow, 1960).

*p*-Values in this study were not evaluated using the traditional $p < 0.05$ as a threshold for defining statistical significance. Likewise, the term "significant" has been avoided. Instead, the present study uses, and recommends the use of, $p < 0.003$ ($3\sigma$) as an indicator of more conclusive results (Courtenay, Herranz-Rodrigo, González-Aguilera & Yravedra, 2021). This *p*-value was robustly defined considering the probability of Type I statistical errors at this threshold to be as low as 4.5 +/- [1.2, 15.9] %, using priors of 0.5 +/- [0.2, 0.8]. Likewise, where necessary ($p < 0.3681$, *sensu* Courtenay et al., 2021), statistical calculations were accompanied by the corresponding probability calculation of each observation being a Type I statistical error. This probability is known as the False Positive Risk (FPR, *sensu* Colquhoun, 2019).

### 2.4. Machine Teaching

As a means of evaluating the advantages of data augmentation, the present study additionally used augmented data for the construction of classification models. For this purpose, a Machine Teaching (MT) approach was adopted (Such, Rawal, Lehman, Stanley & Clune, 2019; Courtenay & González-Aguilera, 2020), to observe classification rates when separating between great apes based on femoral morphology. This entails the training of Machine Learning (ML) algorithms only on the synthetic data produced through data augmentation. Through MT, the real dataset is maintained separate throughout the training process. Once the ML algorithm is trained, it is then used to classify the original data, providing a direct means of evaluating model performance realistically. If models were to overfit on the synthetic data, then their classification performance on the real data would be poor. On the other hand, if algorithms perform well on the real data, then the algorithm can be considered a truly powerful classifier for real world applications.

The ML algorithm employed in the present study included Support Vector Machines (SVM; Cortes & Vapnik, 1995), trained similarly to the algorithms described in Courtenay et al. (2021). For SVMs, a radial basis kernel function was used, tuning both the $c$ and $\gamma$ hyperparameters using Bayesian Optimization Algorithms (BOAs) (Bergstra & Bengio, 2012; Snoek, Larochelle & Adams, 2012; Shahriari, Swersky, Wang, Adams & Freitas, 2016). BOA was initialized using the results from a random optimization algorithm as prior for hyperparameter selection (Bergstra & Bengio, 2012). Next an Expected Improvement (EI) Bayesian model was tuned for 50 iterations so as to define the final optimal $c$ and $\gamma$ values. SVMs were then trained using $k$-fold cross validation ($k = 10$), and a 70:30% *train:test* split.

For model evaluation, careful deliberation was taken in the selection of evaluation metrics. When considering the imbalance observed between each of the samples (≈ 6:1, in the case of chimpanzees and orangutans), the present study considered evaluation metrics less susceptible to changes in sample balance to be more reliable (He & Ma, 2013). These include; balanced accuracy, precision, recall, the F1-Measure, and the Kappa ($\kappa$) statistic where binary classification was performed. Alongside these metrics, model loss was calculated via the Root of the Mean Squared Error (RMSE), so as to calculate an estimation of overall model confidence when classifying new individuals.

So as to assess the power of MT techniques over traditional ML, all experiments were repeated excluding the data augmentation step in the process. For this purpose, each of the samples was divided into 70:30% *train:test* sets, and then subjected to the same processes as described above. As a means of testing the possible contamination that may be present product of performing the *train:test* split after performing PCA (*sensu* Calder et al., 2022), two experiments were devised. Experiment 1 performed *train:test* splits after PCA had been performed on the entire dataset, and was then subjected to MT.

Experiment 2, however, performed the *train:test* split on the Procrustes Superimposed coordinates. The train set was then used to calculate PC scores, which were then augmented and the synthetic values used to train a SVM. Next, the eigenvectors computed from the training data were used to predict the PC scores of the test set, and the trained SVM algorithm was used to predict the corresponding class labels.

Finally, and as a means of testing the advantages augmentation has over bootstrapping of samples, tests were performed training SVMs on bootstrapped datasets, and datasets augmented using MC based methods. To test the possibility of overfitting, the final evaluation of these tests was performed using adversarial examples, as described by Szegedy et al. (2014) and Goodfellow, Schlens & Szegedy (2015). For this purpose, a layer of Gaussian noise ($\epsilon = 0.2$) was added to landmark data prior to any of the analyses. The noisy data was then classified using the SVMs trained on bootstrapped data and MCMC augmented data. To simplify this experiment and see how different approaches affect both imbalance and small-sized samples, tests were performed to differentiate between male ($n = 22$), and female ($n = 8$) AMH femora.

### 2.5. 3D model simulation

For the purpose of simulating 3D models, the present study performed augmentation on geometric morphometric data in form feature space as well, i.e., excluding Procrustes scaling procedures during Procrustes superimposition. Augmented feature spaces were then reverse engineered to reconstruct landmark configurations using Least Square Regressions and Thin Plate Spline (TPS) interpolations (Bookstein, 1989). Once augmented landmark configurations had been calculated, these were used to warp surface meshes and produce a final simulation of a synthetic 3D model.

For mesh warpings, the median individual of each species was used (Rohlf, 1996, 1998). Due to intraspecific differences related to sexual dimorphism in modern great ape groups, calculations were performed using median male and median female meshes, if available.

### 2.6. The AugmentationMC library

In order to ensure reproducibility of the proposed augmentation approaches, all augmentation applications were programmed in the R programming language (v.4.0, with v.3 compatibility), and implemented as a library. The *AugmentationMC* library is available on GitHub (https://github.com/LACourtenay/AugmentationMC). This R library contains functions for both the use of MC and MCMC algorithms, alongside some additional functions for augmented data evaluation. The *AugmentationMC* library has only a single dependency (the *abind* library for more efficient tensor computing tasks), and was written using a mixture of an object-oriented and a functional programming paradigm. Object-oriented programming was performed using S4 type classes.

For the calibration of *p*-values and calculation of FPR values, formulae were implemented in the *pValueRobust* R library (https://github.com/LACourtenay/pValueRobust). A manual and description of the library has been included as Supplementary File 2.

### 2.7. Additional sample pre-processing

To deal with any possible issues that could be produced by sample imbalance (He & Ma, 2013), pre-processing of some samples was performed using an Adaptive Synthetic Sampling (ADASYN) algorithm.

ADASYN is a powerful adaptation of the Synthetic Minority Oversampling Technique (Chawla, Bowyer, Hall & Kegelmeyer, 2002), a popular algorithm used to augment minority samples in imbalanced datasets (He, Bai, Garcia & Li, 2008). ADASYN constructs functions on sample distributions, using calculations of data-density to generate new points that fill in empty regions of the target domain. This technique is powerful for the balancing of small datasets, however, should be used with caution. This is important considering how overuse of such algorithms may enhance a linear nature within the distribution that may not be a true representation of the original target domain (Courtenay & González-Aguilera, 2020). From this perspective, ADASYN was only used to balance the orangutan and gorilla samples from dataset 2 until they were approximately the same size as AMH and chimpanzees. Once balanced, all four samples were passed into MCMCs for more advanced numeric simulation modelling. Only data produced by MCMCs were included in the final study, while all ADASYN produced data were discarded. ADASYN was implemented in the Python (v.3.7.4) programming language.

## 3. Results

### 3.1. Monte Carlo based methods

MC based algorithms proved highly successful in simulating new data for each of the sample distributions, regardless of the data type (categorical or numeric). Beginning with the simplest and most common example of data augmentation, it can be seen how any type of Monte Carlo based algorithms (MC or MCMC) can effectively perform numeric simulations regardless of distribution normality.

For example, Ruff's (1995) $I_x$ variable presents an inhomogeneous distribution ($w$ = 0.99, $p$ = 0.003, FPR = 4.5%), yet numeric simulations of this univariate data produce highly realistic examples of $I_x$ values ($|d|$ = 0.009, $t$(103) = 12.8, $p$ = 1.2e-23, FPR = 1.8e-19%). When separating this variable by sex, normality is observed for both male ($w$ = 0.97, $p$ = 0.21, FPR = 47.1%) and female ($w$ = 0.96, $p$ = 0.10, FPR = 38.5%) femora. In each of these cases, simulations produced almost exact replicas of sample distributions (Fig. 2), with highly realistic examples of $I_x$ values for both males ($|d|$ = 0.040, $t$(108) = -1.0, $p$ = 4.4e-45, FPR = 1.2e-40%) and females ($|d|$ = 0.003, $t$(104) = 0.07, $p$ = 1.2e-44, FPR = 3.2e-40%). Moreover, in all of these examples, exact replicas of the original data are non-existent.

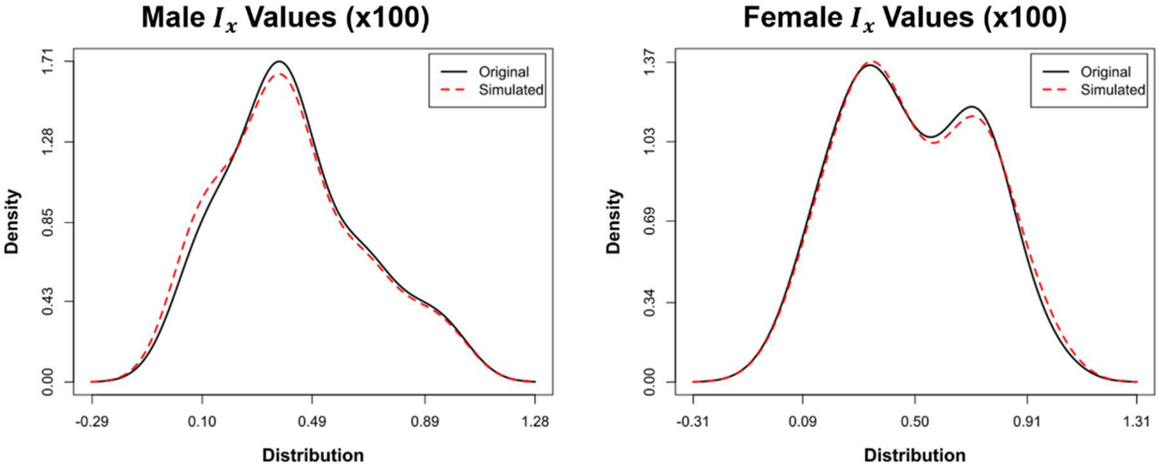


***Figure 2*** *– Example of original and augmented numeric distributions of the $I_x$ variable from Ruff's (1995) dataset, augmented using Monte Carlo based algorithms.*

Similarly, when considering MC algorithm performance for modelling categorical variables, algorithms are seen to successfully produce 1,000 examples of different sets of qualitative observations (Fig. 3), while maintaining the original proportions of observations (Table 1), such that the augmented and original data are almost statistically identical ($p \approx 1$).

***Table 1** – Statistical test results of equal proportion according to Pearson's $\chi^2$ test between different qualitative variables from Ruff's (1995) dataset, when simulating 1,000 examples using MC based algorithms. p Cal. refers to the calibration of this p-Value to a Probability of Null Hypothesis value, ergo, the probability that the analyst is wrong if concluding the alternative hypothesis (sample proportions are different).*

| | Original % | Augmented % | $\chi^2$ | *df* | *p* | *p* Cal. |
|---|---|---|---|---|---|---|
| Eastern Africa | 39.4 | 39.3 | 5.1e-31 | 1 | >0.99 | >97.3 |
| North America | 60.6 | 60.7 | 5.1e-30 | 1 | >0.99 | >97.3 |
| Male | 51.5 | 53.2 | 0.046 | 1 | 0.83 | 70.4 |
| Female | 48.5 | 46.8 | 0.046 | 1 | 0.83 | 70.4 |

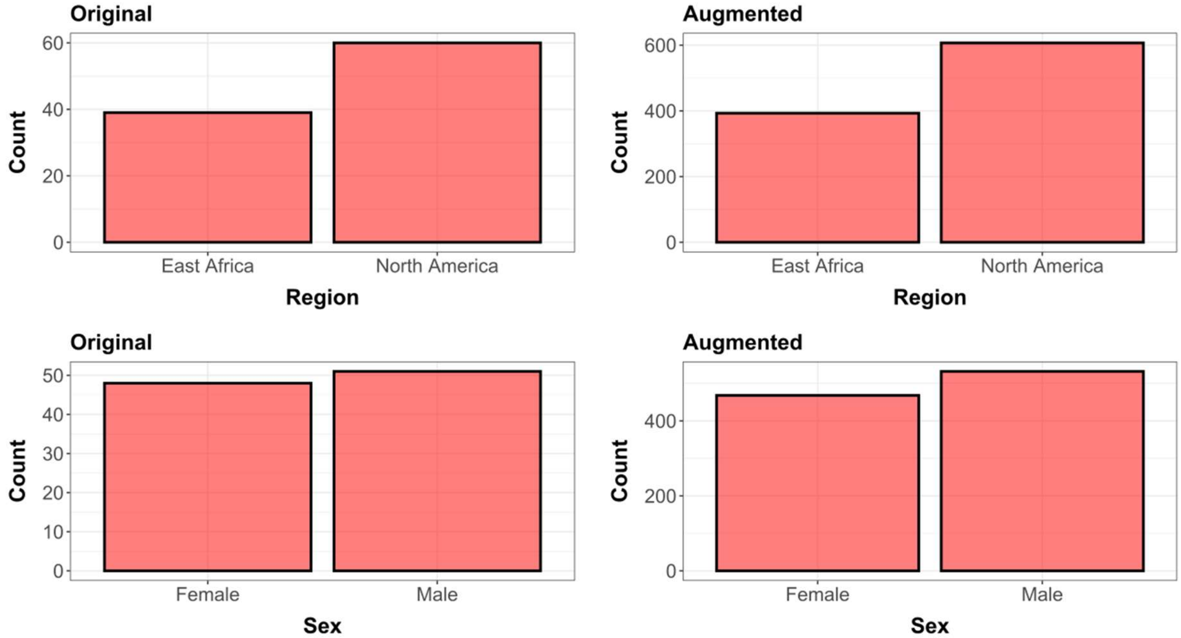


***Figure 3** – Barplots describing the proportion of different categorical variables from Ruff's (1995) dataset before (original), and after (augmented), simulating 1,000 new values via MC based algorithms.*

When confronting a multimodal problem, such as the mixture of both categorical and numeric variables, MC algorithms can be used in a number of ways. One approach considers the conditional augmentation of variables (e.g., Fig. 4), where a qualitative factor (such as sex), can be used to condition the simulation of numeric variables. In all cases, the distribution of both male and female $I_x$ and $I_y$ values appear to be almost identical to the original distributions, while increasing sample variance. Statistically these synthetic distributions are highly equivalent to the original distributions they were modelled from

(Table 2). If this component was to be developed further by augmenting a multitude of both qualitative and quantitative variables, MC performance would continue to prove highly efficient at replicating the original distributions (Fig. 5), with high rates of equivalency (Table 3).

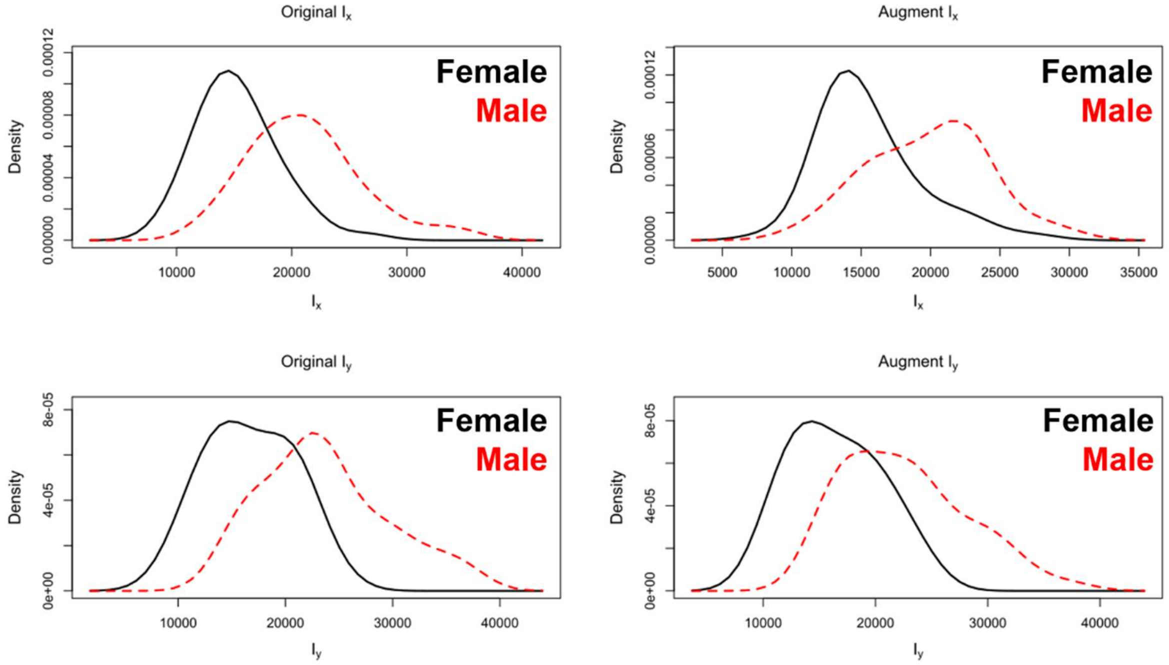


***Figure 4*** *– Density plots comparing the distribution of numeric values from Ruff's (1995) dataset before (Original) and after (Augmented) simulating 100 new values via MC based algorithms, conditioned by a single categorical variable (sex).*

***Table 2*** *– TOST results when augmenting numeric distributions ($I_x$ & $I_y$) from Ruff's (1995) dataset, when conditioned by a qualitative variable (sex), producing 100 synthetic observations using MC based approaches. |d| is the normalised absolute difference. t is the test-statistic. df are degrees of freedom. FPR is the False Positive Risk. Samples marked with an * were tested using TOST, while those without were tested using rTOST.*

| | *\|d\|* | *t* | *df* | *p*-Value | *FPR (%)* |
|---|---|---|---|---|---|
| $I_x$ Male* | 0.0116 | -0.2959 | 97.98 | 6.49E-45 | 1.79E-40 |
| $I_x$ Female | <0.0001 | 8.5627 | 59.93 | 2.73E-12 | 1.97E-08 |
| $I_y$ Male | 0.015 | 5.324 | 59.97 | 7.99E-07 | 0.003 |
| $I_y$ Female | 0.0032 | 5.1289 | 60 | 1.65E-06 | 0.006 |

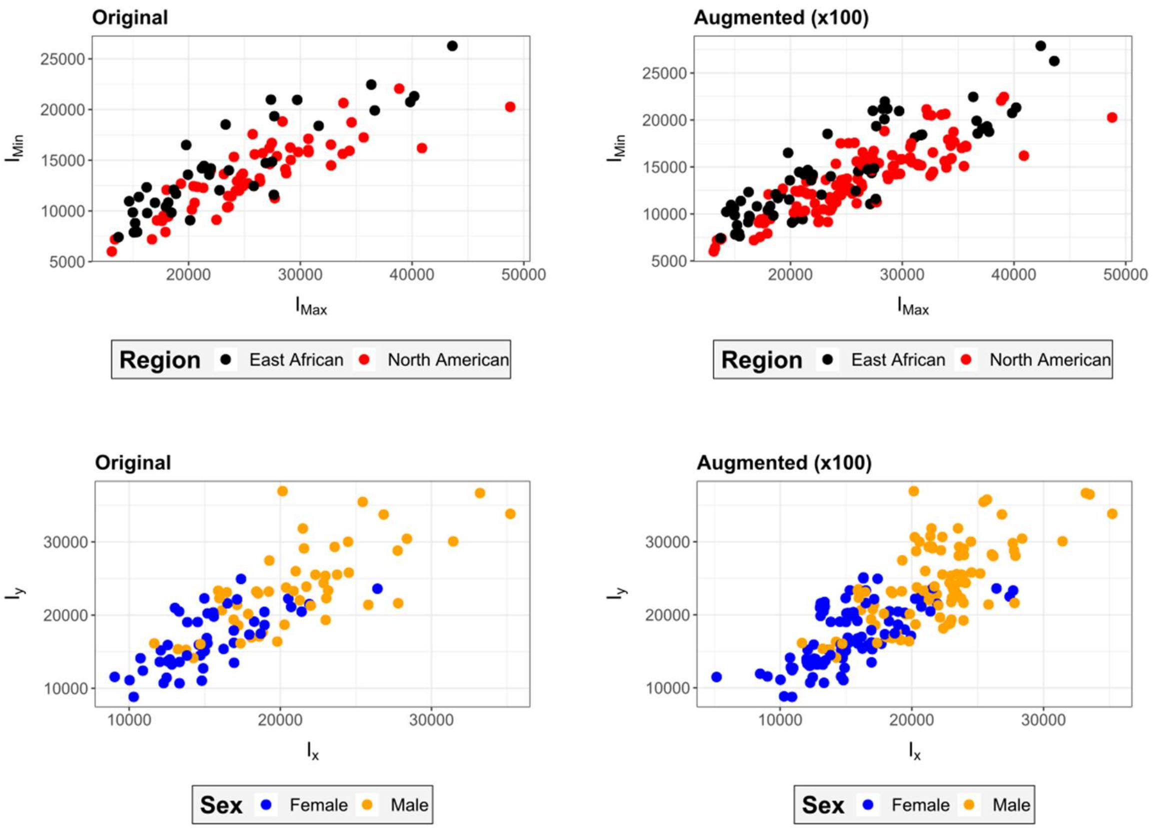


***Figure 5*** *– Examples of the original (left) and augmented (right) scatter plots of Ruff's (1995) dataset, augmented using MC based algorithms of both categorical and numeric data.*

***Table 3*** *– TOST results when augmenting the entire Ruff's (1995) dataset, including both numeric and categorical variables. 100 synthetic observations were produced using MC based approaches. |d| is the normalized absolute difference. t is the test-statistic. df are degrees of freedom. FPR is the False Positive Risk. Samples marked with an * were tested using TOST, while those without were tested using rTOST.*

| | *\|d\|* | *t* | *df* | *p* | *FPR (%)* |
|---|---|---|---|---|---|
| $I_{max}$ | 0.016 | 13.5 | 72.3 | 1.0E-21 | 1.3E-17 |
| $I_{min}$* | 0.002 | 0.08 | 117.6 | 3.5E-84 | 1.8E-79 |
| $I_x$ | 0.008 | 15.8 | 73.5 | 1.3E-25 | 2.0E-21 |
| $I_y$ | 0.012 | 14.0 | 73.8 | 1.0E-22 | 1.4E-18 |
| $J$* | 0.017 | -0.74 | 113.7 | 2.1E-73 | 9.7E-69 |

### 3.2. Markov Chain Monte Carlo

MCMC algorithms proved to be highly efficient for the augmentation of numeric data. When used to augment geometric morphometric data, MCMCs simulated extremely realistic multidimensional distributions in a $\mathbb{R}^{12}$ feature space (Table 4, Fig. 6). The poorest performance was observed when trying to augment the gorilla and orangutan samples. This is logical considering their much smaller sample size, which may not capture the true variability of the population. Nevertheless, sample distributions on all accounts have been tested to be highly equivalent to the original data, maintaining overall descriptive features such as central tendency yet increasing sample variance. This proves MCMCs to be efficient when modelling multivariate data, while maintaining the degree of separation between samples as realistically as possible.

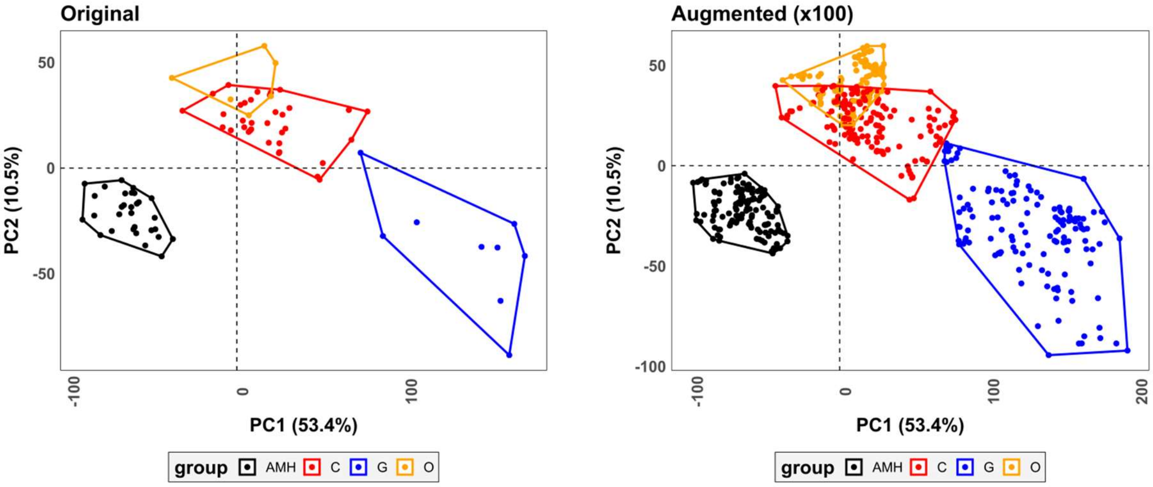


***Figure 6*** *– PCA scatter plots in shape space of great ape femora. Left panel: the original PCA. Right panel: the original PCA alongside 100 simulated individuals using MCMC. AMH = Anatomically Modern Humans, C = Chimpanzees, G = Gorillas, O = Orangutans.*

***Table 4*** *– TOST results when augmenting geometric morphometric data in shape space (Fig. 6). 100 synthetic observations were produced using MCMC. PC scores are displayed alongside the percentage of explained variance. Samples include Anatomically Modern Humans (AMH), Chimpanzees (C), Orangutans (O) and Gorillas (G). |d| is the normalized absolute difference. t is the test-statistic. df are degrees of freedom. FPR is the False Positive Risk. All samples were tested using rTOST.*

| | | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|---|
| | | 53.4% | 10.5% | 6.9% | 4.5% | 3.9% |
| *AMH* | \|*d*\| | 0.004 | 0.000 | 0.008 | 0.002 | 0.001 |
| | *t* | 8.30 | 8.05 | 9.14 | 11.93 | 12.34 |
| | *df* | 140.2 | 140.0 | 142.1 | 142.2 | 138.9 |
| | *p* | 3.9E-14 | 1.6E-13 | 3.0E-16 | 1.8e-23 | 2.2e-24 |
| | FPR (%) | 3.3E-10 | 1.3E-09 | 2.9E-12 | 2.5e-19 | 3.2e-20 |
| *C* | \|*d*\| | 0.005 | 0.011 | 0.008 | 0.001 | 0.003 |
| | *t* | 15.10 | 8.34 | 11.38 | 12.65 | 12.09 |
| | *df* | 133.8 | 131.7 | 131.9 | 135.9 | 132.4 |
| | *p* | 5.5E-31 | 4.3E-14 | 1.3E-21 | 5.1e-25 | 1.9e-23 |
| | FPR (%) | 1.0E-26 | 3.6E-10 | 1.6E-17 | 7.7e-21 | 2.7e-19 |
| *O* | \|*d*\| | 0.000 | 0.001 | 0.001 | 0.000 | 0.001 |
| | *t* | 9.73 | 7.90 | 6.09 | 15.38 | 8.62 |
| | *df* | 129.7 | 129.6 | 129.9 | 129.4 | 130.0 |
| | *p* | 2.0E-17 | 5.1E-13 | 5.9E-09 | 2.7e-31 | 9.7e-15 |
| | FPR (%) | 2.1E-13 | 3.9E-09 | 3.0E-05 | 5.2e-27 | 8.5e-11 |
| *G* | \|*d*\| | 0.002 | 0.003 | 0.005 | 0.001 | 0.001 |
| | *t* | 10.38 | 12.61 | 8.30 | 12.29 | 12.66 |
| | *df* | 137.0 | 136.4 | 136.6 | 136.3 | 136.3 |
| | *p* | 2.7E-19 | 5.8E-25 | 4.4E-14 | 3.9e-24 | 4.4e-25 |
| | FPR (%) | 3.2E-15 | 8.8E-21 | 3.7E-10 | 5.7e-20 | 6.7e-21 |

### 3.3. Machine Teaching

Considering the already existing high degree of separation between great ape groups, most discriminant functions are already capable of separating between the different genera. For instance, a standard cross-validated Linear Discriminant Analysis (LDA) typically used in Geometric Morphometrics on PC scores obtained after PCA dimensionality reduction already reaches 98.7% accuracy when separating between groups, with only a single gorilla individual being assigned to the chimpanzee class. When using more sophisticated algorithms, such as SVM, accuracy drops drastically (Table 5), due to the very small amount of data available which is typically needed for ML applications, while error rates are unacceptably high (RMSE loss = 0.724). MT approaches, however, reach 100% classification with extremely low loss values (RMSE = 0.033).

When performing the second experiment (Exp. 2, Table 5), which included the separation of training data and testing data prior to the calculation of PCA, algorithms maintain high classification rates, with even lower loss values (RMSE = 0.017). This proves that even when trying to reduce the possible contamination of the testing set related to PCA preprocessing as described by Calder et al.

(2022), MT approaches presented here are able to produce highly powerful classification algorithms that may be more realistically applicable to the fossil record.

***Table 5*** *– Results of SVMs when used to differentiate between great ape samples based on geometric morphometric data. SVMs were trained either on the raw data, or datasets augmented using MCMC approaches. RMSE = Root Mean Square Error. Exp 1 = Experiment 1, augmenting PC scores after the train/test split. Experiment 2 = augmenting PC scores before the train/test split.*

| | Raw | Augmented Exp. 1 | Augmented Exp. 2 |
|---|---|---|---|
| Accuracy | 0.55 | 1.00 | 1.00 |
| Precision | 0.38 | 1.00 | 1.00 |
| Recall | 0.31 | 1.00 | 1.00 |
| F1-Score | 0.26 | 1.00 | 1.00 |
| RMSE | 0.724 | 0.033 | 0.017 |

### 3.4. MCMC vs bootstrap

Experiments performed to evaluate the difference between bootstrap methods as opposed to MCMC augmentation reveal that both approaches are able to reach high accuracy values (Table 6). Nevertheless, SVMs suffer more from dataset imbalance as seen in a drop in F1-scores, as well as a considerable difference in the loss values when used to differentiate between male and female femora based on geometric morphometric data. This is logical considering how bootstrap techniques only resample and reshuffle the data that is already present, while MCMCs generate a completely new example of what a male or female femur may look like. Likewise, while accuracy is high for bootstrap models, the authors still feel that this is an inflated observation which, in more complex cases, would be a perfect example of model overfitting.

***Table 6*** *– SVM evaluation results on noisy data when differentiating between male and female AMH femora using bootstrap and MCMC based augmentation techniques. RMSE = Root Mean Square Error.*

| | Bootstrap | MCMC |
|---|---|---|
| Accuracy | 0.90 | 0.90 |
| Precision | 0.86 | 0.78 |
| Recall | 0.75 | 0.88 |
| F1-Score | 0.80 | 0.82 |
| Kappa (κ) | 0.73 | 0.75 |
| RMSE Male | 0.248 | 0.302 |
| RMSE Female | 0.401 | 0.349 |

### 3.5. Data augmentation for the simulation of 3D models

Considering the weight the variable size has on Geometric Morphometric form feature spaces, 13 variables were included for 3D model simulation. In this particular case, 90% of shape variation is described by 12 PC scores, and the first PC score of form space strongly represents the variable size. Once these variables had been defined, MCMCs were found to efficiently augment each of the $\mathbb{R}^{13}$ distributions (Table 7, Fig. 7), similar to the observations made in section 3.2.

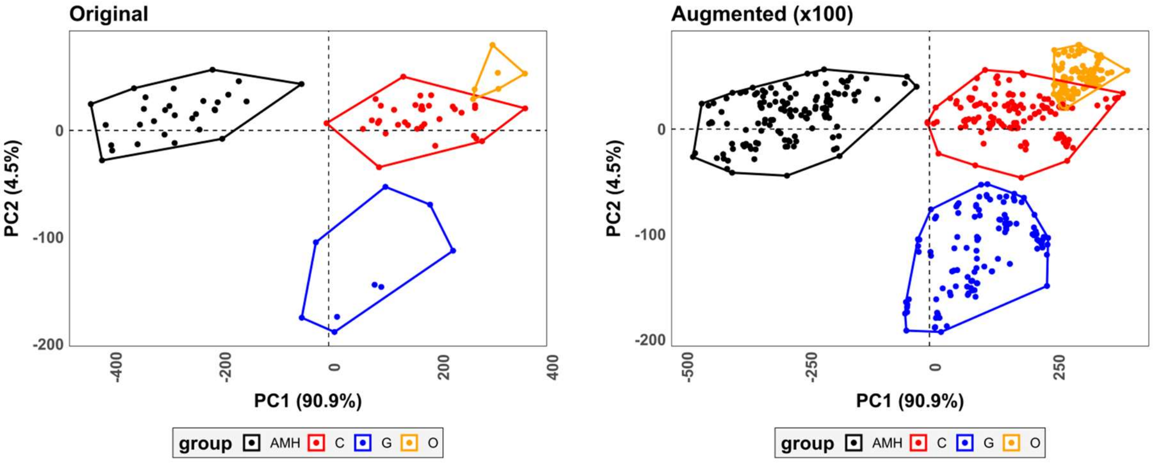


***Figure** 7 – PCA scatter plots in form space of great primate femora. Left panel: the original PCA. Right panel: the original PCA alongside 100 simulated individuals using Monte Carlo Markov Chains. AMH = Anatomically Modern Humans, C = Chimpanzees, G = Gorillas, O = Orangutans.*

***Table** 7 – TOST results when augmenting geometric morphometric data in form space (Fig. 7). 100 synthetic observations were produced using MCMC. PC scores are displayed alongside the percentage of explained variance. Samples include Anatomically Modern Humans (AMH), Chimpanzees (C), Orangutans (O) and Gorillas (G). |d| is the normalized absolute difference. t is the test-statistic. df are degrees of freedom. FPR is the False Positive Risk. All samples were tested using rTOST.*

| | | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|---|
| | | 90.9 | 4.5 | 0.7 | 0.5 | 0.4 |
| *AMH* | \|*d*\| | 0.002 | 0.003 | 0 | 0.007 | 0.004 |
| | *t* | 9.94 | 7.92 | 9.55 | 14.26 | 9.26 |
| | *df* | 129 | 127.4 | 129.9 | 125.9 | 127.9 |
| | *p* | 6.1E-18 | 5.0E-13 | 5.3E-17 | 2.4E-28 | 3.1E-16 |
| | FPR (%) | 6.6E-14 | 3.8E-09 | 5.4E-13 | 4.1E-24 | 3.0E-12 |
| *C* | \|*d*\| | 0.011 | 0.009 | 0.01 | 0.001 | 0.007 |
| | *t* | 8.48 | 13.44 | 11.11 | 10.08 | 13.07 |
| | *df* | 126.3 | 122.2 | 126.5 | 124.1 | 125.2 |
| | *p* | 2.6E-14 | 3.8E-26 | 9.7E-21 | 3.9E-18 | 1.9E-25 |
| | FPR (%) | 2.2E-10 | 6.1E-22 | 1.2E-16 | 4.3E-14 | 3.0E-21 |
| *O* | \|*d*\| | 0.002 | 0.001 | 0 | 0.001 | 0 |
| | *t* | 8.36 | 9.14 | 8.21 | 5.79 | 17.06 |
| | *df* | 127.8 | 127.8 | 128 | 127.6 | 127.7 |
| | *p* | 4.6E-14 | 6.2E-16 | 1.0E-13 | 2.6E-08 | 5.0E-35 |
| | FPR (%) | 3.8E-10 | 5.9E-12 | 8.3E-10 | 1.0E-04 | 1.1E-30 |
| *G* | \|*d*\| | 0.009 | 0.011 | 0.001 | 0.002 | 0.001 |
| | *t* | 6.96 | 6.85 | 2.72 | 8.46 | 8.72 |
| | *df* | 132.9 | 132.7 | 132.6 | 132.6 | 132.5 |
| | *p* | 7.2E-11 | 1.2E-10 | 1.6E-11 | 2.2E-14 | 5.1E-15 |
| | FPR (%) | 4.6E-07 | 7.7E-07 | 1.1E-07 | 1.9E-10 | 4.5E-11 |

Once each of the augmented feature spaces had been reverse engineered to produce a set of synthetic landmark coordinates, mesh warping's and the resultant simulated 3D models were found to appear highly realistic. Detailed inspection of these models found simulations to effectively capture the natural variability of femoral morphology as much as possible based on the samples provided (Fig. 8 & 9). While it is true that the displacements of some fixed landmarks along the femoral shaft cause a slight distortion to small localized areas of the 3D models (e.g., LM34, LM37 and LM40, see Supplementary File 1, Table S2), these changes are negligible, while the majority of general diaphyseal morphological feature traits are still captured by the semilandmark mesh. In light of this, however, it must be stated that the quality of 3D model simulations will be entirely dependent on the number and quality of landmarks included in the model. Meshes cannot be warped if information is lacking on how morphological changes occur in this area. Likewise, the choice of a female or male individual as the reference mesh will also condition results. In the present study, analyses were performed warping meshes to both the male and female reference models, and then inspected to see which models present the least amount of non-biological distortion. Future research, however, should include a much larger sample of both female and male individuals for all species so as to obtain optimal results.

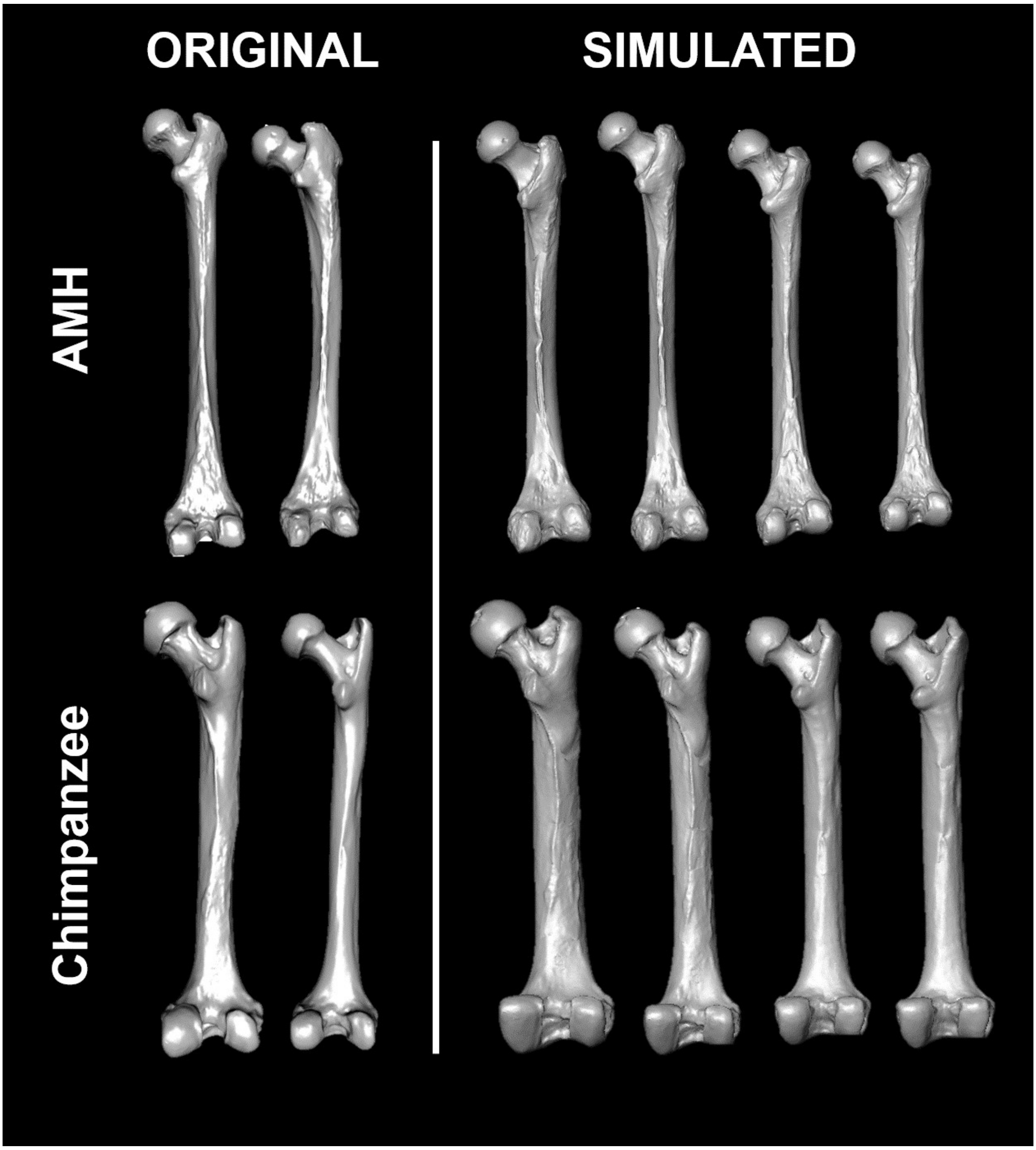


***Figure 8** – Examples of real 3D models of Anatomically Modern Human (AMH) and chimpanzee femora, alongside 4 simulations of synthetic 3D models. Of the original models, the left femora belong to male individuals, while the right femora belong to females. The first two simulated 3D models are based on the model of male femora, while the final two simulated 3D models are based on female femora.*

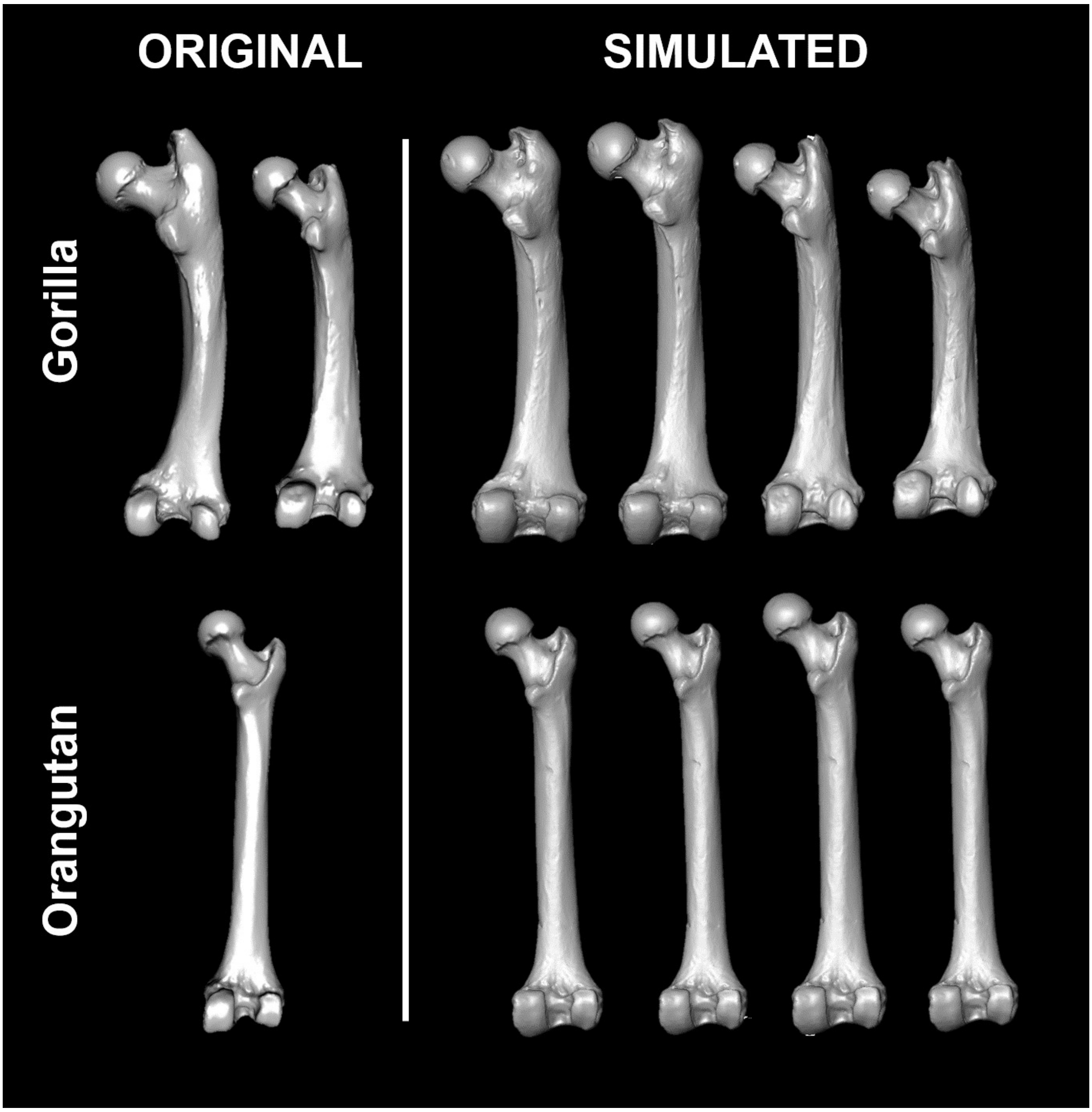


***Figure 9*** *– Examples of real 3D models of gorilla and orangutan femora, alongside 4 simulations of synthetic 3D models. Of the original models, and where available, the left femora belong to male individuals, while the right femora belong to females. The first two simulated 3D models are based on the model of male femora, while the final two simulated 3D models are based on female femora. In the case of orangutans, all femora belong to female individuals.*

### 3.6. Minimum sample sizes and general recommendations

Performing experiments on subsets of samples reveal a negative correlation between sample size and the variation presented within the augmented sample (Pearson's ρ = 0.66), while a positive correlation exists between sample size and TOST's |*d*| (ρ = 0.63). This implies that as sample size decreases, the distribution of simulated data begins to present a higher variance than the original data, and thus begins to lose equivalency. Nevertheless, this observation at present can only be considered of notable importance in the case of |d| values (t = -4.39, *p* = 0.0001, FPR = 0.34%), as the negative correlation between sample size and sample variation was revealed to be inconclusive based on the present data (t = -0.45, $p$ = 0.66). Chow's test revealed that correlation coefficients are continuous (F = 1.56, $p$ = 0.23), indicating no particular decrease in sample size to produce a notable change in the magnitude of equivalency.

While it is logical to assume that the size of the dataset used for augmentation will affect the quality of the augmented data, our intuition when observing these results finds augmenting sample sizes of $n < 10$ to present important differences that no longer represent the variability of the original data. Likewise, samples of $n < 15$ also run a risk of losing reliability. From this perspective, we recommend that either (1) a sample size of $n > 10$ be used for augmentation, (2) if sample sizes are 15 < n > 10 then extreme care be taken, and/or (3) where possible, algorithms such as ADASYN be used to preprocess data that fall within this range.

## 4. Discussion

Sample size is a fundamental component of all elements of science. Both statistical tests and artificially intelligent algorithms are dependent on the amount of data available, which is often hindered by the quality of the fossil record. Multiple techniques exist to overcome some of these limitations. Nevertheless, not all may be suitable for particular types of analyses. Here we have shown how MC based algorithms are highly efficient at simulating probability distributions of categorical, numeric, and multimodal datasets. On all accounts, statistical tests show how the simulated data is (1) highly equivalent to the original data, but (2) not identical. Similar to the observations made by Courtenay & González-Aguilera (2020), augmentation procedures thus increase the represented variability without significantly shifting the distribution's central tendency, while maintaining the general properties of the distribution. This can have great advantages in improving classification algorithm performance, as well as providing more robust measures for statistical applications.

### 4.1. Recruiting a skeleton crew: How can we enhance the fossil record?

In paleoanthropological research, analysts are inevitably restricted to make predictions about the population of a species based usually on very few remains due to the scarcity and fragmentary nature of the fossil record. Difficulties in analyzing large sets of specimens increase due to the common reluctance in data sharing, especially in the case of the very much appreciated fossil hominins, that are usually underrepresented in most studies in comparison to modern data (e.g., Lague, 2002; Harmon, 2009; Tallman, 2012, 2014). The estimation of population parameters such as central tendency and deviation, for example, are the most easily affected by incomplete or imbalanced datasets. Likewise, more sophisticated analyses, such as CVA, bgPCA, or tests commonly employed to address 'modularity *vs* integration' or phylogenetic questions, are highly dependent on the data used as input (Albrecht, 1992; Mitteroecker & Gunz, 2009; Mitteroecker & Bookstein, 2011, Bookstein, 2017, 2019; Cardini et al., 2019), which will essentially condition the type of conclusions drawn from the results.

Fossil sample size is also very much affected by the difficulty in finding either homologous specimens, or specimens that preserve conspicuous portions for morphological, or any other type, of quantitative analysis that might require the integrity of specific bone areas. Two procedures are available in these cases; researchers might opt for restricting the area of study and sample size to incorporate only original data in their analysis (Trinkaus, Ruff, Churchill & Vandermeersch, 1998; Ruff, McHenry & Thackeray, 1999; Tallman, 2012, 2014; Tallman, Almécija, Reber, Alba & Moyà-Solà, 2013; Almécija et al., 2013, 2019; Lague, 2015; Beaudet, Heaton, L'Abbé, Pickering & Stratford, 2018; Daver et al., 2018; Lague et al, 2019a,b), or they can reconstruct the missing or deformed bone portions through (TPS) estimations and/or mirroring and merging tools in case of symmetric structures (Gunz & Harvati, 2007; Friedline, Gunz, Janković, Harvati & Hublin, 2012; Gunz, Mitteroecker, Neubauer, Weber & Bookstein, 2009; Gunz et al., 2020; Claxton, Hammond, Romano & Oleinik, DeSilva, 2016; Mori, Profico, Reyes-Centeno & Harvati, 2020; Davis, Profico & Kappelman, 2021; Beaudet et al., 2022), before data collection for statistical analysis and comparison.

Hominin or other fossil primate remains are not only commonly incomplete but they also often appear in isolation (e.g., Leakey, 1971; Di Vincenzo et al, 2015), or in the form of very partial skeletons (e.g., Wood, 1974; Susman, 1989; Clarke, 1998; Susman, Ruiter & Brain, 2001; Domínguez-Rodrigo et al., 2013; Lague et al, 2019a), providing thus very limited windows to the study of our ancestors. The finding of complete or moderately complete skeletons is very rare (e.g. *Lucy*: Johanson & Edey, 1981; the *Turkana Boy*: Brown, Harris, Leakey & Walker, 1985; DIK-1-1: Alemseged et al., 2006; *Mtoto*: Martinón-Torres et al, 2021), but even more so is the discovery of fossil hominin groups or populations in a single spot (e.g., the First Family: Johanson, Taieb & Coppens, 1982; Sima de los Huesos: Arsuaga, Martínez, Gracia, Carretero & Carbonell, 1993; Dinaledi and Lesedi chambers: Berger et al., 2015, Hawks et al., 2017; Dmanisi, Lordkipanidze et al, 2013; Kanapoi: Ward, Plavcan & Manthi, 2020; Spy Cave: Crevecoeur et al, 2010; Malapa: Berger et al, 2010). The difficulties in finding a group of individuals, or even complete or well-preserved specimens, increases as paleoanthropologists go back in geological time. In those cases, where two or more individuals of the same species have been found, certain parameters related to intraspecific variation regarding sexual dimorphism, or ontogenetic patterns, might be easier to detect. Nevertheless, even in the cases where the number of individuals is relatively high, fully understanding intraspecific variability of long-lasting fossil groups results difficult based on very specific populations.

Additionally, most statistical studies require larger sample sizes than those encountered in the most abundant examples (estimated number of individuals in the Rising Star Cave = 18: Berger et al. (2015), Hawks et al. (2017); estimated number of individuals in Sima de los Huesos = 29: Bermúdez de Castro, Martínez, Gracia-Téllez, Martinón-Torres & Arsuaga, 2020).

Datasets such as those from the Sima de los Huesos or Rising Star might thus provide a more anatomically and morphologically restricted source to generate synthetic data. Nevertheless, they could be more reliable in certain cases as well (though not always, see for instance; Grine, 2019), as they offer a sample of spatially and chronologically well-defined individuals, in contrast to the combination of individuals from different areas, and geological strata, whose taxonomic allocation might still be uncertain. Likewise, it is also logical to assume that it is not the same to estimate data of a specific taxon based on 10 than on 100 individuals, as also highlighted by the experiments performed on sample size. Paleoanthropologists must thus consider all these questions before conducting data augmentation so as to be aware of the possible shortcomings of their simulated group. Additionally, while MC approaches are able to provide a more robust means of making statistical inferences, analysts must remember that augmented data cannot simulate variables or variability that are completely unknown to the analyst and should never be used to replace real datasets.

The present study has taken two datasets describing multiple features of primate femoral attributes, ranging from external morphology, to biomechanical cross-sectional parameters and qualitative data. Using these datasets, results prove the validity of MC based algorithms to generate new synthetic data (minimum of 100 new individuals for each of the case studies), without altering the nature of the probability distribution. Results also show how these methods might be valid for paleoanthropological studies as they perform well in multimodal problems, even those that are affected by inhomogeneous distribution patterns or sample size imbalance. Very small groups, such as those usually encountered in the fossil record, might however present more difficulties, as highlighted by the orangutan ($n = 5$) and gorilla ($n = 9$) samples, and require preprocessing steps before data augmentation.

Augmented samples should thus ideally include more than 10 (or even better, 15) individuals so as to perform more reliable classification analyses.

Though not ideal, even in the cases where sample size is very limited, 3D reproductions of the synthetic data based on reverse engineering (and with the help of preprocessing steps such as ADASYN) indicate that MC calculations provided very realistic results, opening the doors to further analytical possibilities. Results thus suggest that the application of MC based algorithms, though not fully solve the problem of scarcity and fragility of the fossil record, provide important advantages for paleoanthropological data management and statistical robustness.

### 4.2. Methodological reflections on data augmentation for classification tasks

The majority of research into data augmentation is primarily driven by the rising interest in Computer Vision (CV) applications in Artificial Intelligence (AI). In CV, images are "augmented" through a series of transformations; including the scaling, rotation, flipping, cropping and alteration of an image brightness (Goodfellow, Bengio & Courville, 2016). In these approaches, algorithms simply shift and distort the matrix, trying to capture possible changes that could be produced by changes in angle or perspective of the camera or sensor. Nevertheless, this does not truly simulate the variability an object may have, as they are not conditioned by anything more than the original image.

In the context of simulating 3D models, some promising work has been done using Generative Adversarial Network (GAN) based architectures (Wu, Zhang, Xue, Freeman & Tenenbaum, 2017; Achlioptas, Diamanti, Mitliagkas & Guibas, 2018), with exciting contributions from the field of geometric learning (Shu, Park & Kwon, 2019; Valsesia, Magli & Fracastoro, 2019). Nevertheless, in the majority of these cases algorithms were implemented using datasets of ≈ 4,000 (ModelNet dataset from Wu et al., 2015) or ≈ 51,000 (ShapeNet dataset from Chang et al., 2015) point clouds. The present study uses a dataset 50 to 600 times smaller, while still efficiently simulating a set of 400 3D models. This is possible through the augmentation of geometric morphometric shape variables, as opposed to directly augmenting coordinate point clouds. A geometric morphometric based approach thus uses statistical calculations of the most probable example of an individual's femur, and exploits already existent geometric information to simulate what a new individual may look like. This has numerous advantages for the possible integration of Geometric Learning applications into Geometric Morphometrics (Bronstein et al., 2017; Kipf & Welling, 2017; Qi, Yi, Su & Guibas, 2017; Wang et al., 2019).

In other lines of research, the augmentation of datasets using GANs has become very popular, and has been known to be useful for building much more robust classification models (Such et al., 2019; Courtenay & González-Aguilera, 2020; Courtenay et al., 2021). For the simulation of numeric variables, Courtenay & González-Aguilera (2020) presented the use of GANs for the augmentation of three different geometric morphometric datasets with statistically powerful results ($p < 0.001$, FPR = 1.8%). In an applied case study, Courtenay et al. (2021) found these algorithms to work well ($p = 2.4\text{e-}13$, FPR = 1.9e-09%), however were vulnerable to the presence of outliers and still faced some problems if the original training datasets are small. In this same study, these authors presented a comparison with MCMCs for the augmentation of geometric morphometric data with much greater success ($p = 1.2\text{e-}57$, FPR = 4.2e-53%).

At present, both GANs and MCMCs can be considered very powerful algorithms for the modelling of data. The present study displays evidence that MCMCs may be more powerful. Firstly, MC based approaches are less computationally expensive than GANs (Courtenay et al., 2021). Secondly, MC algorithms directly model from the dataset's probability distribution, which implies that outliers are given less weight in the final modelling of the predicted population distribution. Needless to say, and especially when comparing MC based algorithms and MCMCs, MCMCs can prove difficult to implement and train in many cases (Gelman, Roberts & Gilks, 1995; Graves, 2011). This is mostly evident when modelling from highly complex distributions in large dimensional feature spaces (e.g., Fig. 6 and 7). When modelling from complex sets of PC scores, the first set of PC scores are easy to simulate as they contain the majority of distribution variation information. Nevertheless, as we increase the number of PC scores included in a study, the number of residual data increases, which consequently hinders MCMC performance. In some preliminary cases from the present study, the inclusion of all PC scores produced similar results to the "slight mode collapse" observed by Courtenay & González-Aguilera (2021: Fig. 3). In other cases, MCMCs were unable to reach convergence.

Some of the first introductions of AI algorithms into archaeology have been mostly accompanied by the use of bootstrapping procedures to "increase" small sample sizes (Domínguez-Rodrigo, 2018; Courtenay et al., 2019, Courtenay, Huguet, González-Aguilera & Yravedra, 2020, Moclán, Domínguez-Rodrigo & Yravedra, 2019; *inter alia*). While bootstrapping is a valid and useful technique for statistical inference, it has received some criticism in the context of training classification algorithms (Courtenay & González-Aguilera, 2020; McPherron, Archer, Otárola-Castillo, Torquato & Keevil, 2022). Bootstrapping is a resampling technique, therefore is unable to generate new observations. From this perspective, the misconception that x1,000 bootstrapped sample is larger than the original dataset is misinformed, as no new information is available (i.e., the dataset essentially remains the same size, just reshuffled). The objective of classification algorithms is to generate mathematical functions that are generalizable to unknown paleoanthropological cases, however, the bootstrap approach does not complete areas of poor information density (Courtenay & González-Aguilera, 2020). When algorithms are then exposed to new observations, they are likely to perform poorer predictions, or unrealistically high accuracies.

In the present study, we have shown how bootstrapping fails to overcome issues presented by sample imbalance. This can be seen in how the confidence of SVMs drops significantly when predicting the class of adversarial examples of female individuals. In a study by Moclán et al. (2019), experiments performed with and without bootstrapping showed how even without bootstrapping, separation among samples is very high. While bootstrapped samples produce even higher accuracy rates, we strongly believe that this is due to an overrepresentation of resampled (i.e., repeated) case studies, while a MC based approach, as presented here, would produce more favorable results. Finally, although the present authors are also guilty of having presented bootstrapped results (Courtenay et al., 2019), we have since shown how even when using MCMCs and GANs for data augmentation, the construction of highly powerful classification algorithms is still possible for the same type of data (Courtenay et al., 2021).

## 5. Conclusions

This study presents the value of MC based algorithms for the simulation of numeric and categorical data. Here we have applied these methods to two different types of paleoanthropological datasets, with an additional example of how these methods can be leveraged for the simulation of 3D models as well. As can be seen, MC based approaches are a powerful tool to generate synthetic, yet statistically realistic data, that can be used to power more robust statistical tests. This holds numerous distinct advantages over the frequently (and often inaptly) used bootstrapping technique.

There are two main limitations of any data augmentation technique; (1) even though augmentation exists, it should never be used to replace a large sample size; and (2) the quality of augmented data is only as good as the input data. In the present study, this can be seen in the example of simulated orangutan data, which was restricted by the small size of the original sample and its homogeneity (no male individuals were included).

Nevertheless, here we present a toolkit for the handling of multimodal datasets, with the hope to implement more versatile and adaptable versions of MCs and MCMCs in future releases of the

*AugmentationMC* library. Future research will thus focus on improving the computational cost of these algorithms and enhance their precision.

## Acknowledgements

L.A.C. is funded by the Spanish Ministry of Science, Innovation and Universities with an FPI Predoctoral Grant (Ref. PRE2019-089411), associated with the project RTI2018-099850-B-I00 and the University of Salamanca. We would like to express our gratitude to Christopher Ruff for his transparency when sharing his data with us. J.A. is funded by the Euskal Herriko Unibertsitatea through its programme for postdoctoral specialization [ESPDOC21/05].

## Author Contributions

**L.A.C**. Conceptualization, Formal Analysis, Investigation, Methodology (Data Augmentation and Statistics), Software, Visualization, Writing – Original Draft, Review & Editing.
**J.A.** Data Curation, Formal Analysis, Methodology (Geometric Morphometrics), Resources, Validation, Visualization, Writing – Original Draft, Review & Editing.
**D.G.A.** Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing.

## Conflict of Interest Statement

We have no competing interests to declare.

## References

Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L. (2018) Learning representations and generative models for 3D point clouds. *International Conference on Learning Representations*. https://arxiv.org/abs/1707.02392 [Accessed: 17/06/2022]

Albrecht, G.H. (1992) Assessing the Affinities of Fossils using Canonical Variates and Generalised Distances. *Journal of Human Evolution,* 7, 49-69.

Alemseged, Z., Spoor, F., Kimbel, W.H., Bobe, R., Geraads, D., Reed, D., Wynn, J.G. (2006) A juvenile early hominin skeleton from Dikika, Ethiopia. *Nature*, 443, 296-301.

Almécija, S., Tallman, M., Alba, D.M., Pina, M., Moyà-Solà, S., Jungers, W.L. (2013) The femur of Orrorin tugenensis exhibits morphometric affinities with both Miocene apes and later hominins. *Nature Communications*, 4, 2888.

Almécija, S., Tallman, M., Sallam, H.M., Fleagle, J.G., Hammond, A.S., Seiffert, E.R. (2019) Early anthropoid femora reveal divergent adaptive trajectories in catarrhine hind-limb evolution. *Nature Communications*, 10, 4778.

Aramendi, J. (2021) *A new morphometric approach to the study of Plio-Pleistocene hominin biomechanics and adaptation*. PhD Dissertation, Madrid: University Complutense of Madrid.

Arsuaga, J.L., Martínez, I., Gracia, A., Carretero, J-M., Carbonell, E. (1993) Three new human skulls from the Sima de los Huesos Middle Pleistocene site in Sierra de Atapuerca, Spain. *Nature*, 362, 534-537.

Beaudet, A., Dumoncel J., Heaton, J.L., Pickering, T.R., Clarke, R.J., Carlson, K.J., Bam, L., Hoorebeke, L.V., Stratford, D. (2022) Shape analysis of the StW 578 calotte from Jacovec

Cavern, Gauteng (South Africa). *South African Journal of Science*, 118, #11743, DOI: 10.17159/sajs.2022/11743.

Beaudet, A., Heaton, J.L., L'Abbé, E.N., Pickering, T.R., Stratford, D. (2018) Hominin cranial fragments from Milner Hall, Sterkfontein, South Africa. *South African Journal of Science*, 114, #5262, DOI: 10.17159/sajs.2018/5262

Berger, L.R., de Ruiter, D.J., Churchill, S.E., Schmid, P., Carlson, K.J., Dirks, P.H.G.M., Kibii, J.M. (2010) Australopithecus sediba: A New Species of Homo-Like Australopith from South Africa. *Science*, 328, 195-204.

Berger, L.R., Hawks, J., de Ruiter, D.J., Churchill, S.E., Schmid, P., Delezene, L.K., Kivell, T.L., Garvin, H.M., Williams, S.A., DeSilva, J.M., Skinner, M.M., Musiba, C.M., Cameron, N., Holliday, T.W., Harcourt-Smith, W., Ackermann, R.R., Bastir, M., Bogin, B., Bolter, D., Brophy, J., Cofran, Z.D., Congdon, K.A., Deane, A.S., Dembo, M., Drapeau, M., Elliott, M.C., Feuerriegel, E.M., García-Martínez, D., Green, D.J., Gurtov, A., Irish, J.D., Kruger, A., Laird, M.F., Marchi, D., Meyer, M.R., Nalla, S., Negash, E., Orr, C.M., Radovcic, D., Schroeder, L., Scott, J.E., Throckmorton, Z., Tocheri, M.W., VanSickle, C., Walker, C.S., Wei, P., Zipfel, B. (2015) Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa. *eLife*, 4, e09560.

Bergstra, J., Bengio, Y. (2012) Random search for hyper-parameter optimization. *Journal of Machine Learning Research,* 13, 281–305.

Bermúdez de Castro, J.M., Martínez, I., Gracia-Téllez, A., Martinón-Torres, M., Arsuaga, J.L. (2020) The Sima de los Huesos Middle Pleistocene hominin site (Burgos, Spain). Estimation of the number of individuals. *The Anatomical Record,* 304, 1463-1477.

Bishop, C. (2006) *Pattern Recognition and Machine Learning*. Springer, Singapore.

Bookstein, F.L. (1989) Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11, 567-585.

Bookstein, F.L. (1991) *Morphometric Tools for Landmark Data*. Cambridge University Press, New York.

Bookstein, F.L. (2017) A newly noticed formula enforces fundamental limits on geometric morphometric analyses. *Evolutionary Biology*, 44, 522-541.

Bookstein, F.L. (2019) Pathologies of between-groups Principal Components Analysis in Geometric Morphometrics. *Evolutionary Biology*, 46, 271-302.

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P. (2017) Geometric Deep Learning. *IEEE Signal Processing Magazine*, 34, 18-42.

Brown, F., Harris, J., Leakey, R., Walker, A. (1985) Early Homo erectus skeleton from west Lake Turkana, Kenya. *Nature*, 316, 788-792.

Calder, J., Coil, R., Melton, A., Olver, P.J., Tostevin, G., Yezzi-Woodley, K. (2022) Use and misuse of Machine Learning in Anthropology. *arXiv Preprint*, https://arxiv.org/abs/2209.02811 [Accessed: 19/09/2022]

Cardini, A., O'Higgins, P., Rohlf, F.J. (2019) Seeing distinct groups where there are none: spurious patterns from Between-Group PCA. *Evolutionary Biology*, 46, 303-316.

Carlin, B.P., Louis, T.A. (2008) *Bayesian Methods for Data Analysis*. Chapman and Hall, Boca Raton.

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yu, L., Yu, F. (2015) ShapeNet: An Information-Rich 3D Model Repository. *arXiv Preprint*. https://arxiv.org/abs/1512.03012 [Accessed: 17/06/2022]

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357

Chow, G.C. (1960) Tests for equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3), 591-605.

Clarke, R.J. (1998) First ever discovery of a well-preserved skull and associated skeleton of Australopithecus. *South African Journal of Science*, 94, 460-463.

Claxton, A.G., Hammond, A.S., Romano, J., Oleinik, E., DeSilva, J.M. (2016) Virtual reconstruction of the Australopithecus africanus pelvis Sts 65 with implications for obstetrics and locomotion. *Journal of Human Evolution*, 99, 10-24.

Cohen, J. (1988) *Statistical Power Analysis for Behavioural Sciences*. Routledge, New York.

Colquhoun, D. (2019) The False Positive Risk: A proposal concerning what to do about *p*-values. *The American Statictician*, 73, 192–201.

Cortes, C., Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, 20, 273-297.

Courtenay, L.A., Yravedra, J., Huguet, R., Aramendi, J., Maté-González, M.Á., González-Aguilera, D., Arriaza, M.C. (2019) Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks. *Paleogeography, paleoclimatology, paleoecology*, 522, 28-39.

Courtenay, L.A., González-Aguilera, D. (2020) Geometric Morphometric Data Augmentation using Generative Computational Learning Algorithms. *Applied Sciences*, 10, 9133.

Courtenay, L.A., Herranz-Rodrigo, D., González-Aguilera, D., Yravedra, J. (2021) Developments in Data Science Solutions for Carnivore Tooth Pit Classification. *Scientific Reports,* 11, 10209.

Crevecoeur, I., Bayle, P., Rougier, H., Maureille, B., Higham, T., van der Plicht, J., De Clerck, N., Semal, P. (2010) The Spy VI child: A newly discovered Neandertal infant. *Journal of Human Evolution*, 59, 641-656.

Daver, G., Berillon, G., Jacquier, C., Ardagna, Y., Yadeta, M., Maurin, T., Souron, A., Blondel, C., Coppens, Y., Boisserie, J.R. (2018) New hominin postcranial remains from locality OMO 323, Shungura Formation, Lower Omo Valley, Southwestern Ethiopia. *Journal of Human Evolution,* 122, 23-32.

Davis, C.A., Profico A., Kappelman, J. (2021) Digital restoration of the Wilson-Leonard 2 Paleoindian skull (~10,000 BP) from central Texas with comparison to other early American and modern crania. *American Journal of Biological Anthropology*, 176, 486-503.

Di Vincenzo, F., Rodríguez, L., Carretero, J.M., Collina, C., Geraads, D., Piperno, M., Manzi, G. (2015) The massive fossil humerus from the Oldowan horizon of Gombore I, Melka Kunture (Ethiopia, >1.39 Ma). *Quaternary Science Reviews*, 122, 207-221.

Domínguez-Rodrigo, M. (2018) Successful classification of experimental bone surface modifications (BSM) through machine learning algorithms: a solution to the controversial use of BSM in palaeoanthropology. *Archaeological and Anthropological Sciences*, 11, 2711-2725.

Domínguez-Rodrigo, M., Pickering, T.R., Baquedano, E., Mabulla, A., Mark, D.F., Musiba, C., Bunn, H.B., Uribelarrea, D., Smith, V., Diez-Martin, F., Pérez-González, A., Sánchez, P., Santonja, M., Barboni, D., Gidna, A., Ashely, G., Yravedra, J., Heaton, J.L., Arriaza, M.C. (2013) First Partial Skeleton of a 1.34-Million-Year-Old Paranthropus boisei from Bed II, Olduvai Gorge, Tanzania. *PLoS ONE*, 8, e80347.

Dryden, I.L., Mardia, K.V. (1998) *Statistical Shape Analysis*. John Willey, Chichester.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics,* 7, 1-26

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. (2018) *Learning from Imbalanced Data Sets*. Springer, Heidelberg.

Fisher, R.A., (1935) The Design of Experiments. Hafner Pub, New York.

García-Martin, R., Bautista-De Castro, Á., Sánchez-Aparicio, L.J., Fueyo, J.G., González-Aguilera, D. (2019) Combining digital image correlation and probablistic approaches for the reliability analysis of compositve pressure vessels. *Archives of Civil and Mechanical Engineering,* 19, 224-239.

Gelman, A., Roberts, G.O., Gilks, W.R.(1995) Efficient Metropolis jumping rules. In: Bernardo, J.M., Berger, J., Dawid, A.P., Smith, A.F.M. (Eds.) *Bayesian Statistics Vol. 5*. Oxford University Press, Oxford, pp. 599-608.

Giró, A., Padró, J.A., Valls, J., Wagensberg, J. (1985) Monte carlo simulation of an ecosystem: a matching between two levels of observation. *Bulletin for Mathematical Biology*, 47, 111-122.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014) Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27, 1-9.

Goodfellow, I.J., Shlens, J., Szegedy, C. (2015) Explaining and Harnessing Adversarial Examples, International Conference on Learning Representations. *arXiv Preprint*, https://arxiv.org/abs/1412.6572v3 [Accessed: 17/06/2022]

Goodfellow, I., Bengio, Y., Courville, A. (2016) *Deep Learning.* MIT Press, Cambridge.

Gordon, N.J., Salmond, D.J., Smith, A.F.M. (1993) Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEEE Proceedings F: Radar and Signal Processing*, 140, 107-113

Gosset, W.S. (1908) The probable error of a mean. Biometrika 6, 1-25

Graves, T.L. (2011) Automatic step size selection in Random Walk Metropolis algorithms. *arXiv Preprint*, https://arxiv.org/abs/1103.5986 [Accessed: 17/06/2022]

Greenwood, J.A., Sandomire, M.M (1950) Sample Size required for estimating the standard deviation as a percent of its true value. *Journal of The American Statistical Association,* 45, 257-260

Grine, F.E. (2019) The alpha taxonomy of Australopithecus at Sterkfontein: The postcranial evidence. *Comptes Rendus Palevol*, 18, 335-352.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. (2017) Improving training of Wasserstein GANs. *Neural Information Processing Systems*, 31, 5769-5779.

Gunz, P., Harvati, K. (2007) The Neanderthal "chignon": Variation, integration, and homology. *Journal of Human Evolution,* 52, 262-274

Gunz, P., Neubauer, S., Falk, D., Tafforeau, P., La Cabec, A., Smith, T.M., Kimbel, W.H., Spoor, F., Alemseged, Z. (2020) Australopithecus afarensis endocasts suggest brain like organization and prolonged brain growth. *Scientific Advances*, 6, eaaz4729

Gunz, P., Mitteroecker, P., Bookstein, F.L., Weber, G.W. (2004) Computer Aided Reconstruction of Human Crania. In: Wien, M., Erbe, R.K., Wien, S. (Eds.) *Enter the Past: Computer Applications and Quantitative Methods in Archaeology*. BAR International Series Vol. 1227, Oxford, pp. 92-94

Gunz, P., Mitteroecker, P., Neubauer, S., Weber, G.W., Bookstein, F.L. (2009) Principles for the Virtual Reconstruction of Hominin Crania. *Journal of Human Evolution*, 57, 48-62.

Gupta, P.L., Gupta, R.D. (1987) Sample Size Determination in Estimating a Covariance Matric. *Computational Statistics and Data Analysis*, 5, 185-192

Harmon, E. (2009) The shape of the early hominin proximal femur. *American Journal of Physical Anthropology*, 139, 154–171.

Hastings, W.K. (1970) Monte-Carlo sampling methods using Markov Chains and their applications, *Biometrika*, 57, 97-109

Hawks, J., Elliott, M., Schmid, P., Churchill, S.T., de Ruiter, D.J., Roberts, E.M., Hilbert-Wolf, H., Garvin, H.M., Williams, S.A., Delezene, L.K., Feuerriegel, E.M., Randolph-Quinney, P., Kivell, T.L., Laird, M.F., Tawane, G., DeSilva, J.M., Bailey, S.E., Brophy, J.K., Meyer, M.R., Skinner, M.M., Tocheri, M.W., VanSickle, C., Walker, C.S., Campbell, T.L., Kuhn, B., Kruger, A., Tucker, S., Gurtov, A., Hlophe, N., Hunter, R., Morris, H., Peixotto, B., Ramalepa, M., van Rooyen, D., Tsikoane, M., Boshoff, P., Dirks, PHGM, Berger, L.R. (2017) New fossil remains of Homo naledi from the Lesedi Chamber, South Africa. *eLife*, 6, e24232.

He, H., Bai, Y., Garcia, E.A., Li, S. (2008) ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks*. 1322-1328

He, H., Ma, Y. (2013) *Imbalanced Learning*. IEEE Press, New Jersey.

Höhle, J., Höhle, M. (2009) Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64, 398-406.

Johanson, D.C., Edey, M.A. (1981) *Lucy: The beginnings of humankind*. Simon and Schuster, New York.

Johanson, D.C., Taieb, M., and Coppens, Y. (1982) Pliocene hominids from the Hadar Formation, Ethiopia (1973–1977): Stratigraphic, chronologic, and paleoenvironmental contexts, with notes on hominid morphology and systematics. *American Journal of Physical Anthropology*, 57, 373–402.

Jollife, I. (2002) *Principal Component Analysis*. Springer, New York.

Kipf, T.N., Welling, M. (2017) Semi-supervised classification with Graph Convolutional Networks, *International Conference on Learning Representations*. Online: https://arxiv.org/abs/1609.02907 [Accessed: 17/06/2022]

Lague, M.R. (2002). Another look at shape variation in the distal femur of Australopithecus afarensis: implications for taxonomic and functional diversity at Hadar. *Journal of Human Evolution*, 42, 609–626.

Lague, M.R. (2015) Taxonomic identification of Lower Pleistocene fossil hominins based on distal humeral diaphyseal cross-sectional shape. *PeerJ*. 3, e1084.

Lague, M.R., Chirchir, H., Green, D.J., Mbua, E., Harris, J.W.K., Braun, D.R., Griffin, N.L., Richmond, B.G. (2019a) Humeral anatomy of the KNM-ER 47000 upper limb skeleton from Ileret, Kenya: Implications for taxonomic identification. *Journal of Human Evolution*, 126, 24-38.

Lague, M.R., Chirchir, H., Green, D.J., Mbua, E., Harris, J.W.K., Braun, D.R., Griffin, N.L., Richmond, B.G. (2019b) Cross-sectional properties of the humeral diaphysis of Paranthropus boisei: Implications for upper limb function. *Journal of Human Evolution*, 126, 51-70.

Lakens, D. (2017) Equivalence tests: a practical primer for t tests, correlations and meta analyses. *Social Psychological and Personality Science,* 8(4), 355-362.

Leakey, M. D. (1971) *Olduvai Gorge*, Vol. 3, Cambridge University Press, Cambridge.

LaPlace, P.S. (1823) *Théorie Analytique des Probabilités*. Courcier, Paris.

Lordkipanidze, D., Ponce de León, M.S., Margvelashvili, A., Rak, Y., Rightmire, P., Vekua, A., Zollokofer, C.P.E. (2013) A Complete Skull from Dmanisi, Georgia, and the Evolutionary Biology of Early Homo. *Science*, 342, 326-331.

Martin, O. (2018) *Bayesian Analysis with Python*. Packt, Birmingham.

Martinón-Torres, M., d'Errico, F., Santos, E., Álvaro Gallo, A., Amano, N., Archer, W., Armitage, S.J., Arsuaga, J.L., Bermúdez de Castro, J.M., Blinkhorn, J., Crowther, A., Douka, K., Dubernet, S., Faulkner, P., Fernández-Colón, P., Kourampas, N., González García, J., Larreina, D., Le Bourdonnec, F.X., MacLeod, G., Martín-Francés, L., Massilani, D., Mercader, J., Miller, J.M., Ndiema, E., Notario, B., Pitarch Martí, A., Prendergast, M.E., Queffelec, A., Rigaud, S., Roberts, P., Shoaee, M.. J., Shipton, C., Simpson, I., Boivin, N., Petraglia, M.D. (2021) Earliest known human burial in Africa. *Nature* 593, 95–100.

McPherron, S.P., Archer, W., Otárola-Castillo, E.R., Torquato, M.G., Keevil, T.L. (2022) Machine Learning, Bootstrapping, Null Models, and why we are still not 100% sure which Bone Surface Modification were made by crocodiles. *Journal of Human Evolution*, 164, 103071

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.

Mitteroecker, P., Bookstein, F. (2011) Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evolutionary Biology*, 38, 100–114.

Mitteroecker, P., Gunz, P. (2009) Advances in Geometric Morphometrics. *Evolutionary Biology*, 36, 235-247.

Moclán, A., Domínguez-Rodrigo, M., Yravedra, J. (2019) Classifying agency in bone breakage: an experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using Machine Learning (ML) algorithms. *Archaeological and Anthropological Sciences*, 11, 4663-4680

Mori, T., Profico, A., Reyes-Centeno, H., Harvati, K. (2020) Frontal bone virtual reconstruction and geometric morphometric analysis of the mid-Pleistocene hominin KNM-OG 45500 (Olorgesailie, Kenya). *Journal of Anthropological Sciences*, 98, 49-72.

Pearson, K.X. (1900) On the Criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 50, 151-175.

Pisonero, J., López-Rebollo, J., García-Martín, R., Rodríguez-Martín, M., Sánchez-Aparicio, L.J., Muñoz-Nieto, A., González-Aguilera, D. (2021) A comparative study of 2D and 3D digital image correlation approaches for the characterization and numerical analysis of composite materials. *IEEE Access,* 9, 160675-160687.

Qi, C.R., Yi, L., Su, H., Guibas, L.J. (2017) PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Proceedings of the Neural Information Processing Systems*. Online: https://arxiv.org/abs/1706.02413 [Accessed: 17/06/2022]

Razali, N.M. (2011) Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2, 21-33

Rodríguez-Martín, M., Rodríguez-Gonzálvez, P., Ruiz de Oña Crespo, E., González-Aguilera, D. (2019) Validation of portable mobile mapping system for inspection tasks in termal and fluid-mechanical facilities. *Remote Sensing,* 11, 2205.

Rohlf, F.J. (1996) Morphometric spaces, shape components and the effects of linear transformations. In Marcus, L.F., Corti, M., Loy, A., Naylor, G.J.P., Slice, D.E. (Eds.) *Advances in Morphometrics*. Springer, The Netherlands, pp. 117-129.

Rohlf, F.J. (1998) On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology*, 47, 147-158.

Ruff, C.B., Hayes, W.C. (1983) Cross-sectional geometry of Pecos Pueblo femora and tibiae – a biomechanical investigation: 1. Method and General Patterns of Variation. *American Journal of Physical Anthropology*, 60, 359-381

Ruff, C. (1987) Sexual dimorphism in human lower limb bone structure: relationship to subsistence strategy and sexual division of labor. *Journal of Human Evolution*, 16, 391-416

Ruff, C.B. (1991) *Aging and Osteoporosis in Native Americans from Pecos Pueblo*, New Mexico. Garland, New York.

Ruff, C.B. (1995) Biomechanics of the Hip and Birth in Early Homo. *American Journal of Physical Anthropology*, 98, 527-574

Ruff, C.B., McHenry, H.M., Thackeray, J.F. (1999) Cross-sectional geometry of the SK 82 and 97 proximal femora. *American Journal of Physical Anthropology*, 109, 509–521.

Ruff, C.B., Burgess, M.L., Squyres, N., Junno, J.A., Trinkaus, E. (2018) Lower limb articular scaling and body mass estimation in Pliocene and Pleistocene hominins. *Journal of Human Evolution*, 115, 85-111

Schurimann, D.L. (1987) A comparison of the two one-sided test procedure and the power approach for assessing the equivalence of average biovariability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657-680.

Shahrirari, B., Swersky, K., Wang, Z., Adams, R.P., Freitas, N. (2016) Taking the Human out of the loop: a review of Bayesian optimization. *Proceedings of the IEEE*, 104, 148-175.

Shu, D.W., Park, S.W., Kwon, J. (2019) 3D Point Cloud Generative Adversarial Network based on Tree Structured Graph Convolutions. *IEEE International Conference on Computer Vision*. https://arxiv.org/abs/1905.06292 [Accessed: 17/06/2022]

Snoek, J., Larochelle, H., Adams, R.P. (2012) Practical Bayesian Optimization of Machine Learning Algorithms. *Proceedings of the International Conference on Neural Information Processing Systems*, 25, 2951-2959.

Souza, D.M., Matos, M.P.S., Oliveira-Neto, N.M., Albuquerque, R.V.T. (2020) An analysis of pseudo-first-order behaviour in biomolecular chemical reactions via stochastic computational simulation. *Revista Virtual de Quimica*, 12, 598-607

Such, F.P., Rawal, A., Lehman, J., Stanley, K.O., Clune, J. (2019) Generative Teaching Networks: Accelerating Neural Architecture search by Learning to Generate Synthetic Training Data. *arXiv Preprint*, https://arxiv.org/abs/1912.07768 [Accessed: 17/06/2022]

Susman, R.L. (1989) New Hominid Fossils from the Swartkrans Formation (1979-1 986 Excavations): Postcranial Specimens. *American Journal of Physical Anthropology*, 79, 451-474.

Susman, R.L., de Ruiter, D., Brain, C.K. (2001) Recently identified postcranial remains of Paranthropus and Early Homo from Swartkrans Cave, South Africa. *Journal of Human Evolution*, 41, 607-629.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2014) Intriguing properties of neural networks. *International Conference on Learning Representations*, https://arxiv.org/abs/1312.6199 [Accessed: 17/06/2022]

Takeshita, T., Nozawa, S., Kimura, F. (1993) On the bias of Mahalanobis distance due to limited sample size effect. *Proceedings of the International Conference on Document Analysis and Recognition*, 2, 171-174.

Tallman, M. (2012) Morphology of the Distal Radius in Extant Hominoids and Fossil Hominins: Implications for the Evolution of Bipedalism. *The Anatomical Record*, 295, 454-464.

Tallman, M., Almécija, S., Reber, S.L., Alba, D.M., Moyà-Solà, S. (2013) The distal tibia of Hispanopithecus laietanus: More evidence for mosaic evolution in Miocene apes. *Journal of Human Evolution*, 64, 319-327.

Tallman, M. (2014) Phenetic and Functional Analyses of the Distal Ulna of Australopithecus afarensis and Australopithecus africanus. *The Anatomical Record*, 298, 195-211.

Tanaka, F.H.K.S., Aranha, C. (2019) Data Augmentation using GANs. *arXiv Preprint*. arXiv: https://arxiv.org/abs/1904.09135 [Accessed: 17/06/2022]

Trinkaus, E., Ruff, C.B., Churchill, S.E., Vandermeersch, B. (1998) Locomotion and body proportions of the Saint-Césaire 1 Châtelperronian Neandertal. *Proceedings of the National Academy of Sciences*, 95, 5836–5840.

Valsesia, D., Magli, E., Fracastoro, G. (2019) Learning localized generative models for 3D point clouds via Graph Convolution. *International Conference on Learning Representations*. https://openreview.net/pdf?id=SJeXSo09FQ [Accessed: 17/06/2022]

Ward, C.V., Plavcan, J.M., Manthi, F.K. (2020) New fossils of Australopithecus anamensis from Kanapoi, West Turkana, Kenya (2012-2015). *Journal of Human Evolution*, 140, 102368

Wang, Y., Sun, Y., Liu, Z.; Sarma, S.E., Bronstein, M.M., Solomon, J.M. (2019) Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphs*, 1, 1-13. DOI: 10.1145/3326362

Wood, B.A. (1974) Olduvai Bed I Post-cranial Fossils: A Reassessment. *Journal of Human Evolution*, 3, 373-378.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J. (2015) 3D ShapeNets: A Deep Representation for Volumetric Shapes. *IEEE Conference on Computer Vision and Pattern Recognition*, https://arxiv.org/abs/1406.5670 [Accessed: 17/06/2022]

Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B., 2017. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative Adversarial Modeling. *Conference on Neural Information Processing Systems*. https://arxiv.org/abs/1610.07584 [Accessed: 17/06/2022]

Yuen, K.K., Dixon, W.J. (1973) The approximate behaviour and performance of the two-sample trimmed t. *Biometrika*, 60, 349-374.

Yuen, K.K. (1974) The two-sample trimmed t for unequal population variances. *Biometrika*, 61, 165-170.

*Publications: Section 3*

# Artificially Intelligent Classification Algorithms

*Spanish Translation of Title and Abstract*

# Una aproximación híbrida a través de morfometría geométrica y aprendizaje profundo para la clasificación de marcas de corte y pisoteo.

El concepto de equifinalidad es actualmente uno de los mayores problemas de la tafonomía, que con frecuencia lleva a los analistas a interpretar erróneamente la formación y funcionalidad de los yacimientos arqueológicos y paleontológicos. Un ejemplo de esta equifinalidad se encuentra en la diferenciación entre las marcas de corte antrópicas, y otros rastros en el hueso producidos por agentes naturales, como la abrasión sedimentaria y el pisoteo. Estas cuestiones son un componente clave para la comprensión de la evolución humana temprana, pero a menudo se basan en rasgos cualitativos para su identificación. Desafortunadamente, los datos cualitativos suelen ser susceptibles a errores debido a la subjetividad, a veces fruto de la experiencia del analista. El presente estudio pretende afrontar estas cuestiones mediante un enfoque metodológico híbrido. En este caso, combinamos datos de morfometría geométrica, microscopía digital 3D, y aprendizaje profundo mediante redes neuronales, para proporcionar un medio para la clasificación empírica de las huellas tafonómicas en el hueso. Los resultados obtenidos alcanzan una tasa de clasificación superior al 95%, proporcionando un posible medio para superar la equifinalidad tafonómica en el registro arqueológico y paleontológico.

**Supplementary Information available from:**
`https://www.mdpi.com/2076-3417/10/1/150#supplementary`

**Code available from:**
`https://github.com/LACourtenay/Deep-Neural-Network-for-Cut-Mark-Classification`

*applied sciences*

MDPI

*Article*

# A Hybrid Geometric Morphometric Deep Learning Approach for Cut and Trampling Mark Classification

**Lloyd A. Courtenay [1,2,3,*], Rosa Huguet [2,3,4], Diego González-Aguilera [1] and José Yravedra [5,6]**

[1] Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003 Ávila, Spain; daguilera@usal.es

[2] Área de Prehistoria, Universitat Rovira i Virgili (URV), Avignuda de Catalunya 35, 43002 Tarragona, Spain; rhuguet@iphes.cat

[3] Institut Català de Paleoecologia Humana i Evolució Social (IPHES), C/ Marcellí Domingo s/n, Campus Sescelades URV (Edifici W3) E3, 43700 Tarragona, Spain

[4] Unit Associated to CSIC, Departamento de Paleobiologia, Museo de Ciencias Naturales, C/ José Gutiérrez Abascal, s/n, 28006 Madrid, Spain

[5] Department of Prehistory, Complutense University, Prof. Aranguren s/n, 28040 Madrid, Spain; joyravedra@hotmail.com

[6] Director of the C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren s/n, 28040 Madrid, Spain

* Correspondence: ladc1995@gmail.com; Tel.: +34-633-647-825

Received: 28 October 2019; Accepted: 19 December 2019; Published: 23 December 2019

**Featured Application: Cut mark identification and analysis is a fundamental component for archaeological investigation. Cut mark analysis, however, has been the root of great debates, with some authors claiming to have the oldest cut marks in or outside of Africa. If these marks were to truly be anthropic in nature, then the repercussions of these findings would produce a paradigm shift for our understanding of human evolution. Unfortunately, the majority of methods available for cut mark classification are namely qualitative in nature. Here we provide a new, highly powerful artificially intelligent neural network classification model that can be used to quantitatively and more objectively overcome these issues, using 3D digital microscopy, Deep Learning and Geometric Morphometrics to obtain up to 100% accuracy in some cases.**

**Abstract:** The concept of equifinality is currently one of the largest issues in taphonomy, frequently leading analysts to erroneously interpret the formation and functionality of archaeological and paleontological sites. An example of this equifinality can be found in the differentiation between anthropic cut marks and other traces on bone produced by natural agents, such as that of sedimentary abrasion and trampling. These issues are a key component in the understanding of early human evolution, yet frequently rely on qualitative features for their identification. Unfortunately, qualitative data is commonly susceptible to subjectivity, producing insecurity in research through analyst experience. The present study intends to confront these issues through a hybrid methodological approach. Here, we combine Geometric Morphometric data, 3D digital microscopy, and Deep Learning Neural Networks to provide a means of empirically classifying taphonomic traces on bone. Results obtained are able to reach over 95% classification, providing a possible means of overcoming taphonomic equifinality in the archaeological and paleontological register.

**Keywords:** taphonomy; microscopy; equifinality; archaeological data science

## 1. Introduction

The publication of the 'oldest' anthropic evidence of any type is always a problematic issue, usually drawing attention, criticism, and eventual debate on the quality of these findings from the entire archeological and paleontological community. Perfect examples of such debates can be observed in the claims of ~3.4 Ma cut marks from Dikika, Ehtiopia [1], which have since been heavily criticized and rejected [2,3]. Likewise, sites claiming to have ~2.6 Ma cut marks outside of Africa in the province of Quranwala, India [4], have drawn some speculation to their authenticity. In the Americas, ~130 ka anthropic bone breakage [5] are also noted to be located in areas with highly abrasive sediments and problematic taphonomic contexts. The current consensus for the oldest cut marks in Africa, however, remains to be those of Gona [6], dated to approximately between 2.1 and 2.58 Ma, while other promising results have been localized with 1.9 and 2.4 Ma in Northern Africa [7].

Taphonomic debates revolving around these topics are essential in understanding features of human evolution, considering how current theories argue meat-eating to be a fundamental component of our evolution [8–11]. The concept of butchery contains a multitude of different implications beginning with resource acquisition [8,12–15], as well as the cognitive technical capacities to manufacture the instruments used for such activities [16–20]. Dates of cut marks at 3.3 Ma implicate Australopithecine populations to be the first users of tools and butcherers in hominin pre-history [1], however authors are yet to come to an agreement as to whether these individuals were physically capable of such practices [16–20]. While an argument has been proposed to say that natural edges of unknapped stones could be used for butchery practices [1], other authors argue that experimentation is yet to be found that supports this claim [3]. Nevertheless, if these findings were to be real, then strong empirical evidence would be needed in support of such a hypothesis.

Recent advances in the development of new methodologies for the study of Bone Surface Modifications (BSMs) have been able to reveal interesting patterns in the in-depth study of taphonomic traces. The implementation of Geometric Morphometric studies has been able to reveal a means of inferring different tool use [21] as well as raw material management [21–23] through cut mark morphologies. Moreover, when applied to the carnivore induced BSMs, analysts have been able to differentiate between carnivore agents based on tooth mark morphologies [24–27]. The innovative introduction of Artificial Intelligence (AI) in taphonomy [26–31] has additionally been able to overcome multiple barriers imposed by subjectivity [32]. This presents a powerful tool for the construction of classification models, presenting a series of efficient tools for the processing of complex data.

Here we present the power of Feed Forward Neural Networks (FFNN) trained through Deep Learning (DL) for the processing of morphological data obtained with advanced 3D digital microscopy. These efforts attempt to overcome issues imposed by equifinality and subjectivity in taphonomic research, as well as complement previously obtained data regarding the effectivity of Machine Learning (ML) algorithms for the processing of Geometric Morphometric information [26]. Through this, we present a new means of classifying cut marks and trampling marks through their morphological attributes, as well as an empirically objective and quantitative approximation to their morphological description and characterization.

## 2. Materials and Methods

Experimental samples consisted of 80 cut marks and 251 trampling marks (Figure 1). Sample sizes where chosen in accordance with statistical power tests [33,34], defining a minimum sample size of 59 individuals as significant for the type of analysis performed within this paper (Cohen's $d = 0.52$, $\alpha = 0.05$, power = 0.8).

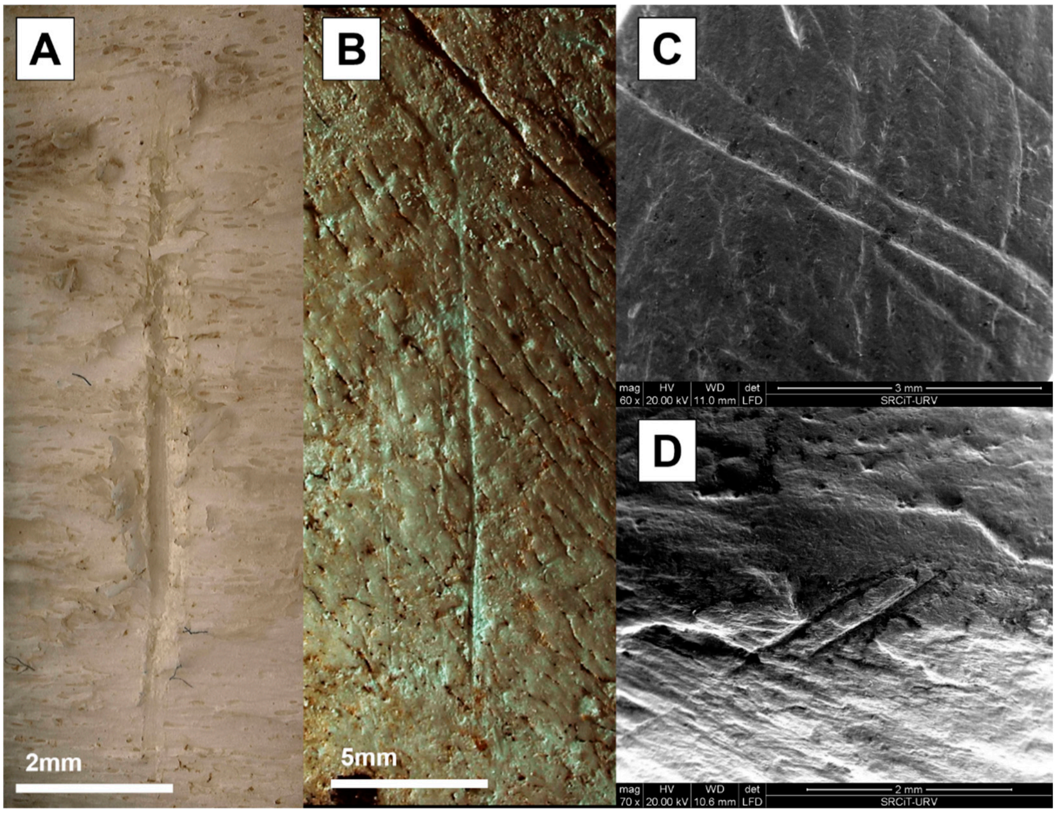


**Figure 1.** Photographic examples of (**A**,**B**) experimental and (**C**,**D**) archaeological taphonomic traces. Photographic documentation using the HIROX KH-8700 microscope of experimental (**A**) cut and (**B**) trampling marks. Scanning Electron Microscope photographs of (**C**) inconspicuous taphonomic traces that could be cut marks and (**D**) clear cut marks found on antelope bone from Frida Leakey Korongo West of the Olduvai Gorge (Tanzania). Photos by L.A.C. Archaeological remains studied first by J.Y. [35] and later by L.A.C.

Cut marks (Figure 1A) were produced using simple flakes knapped by a single right-handed individual, experienced and familiar with lithic materials from the Olduvai Gorge and other Lower Pleistocene sites. The raw material used for these experiments was obtained directly from the Naibor Soit Inselberg of the Olduvai Gorge (Tanzania) [23]. This raw material consists of a coarse-grained quartzite frequently found in multiple sites of Beds I and II of the Olduvai Gorge. Cut marks were produced on a mixture of adult bovid and suid individuals on a number of different anatomical elements, including femora, tibiae, and humerii. All cut marks were produced by a single right-handed individual, perpendicular to the bone while the bone was fresh and the meat intact.

Trampling mark samples (Figure 1B) were obtained from Domínguez-Rodrigo et al. [36]'s sample. These traces were produced under a number of different experimental conditions. Experiments were carried out using cervid bones obtained from a legal organized hunting party. Anatomical elements present a mixture of axial and appendicular elements, including femora, tibiae, radii, ulnae, humerii, vertebrae, ribs, and scapulae. The majority of the meat from these bones were removed with metal knives, then sectioned into smaller pieces using an electric saw. Each bone was then examined to avoid misclassifying BSMs produced by the defleshing and sectioning processes. The sample was then separated into multiple subsamples that were subjected to different experimental conditions. The first variable considered was the sediment type. Five different sedimentary conditions were used, the first consisting in fine-grained sands (60–200 μm), followed by medium-grained sands (200–600 μm), coarse-grained sands (0.6–2 mm), a combination of the different sand types in a clay stratum, and finally gravels (>2 mm). Additional variables considered the time exposed to trampling

(10 s or 2 min), the individuals producing the trampling (all students of varying weights), and whether the bones were dry or fresh when buried. For more details consult citation reference [36].

## 2.1. Digital Reconstruction Technique

A combination of two different methodological approaches was used for this study, the first concerning the 3D digital reconstruction protocol via advanced digital microscopy [37] followed by the processing of this data via a 3D 13-landmark model [21].

The digitalization process was performed using the HIROX KH-8700 3D Digital Microscope with an MXG-5000 REZ triple objective revolving lens located in the *Institut Català de Paleoecologia Humana I Evolució Social* (IPHES), Tarragona, Spain. The HIROX is equipped with a high intensity LED light source that can be positioned around the subject of study. For this study, the light source was positioned directly above the object, combining both coaxial and ring lighting conditions without the use of any polarized filters. Digital reconstructions of each trace were performed between 100× (Field of View (FOV) = 1516 µm) and 200x (FOV = 3032 µm) magnification, using either the low or medium range lens. Three-dimensional reconstructions were produced using the HIROX's mosaic tiling function, specifying a minimum of 30 photos per tile. This process takes approximately 13 min to complete per mark [37]. Collection of 3D landmark data was performed directly within the HIROX's system software, employing the use of multiple measurement systems to obtain x, y, and z coordinates for the position of each landmark. This landmark data was then formatted and imported into R (https://www.r-project.org/) for further statistical analysis.

Further technical details regarding the microscope and a detailed description of the reconstruction protocol can be consulted in [37], as briefly and graphically described in Figure 2.

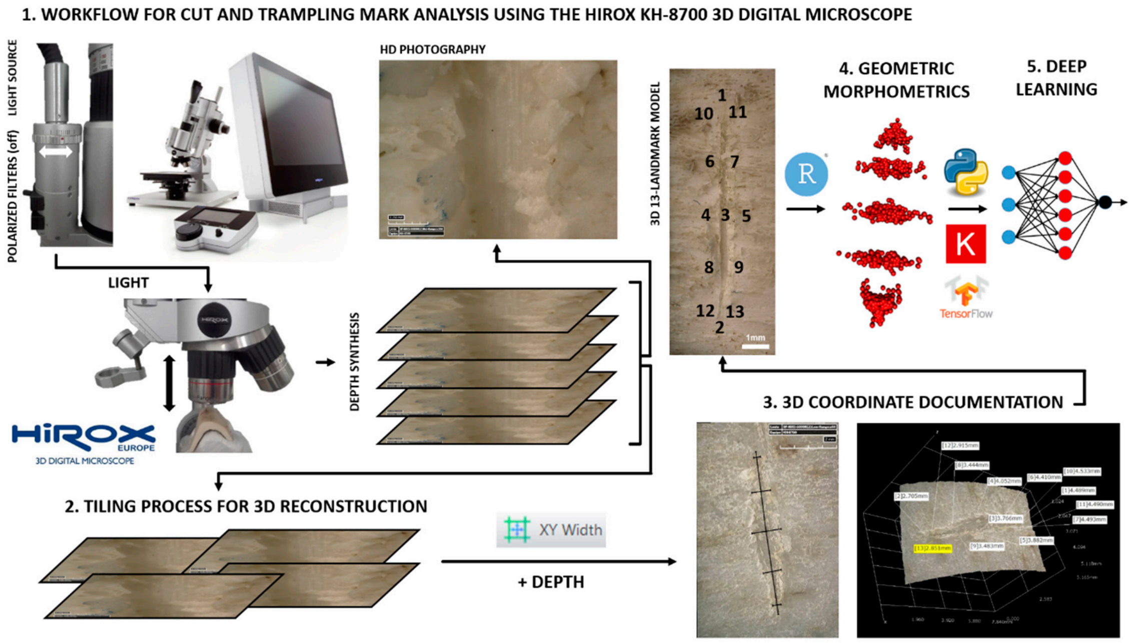

1. WORKFLOW FOR CUT AND TRAMPLING MARK ANALYSIS USING THE HIROX KH-8700 3D DIGITAL MICROSCOPE
POLARIZED FILTERS (off) LIGHT SOURCE
LIGHT
HIROX EUROPE
3D DIGITAL MICROSCOPE
HD PHOTOGRAPHY
DEPTH SYNTHESIS
3D 13-LANDMARK MODEL
1
10 11
6 7
4 3 5
8 9
12 13
2
1mm
R
4. GEOMETRIC MORPHOMETRICS
K
TensorFlow
5. DEEP LEARNING
3. 3D COORDINATE DOCUMENTATION
2. TILING PROCESS FOR 3D RECONSTRUCTION
XY Width
+ DEPTH

**Figure 2.** Graphical description of the proposed methodological workflow using the HIROX KH-8700 3D Digital Microscope, as described in detail through Courtenay et al. [21,37]. Figure by L.A.C.

## 2.2. Geometric Morphometrics

Geometric morphometric analysis was performed in the free statistical software R (https://www.r-project.org/), employing the use of multiple packages that can be consulted in Appendix A Table A1.

For the Geometric Morphometric analysis of both linear traces, a 3D 13-landmark model was used [21]. This model combines landmark types I and II to capture the internal as well as external morphological features of each trace. Landmark data is fist processed using a full Procrustes fit and an orthogonal tangent projection [38], known as Generalized Procrustes Analysis (GPA), normalizing data for further multivariate statistical analyses. GPA is a common practice in Geometric Morphometrics for the standardization of form information through multiple superimposition procedures, including translation, rotation, and scaling. Differences through this are revealed through patterns of variation and covariation that can be assessed statistically [39,40]. Principal Components Analyses (PCA) are performed on this data to reduce the complex combination of variables to fewer dimensions [39]. Additional thin plate splines, grid warpings, and mean shapes were calculated to visualize morphological variation across the Principal Component (PC) scores [39]. Degree of variance is then assessed using pairwise Multiple Variance Analyses (MANOVA). Depending on the inter-group homogeneity within each sample, MANOVA calculations were adjusted either using the Wilks or Hotelling–Lawley formula for inhomogeneous and homogeneous samples, respectively.

Samples were also processed using a Canonical Variance Analysis (CVA). CVA consists in the transformation of the raw PCA data, whereby pooled within-group dispersion are manipulated in a scaling process, thus standardizing within-group variance, and finally rotating the axes to be redrawn as a CVA graph [40]. Distances were then calculated between the groups with permutated *p*-values from the pooled within-group covariance matrices, calculating the degree of separation between samples in the form of Procrustes and Mahalanobis distances with their associated *p*-values of significance.

To ensure the efficiency of the learning process, PC scores were bootstrapped 1000× and extracted for the construction of Deep Learning Feed Forward Neural Network models.

### 2.3. Deep Learning

Deep Learning applications were programmed in Python (https://www.python.org/), using a number of different packages that can be consulted in Appendix A Table A2. Algorithms conceptualized for supervised training and classification of samples consisted in the development of Feed Forward Neural Networks (FFNN). Neural Networks are modeled and coded to replicate brain patterns that are able to process highly complex and large sets of data [41], consisting of multiple nodes or perceptrons, which are connected by weighted axons or edges. These networks are designed to recognize patterns in order to interpret data, utilizing components of mathematics, calculus, linear algebra, and statistics to train and perform different tasks [42,43]. Likewise, this can be performed in a supervised, semisupervized, or unsupervised manner.

The FFNN designed for this study was constructed in TensorFlow 2.0 using the Keras API [41]. All Deep Learning implications were therefore run using TensorFlow as a backend engine on a portable laptop's CPU (Intel® Core$^{\mathrm{TM}}$ i5 6300HQ), executed in a Conda (https://www.anaconda.com) virtual environment. The network was trained using the PC scores as dependencies, employing NumPy to convert PC scores into 64-bit floating point matrices. The associated class labels were indexed as separate 64-bit floating point vectors. PC scores are frequently used in Machine and Deep Learning as a method for projecting high-dimensional data into a new feature space that is useful for the training of AI models. Here we used the top 10 PC scores, representing 93% of the sample's variance, using these PC scores to train models to classify unknown individuals.

A combination of neurons in a mixture of hidden layers are then used to map out the relationships between the dependency inputs ($x$) and the label outputs ($y$) [42,43]. A generalized mathematical representation of a single neuron can thus be represented as [42]:

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where $w$ are the weights connecting each neuron and *f()* represents an activation function that varies according to the position within the network [42]. Considering the case at hand consists in a binary classification problem, the label values were converted into a string of 1s and 0s indicating whether the taphonomic trace is anthropic (1) or not (0). In order to ensure that the model therefore only produces an output between 1 and 0, a sigmoid activation function is used for the final layer, described mathematically as [42]:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

While all hidden layers were activated using the Rectified Linear Unit (ReLU) function [42]:

$$f(x) = \max(0, x) \tag{3}$$

During the learning process, the training of the model searches for the optimal combination of weights that can efficiently map out the *y(x)* relationship. Neural Networks are stochastic; therefore, weight initialization is performed at random [41–44]. The tuning of these weights was performed using back propagation and gradient descent via a stochastic optimization algorithm [42,43,45–48] and a binary cross entropy loss function [42,43].

For the purpose of configuring the neural network and finding the right hyperparameters during optimization, a series of trial runs were performed. These alternated between different combinations of hyperparameters, searching for the best results without overfitting. These trials employed typical practices of Deep Learning techniques [44], including changing the number of hidden layers, number of neurons per layer, batch size, epoch size, kernel constraints, weight regularization, the presence, position, or threshold of dropout layers as well as different optimization algorithms and learning rates. A summary of the hyperparameters tested can be consulted in Table 1.

**Table 1.** List of hyperparameters and settings tested during optimization of final Feed Forward Neural Networks (FFNN) model architecture and configurations.

| Hyperparameter | Tested Settings | References |
|---|---|---|
| Number of Layers | Between 1 and 7 | [41,42,44] |
| Node Density * | Between 3 and 20 | [41,42,44] |
| Activation Function * | ReLU, Leaky ReLU, Tanh | [49–52] |
| Kernel Initializer | None, Uniform | [51,53] |
| Dropout ** | None, Present with a threshold of 0.5 between 0.5 and 0.9 | [54] |
| Weight Regularizer | None, l2 with a threshold between 0.01 and 0.00001 | [53] |
| Weight Constraint | UnitNorm, MaxNorm, MinMaxNorm | [54] |
| Training Epochs | Between 150 and 2000 | [41,42,44] |
| Batch Size | 4, 8, 16, 32, 64, 128, 256 | [41,42,44] |
| Optimizer | Stochastic Gradient Descent, RMSProp, AdaGrad, Adam | [45–48] |
| Learning Rate | Between 0.1 and 0.00001 | |
| Decay | Between 0.9 and 0.0001 | |
| Momentum | Between 0.99 and 0.1 | |

* With exception of the final layer of the network which always consisted of 1 neuron with a sigmoid activation function. ** Positioning of dropout layers within the model's architecture was also tested.

Models were trained and evaluated using training, validation, and test splits. This is common practice in both Machine and Deep Learning [41,44,55]. The train-test split consisted of a 70:30% split ratio, respectively. During training, the training sample was further split using 30% for validation. FFNN were then trained on training and validation data, optimizing weights to improve the accuracy and reduce the loss. Learning curves were plotted to evaluate the increase/decrease of accuracy and loss over each iteration epoch. The metric used to evaluate the learning process while the model was being fit was set to 'accuracy'. These learning curves could then be used to diagnose model behavior, thus evaluating whether the model was under- or overfitting on the training and validation data [44].

Final evaluation of the model was performed using the test set. The model was used to predict this 'unknown' data, recording both the accuracy and the loss obtained when predicting the missing label values from this data. The final metrics employed and evaluated using the test set consisted of sensitivity, specificity, and kappa values obtained via confusion matrices. The kappa ($\kappa$) statistic adjusts accuracy by considering the possibility of a correct prediction by change alone [55]. The resulting value is presented between −1 and 1, with a κ > 0.8 considered as a powerful predictive model. Sensitivity and specificity tests combine the frequencies and ratios of Type I and Type II statistical errors in proportion with the rest of the confusion matrix [56]. Values between 0 (poor) and 1 (high performing) indicate the predictive power of the model [55–57]. Further examination of the relationship between sensitivity and specificity in models was performed through the plotting of Receiver Operating Characteristic (ROC) curves and calculation of Area Under Curve (AUC) values [55,56]. ROC curves and AUC results are interpreted through the amount of space represented underneath the curve: the larger the area (AUC ≈ 1), the more accurate the model is when making predictions [55].

Considering the stochastic nature of FFNNs, evaluation and training was performed 30 times, taking averages of each numeric result to provide the final results. Results across all 30 iterations are provided as Tables S1 and S2. Confidence intervals were then calculated using the 2nd Standard Deviation (±2SD), thus representing approximately 95% of the deviation from the mean.

The final Python code used for this study is available in the form of a Jupyter Notebook online at https://github.com/LACourtenay/Deep-Neural-Network-for-Cut-Mark-Classification.

## 3. Results

### 3.1. Geometric Morphometrics

PCA was able to produce up to 32 PC scores, with the first 10 representing up to 93% of the total variance (Figure 3). The first two components of this analysis represent a cumulative variance of 52% of the sample, displaying a high degree of overlapping among trampling samples and anthropogenic cut marks. Regardless, trampling marks can be seen to represent a much greater degree of variability, with a much larger proportion of the sample displaying a trend toward a wider and more superficial morphology. Cut marks occupy a much smaller percentage of the overall feature space, leaning much closer to the end of PC1 that is represented by a finer groove. Additionally, trampling marks are also seen to vary greatly across the second principal component, which is represented by variations in groove trajectory. Contrarily, cut marks display a restricted distribution. Analysis of thin plate splines across PC3 displays a much clearer tendency for cut marks to lean toward deeper marks (Figure 4A), with trampling marks occupying primarily a percentage of feature space that is represented by very superficial traces.

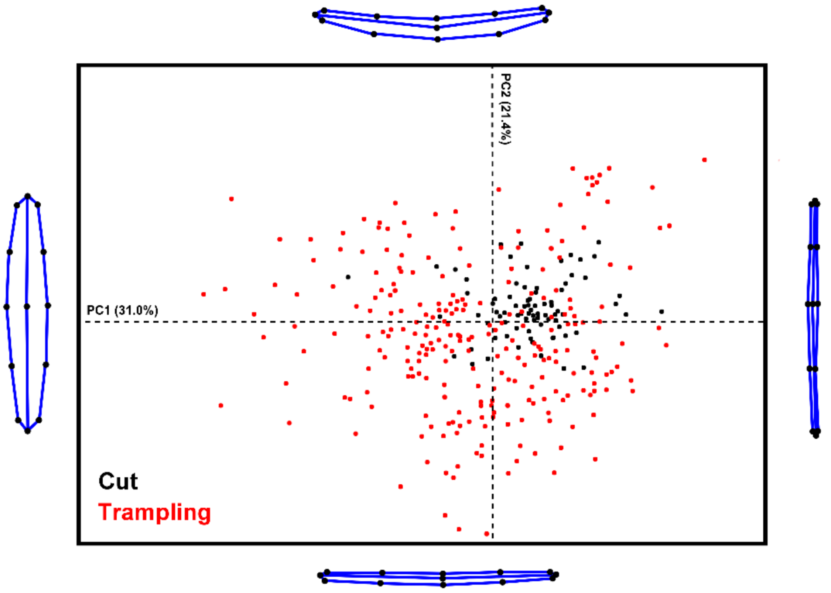


**Figure 3.** Principal Components Analysis comparing Cut and Trampling Mark morphologies. Extreme shape changes can be observed at the extremity of each PC score.

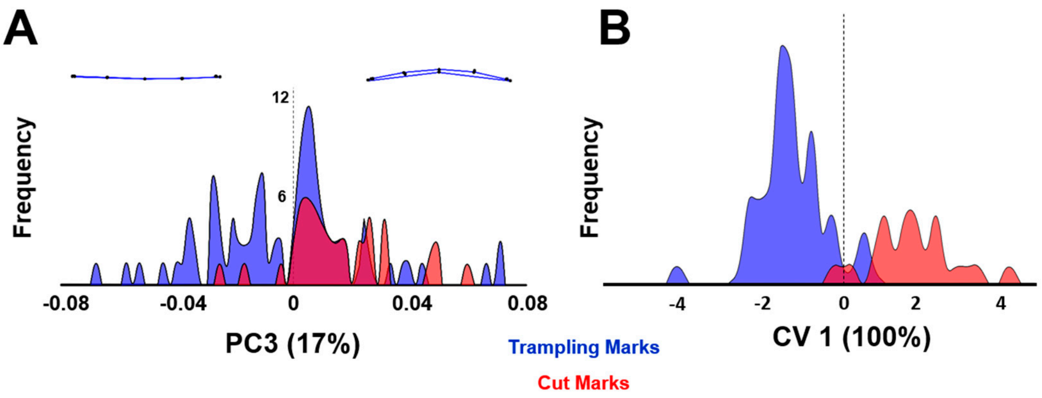


**Figure 4.** Principal components and Canonical Variance Analysis comparing Cut and Trampling Mark morphologies. (**A**) Distribution of samples across Principal Component 3 with extreme shape changes graphically presented at the extremity of each side of the axis. (**B**) Distribution of samples in Canonical Variance Analyses.

Exploring these variations through numerical results highlight significant differences between samples, with MANOVA of $p = 0.001$ between both groups. Mahalanobis (D = 3.4403, $p < 0.0001$) and Procrustes (D = 0.0297, $p = 0.0001$) distance calculations also concur, with a clear separation between groups in CVA graphs (Figure 4B), represented by a total of 100% across the single axis of this figure.

### 3.2. Deep Learning

Initial trials prior to hyperparameter optimization and tuning began by achieving a model accuracy of approximately 70%, while overfitting proved to be a considerable issue with most model training and validation sets, even at this low degree of accuracy. After hyperparameter optimization, the final model obtained between 97.63% and 100% accuracy differentiating between trampling and cut marks (Figure 5, Table 2), presenting variation due to the stochastic nature of the model during weight initialization. The final model employed the use of 6 layers, 5 standard neural layers, and 1 dropout layer (Figure 5A and Figure S1). The inclusion of a larger density layer after the input (number of neurons = 20) produced a significant boost in accuracy, yet in order to prevent this additional layer from producing an over generalization of the data, a dropout layer with a constraint threshold of 0.5 was included directly afterwards (Figure 5A and Figure S1). A number of different positions for the dropout layer were tried and tested, yet the best results were obtained positioning said dropout in-between layers 3 and 5. An additional "UnitNorm" weight constraint was used to reduce overfitting, while the best training performance was obtained using the Adam optimization algorithm (learning rate ($\alpha$) = 0.001, decay ($\beta_1$) = 0.9). No additional regularization or kernel initialization techniques were found necessary for the final model.

**Table 2.** Performance accuracy and loss of training, validation, and testing of Neural Network after 30 iterations. All accuracy values are presented in percentages.

| | | Training | Validation | Testing |
|---|---|---|---|---|
| Accuracy | Max | 100.00 | 100.00 | 100.00 |
| | Mean | 99.56 | 99.50 | 99.59 |
| | Upper CI | 100.00 | 100.00 | 100.00 |
| | Lower CI | 98.97 | 98.74 | 98.95 |
| | Min | 98.43 | 97.63 | 98.00 |
| Loss | Max | 0.13 | 0.09 | 0.02 |
| | Mean | 0.05 | 0.02 | 0.01 |
| | Upper CI | 0.08 | 0.05 | 0.01 |
| | Lower CI | 0.02 | 0.00 | 0.00 |
| | Min | 0.00 | 0.00 | 0.00 |

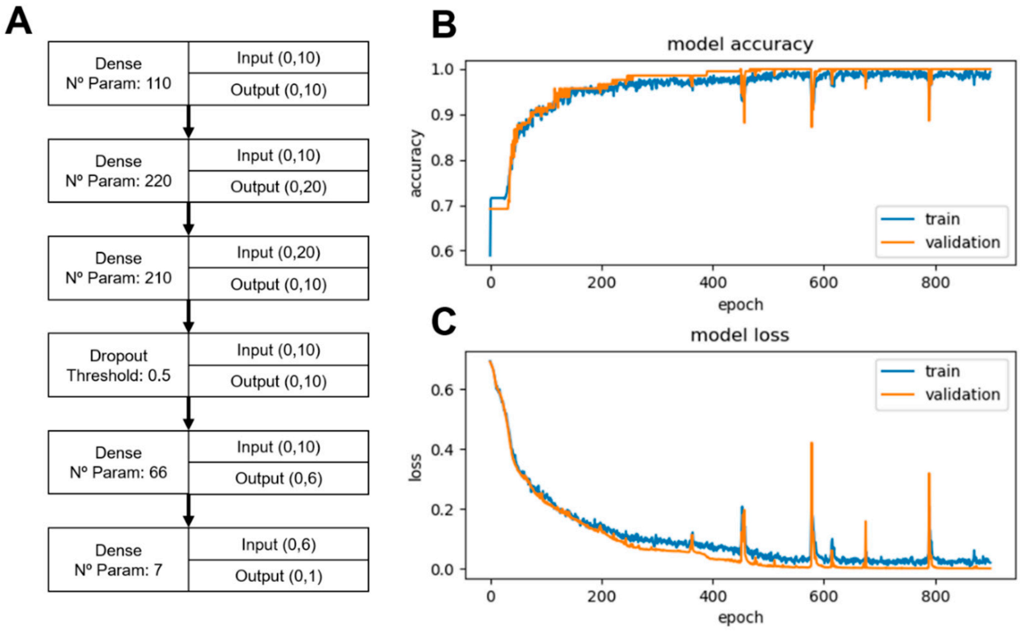


**Figure 5.** Feed Forward Neural Network Architecture and Learning Curve. (**A**) Visualization of the Neural Network Architecture including input and output shape for each layer, n° of adjustable parameters, and the type of layer. A graphical representation of this has been provided as Supplementary Figure S1. (**B**) Accuracy learning curve for validation and training over epochs. (**C**) Loss learning curve for validation and training over epochs.

The final training process used 900 epochs and a microbatch size of 64, obtaining an average accuracy of 99.55 ± 1.32% across training, testing, and validation samples (Figure 5B, Table 2 & Table S1). Loss on all accounts highlights the FFNN to be a powerful classifier with high confidence when assigning class labels to new individuals (Figure 5C). In training the average loss was recorded at 0.05, while 0.02 was recorded for validation and 0.01 for testing (Table 2, Table S1).

Further model evaluation through confusion matrices obtained on model testing was able to confirm the FFNN to be a highly efficient classification model, differentiating between cut and trampling marks with κ values of 1 ± 0.008 (Table 3, Table S2). Likewise, both sensitivity and specificity values averaged at 1 with the lowest specificity value being recorded at 0.995 and all sensitivity values obtaining 1 as well (Figure 6, Table 3, Table S2). ROC graphs almost always display a perfect right hand angle rather than a curve (Figure 6), with AUC values averaging at 1 ± 1.2 x $10^{-4}$.

**Table 3.** Average Neural Network performance evaluation on test sets.

| | **Mean** | **2 SD** |
|---|---|---|
| Sensitivity | 1.000 | 0.000 |
| Specificity | 0.999 | 0.004 |
| Kappa | 0.997 | 0.008 |
| AUC | 1.000 | 0.000 |

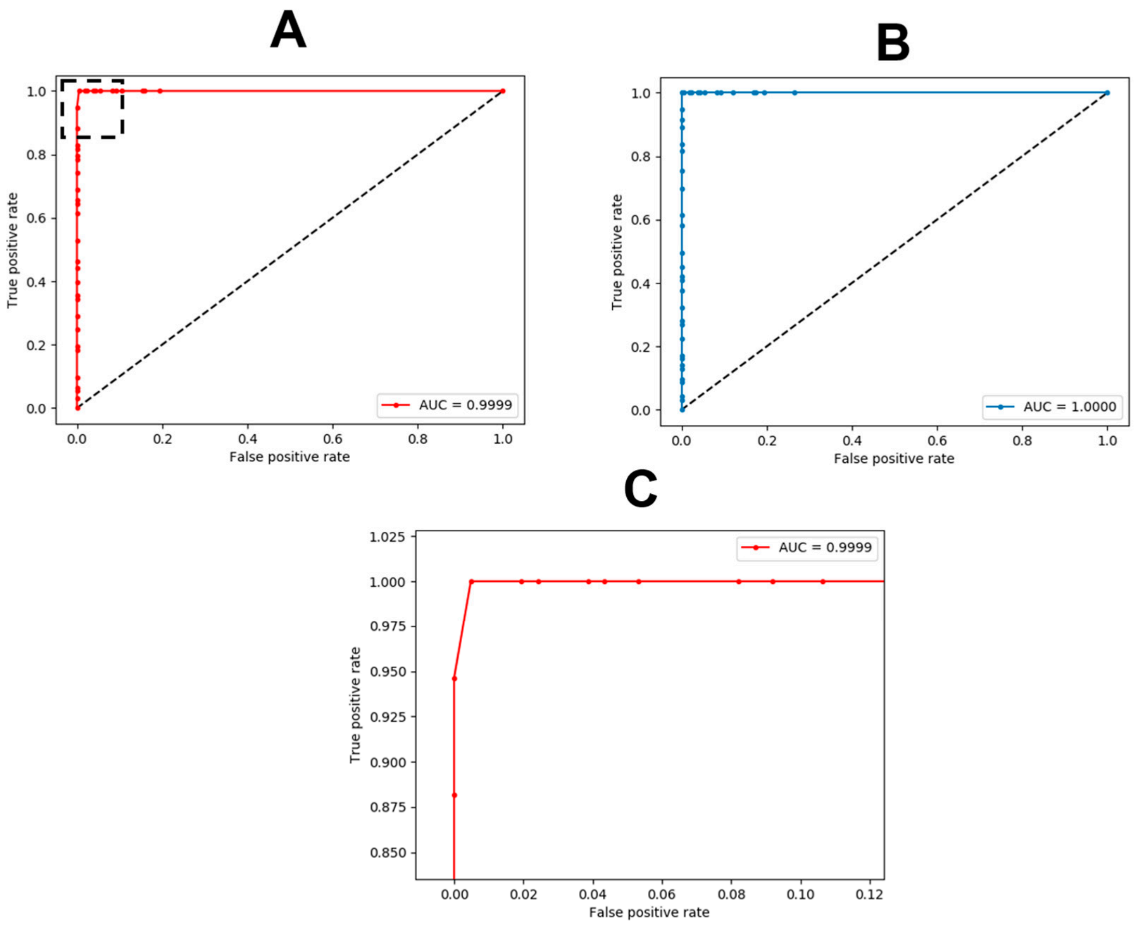


**Figure 6.** ROC curves with AUC values for Neural Network performance. (**A**) Worst recorded performance of the Neural Network model, obtaining an AUC of 0.9999. (**B**) best recorded performance of the Neural Network model, obtaining an AUC of 1. (**C**) Detail of dotted-square in (**A**), showing the slight perfection affecting this model's achievement of a perfect right-angle "curve".

Finally, FFNN training time averaged at 10.87 s while taking as little as ~17 milliseconds when making predictions.

## 4. Discussion and Conclusions

BSM analysis remains to be a very important component of taphonomic studies, whereby their identification and in-depth analysis can reveal multiple components regarding early hominin populations [11–15], their development [6,7,23], and their associated paleoecologies [14,24,28]. Nevertheless, issues imposed by equifinality have led to complications in their identification and interpretation [1–5], requiring more objective and empirical methods that can be used for BSM classification and characterization [21–27,30,58].

In recent years, debates regarding the protocol used to identify cut marks have ranged from simple observational criteria [58] to developments with a more complex multivariate protocol [36] and advanced microscopic studies [59]. With the integration of ML to the processing of qualitative data, analysts have been able to improve the processing of archaeological data sets to a considerable degree [29]; nevertheless, the subjective nature upon which this data is obtained makes some of these advances debatable [32]. One alternative has been proposed utilizing DL convolutional neural network architectures for image processing and classification [30], presenting promising results for automated BSM identification. Here, we additionally present the combined usage of advanced digital microscopy, Geometric Morphometrics, and artificially intelligent computational algorithms for cut mark identification and characterization. On both accounts, ML and DL have effectively proven to outperform human performance [29,30,32], presenting a more objective and precise means of studying microscopic traces.

The present results are able to develop data observed by multiple authors [12,36,37,58], yet employing new empirical means of quantifying these conclusions. Geometric morphometric characterization of trampling and cut marks concur that the most significant features of cut marks are their depth and straight trajectory, while trampling marks are more variable presenting much more superficial morphologies alongside other irregularities [37]. Furthermore, the ability of high-resolution 3D digital microscopy to overcome limitations imposed by the superficial nature of some traces [37] can also be considered a significant improvement from previous efforts [21,22]. While equifinality can still be observed to a certain degree, considering the high degree of overlap in most of these samples, it is important to point out the high dimensionality of PCA results derived from morphological data, as seen in how MANOVA testing is still able to identify significant differences between samples. Moreover, FFNN efficiently differentiates experimental samples with high levels of confidence on all accounts, considering their ability to extract complex patterns from difficult data [60].

The present study additionally complements previous efforts to implement ML algorithms in Geometric Morphometric analyses, expanding the available toolbox for morphological studies [26,61–66]. Courtenay et al. [26]'s original attempts to implement neural network architectures for tooth mark classification performed poorly, attributed by the authors to the model's superficial nature. The complexity of the model here supports this observation. These results thus confirm model configuration and tuning to be essential for efficient classification, requiring extensive experimentation to find the optimal model. This would also explain the mixed results obtained by similarly superficial models in applications for systematic biology [61–66].

The field of AI can be seen to have exponentially grown since its conceptualization, providing algorithmic computational means of processing complex data sets. Many of these algorithms have presented significant advances for other disciplines, including medical research [67], pharmaceutics [68], business studies [69], engineering [70], and any other fields that require the advanced processing of large and complex data sets [60]. In prehistoric archaeology, ML and DL have arrived relatively late, yet present promising results. Nevertheless, problems of true experimental analogy are needed before these approaches can be applied on a broader scale. Here, Naibor Soit quartzite is used, considering this raw material's importance in many Pleistocene sites of the Olduvai Gorge; however, if this approach were to be applied to other sites in Europe, Asia, or the Americas, then the experimental protocol and reference sample should be adjusted accordingly. Moreover, analysts should be aware of the possible overlapping traces that may increase the effects of taphonomic equifinality over time, such as fluvial abrasion, chemical alterations, and general loss of cortical surfaces [71,72], to name a few.

In other practical cases, the drawbacks of DL can be presented by limited sample sizes as well as the cost of training. This is especially apparent in archaeology and palaeoanthropology considering the conservation and preservation of the fossil record present considerable limitations. Nevertheless, Geometric Morphometric data has still proven to be a powerful type of input data for training, proving relatively fast to learn patterns from >1 min. Furthermore, considering the nature of the landmark data involved and its consequent transformation through GPA and PCA dimensionality reduction methods, this type of input data is less prone to issues presented by sample size as opposed to studies concerning, for example, Computer Vision and image processing-based techniques [54,73–76], the latter requiring large amounts of parameters (usually in the millions), which are hard to learn from small datasets [76].

Needless to say, as with the case of any innovative methodological introduction in archaeological and paleontological research, a large-scale use of these techniques is usually slow and requires large experimental programs to truly fine tune these results and examine their limitations.

DL and ML provide a significant advance for classification problems and predictive modeling [55], almost regardless of the type of data being analyzed. Advances in data science are presenting new means of automating data collection and processing, presenting a new empirical basis that can be used to confirm or reject cases of controversial taphonomic interpretations [1,4,5]. Neural Networks are highly versatile computational algorithms and can be adapted to most data sets [41,60]. Their success, however, is highly dependent on the tuning of their configuration and the developments available when considering the options feature engineering and hyperparameter optimization may provide [44,54,60,75,76]. Here, we have tested the potential of Deep Learning on the processing of morphological data to provide a hybrid approach that efficiently overcomes one of the taphonomy's biggest questions. The present work thus demonstrates an example of how advanced microscopy and developed artificially intelligent algorithms may provide a promising future for archaeological and paleontological science.

**Supplementary Materials:** The following are available online at http://www.mdpi.com/2076-3417/10/1/150/s1: Figure S1: Visualization of the Neural Network Architecture. Bias neurons are represented in green, hidden neurons in blue and the dropout layer in red., Table S1: Accuracy and Loss Results from 30 prediction iterations on training, validation and test data, and Table S2: Results from 30 prediction iterations on test data

**Author Contributions:** Conceptualization, L.A.C., R.H., and J.Y.; methodology, L.A.C.; software, L.A.C.; validation, L.A.C.; formal analysis, L.A.C.; investigation, L.A.C.; resources, L.A.C., R.H., and J.Y.; data curation, L.A.C.; writing—original draft preparation, L.A.C.; writing—review and editing, L.A.C., R.H., D.G.-A., and J.Y.; visualization, L.A.C.; supervision, R.H. and J.Y.; project administration, R.H., D.G.-A., and J.Y.; funding acquisition, R.H. and D.G.-A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** First we would like to thank Manuel Domínguez-Rodrigo for providing the experimental samples used within this study. We would also like to acknowledge his inspirational work in this field. We would like to thank all the staff and members of both the IPHES and the Rovira I Virgili Univeriisty. The corresponding author would also like to acknowledge the support provided by the TIDOP Groups 1 & 2 from the Department of Cartographic and Land Engineering of the Higher Polytechnics School of Avial, University of Salamance, and especially the help, advice, and support provided by Diego González-Aguilera and Miguel Ángel Maté-González. We would also like to thank Andreu Ollé for his advice, support, and interest in this research, as well as his technical support regarding the microscope and the lithic implements used within this study. We would also like to thank the helpful comments and suggestions provided by two anonymous reviewers and the editorial staff of MDPI. Finally, this work was supported by MICINN-FEDER PGC2018-093925-B-C32 and the AGUAR project number SGR 2017-1040: the Universitat Rovira I Virgili (2014, 2015, and 2016 PFR-URV-B2-17).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A

For the purpose of this study, a mixture of R (https://www.r-project.org/) (Table A1) and Python (https://www.python.org/) (Table A2) were used for data science applications.

**Table A1.** Table presenting the R libraries used for geometric morphometric applications.

| Library | Used For: | Link |
|---|---|---|
| Geomorph | Generalized Procrustes Analysis<br>Principal Components Analysis<br>Thin Plate Splines | https://cran.r-project.org/web/packages/geomorph/geomorph.pdf |
| Shapes | Canonical Variate Analysis | https://cran.r-project.org/web/packages/shapes/shapes.pdf |
| RVAideMemoire | Multivariate Variance Analysis | https://cran.r-project.org/web/packages/RVAideMemoire/RVAideMemoire.pdf |

**Table A2.** Table presenting the Python libraries used for Deep Learning applications.

| Library | Used for: | Link |
|---|---|---|
| TensorFlow 2.0.<br>Keras API | Neural Network Construction<br>Hyperparameter Optimization | https://www.tensorflow.org/<br>https://keras.io/ |
| Numpy | Numerical applications and operations<br>Slicing, indexing and transformation of data | https://www.numpy.org/ |
| Pandas | Loading data | https://pandas.pydata.org/ |
| Matplotlib | Plotting learning curves and ROC results | https://matplotlib.org/ |
| Scikit-Learn | Model evaluation | https://scikit-learn.org/ |

For handling of Python and Deep Learning applications, the open source Anaconda (https://www.anaconda.com/) software was used to manage libraries and internal environments for Deep Learning. For debugging of Python code Jupyter Notebook (https://jupyter.org/) was used. For debugging of R code, R-Studio was used (https://www.rstudio.com/).

## References

1. McPherron, S.P.; Alemseged, Z.; Marean, C.W.; Wynn, J.G.; Reed, D.; Geraads, D.; Bobe, R.; Béarat, H. Evidence for stone-tool-assisted consuption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. *Nat. Lett.* **2010**, *466*, 857–860. [CrossRef]
2. Domínguez-Rodrigo, M.; Pickering, T.R.; Bunn, H.T. Experimental study of cut marks made with rocks unmodified by human flaking and its bearing on claims of ~3.4-Million-Year-Old butchery evidence from Dikika. *J. Arch. Sci.* **2012**, *39*, 205–214. [CrossRef]
3. Domínguez-Rodrigo, M.; Alcalá, L. 3.3 million year old stone tools and butchery traces? More evidence needed. *Paleoanthropology.* **2016**, *2016*, 46–53.
4. Malassé, A.D.; Moigne, A.M.; Singh, M.; Calligaro, T.; Karir, B.; Gaillard, C.; Kaur, A.; Bharwaj, V.; Pal, S.; Abdessadok, S.; et al. Intentional cut marks on bovid from the Quranwala Zone, 2.6 Ma, Siwalik Frontal Range, Northwestern India. *Comptes Rendus Palevol* **2016**, *15*, 317–339. [CrossRef]
5. Holen, S.R.; Deméré, T.A.; Fisher, D.C.; Fullagar, R.; Paces, J.B.; Jefferson, G.T.; Beeton, J.M.; Cerutti, R.A.; Rountret, A.N.; Vescera, L.; et al. A 130,000 year old archaeological site in Southern California, USA. *Nature* **2017**, *544*, 479–483. [CrossRef] [PubMed]
6. Domínguez-Rodrigo, M.; Pickering, T.R.; Semaw, S.; Rogers, M.J. Cutmarked bones from Pliocene archaeological sites at Gona, Afar, Ethiopia: Implications for function of the world's oldest stone tools. *J. Hum. Evol.* **2005**, *48*, 109–121. [CrossRef] [PubMed]
7. Sahnouni, M.; Parés, J.M.; Duval, M.; Cáceres, I.; Harichane, Z.; van der Made, J.; Pérez-González, A.; Abdessadok, S.; Kandi, N.; Derradji, A.; et al. 1.9-Million-Year and 2.4-Million-Year-Old Artifacts and Stone Tool-Cutmarked bones from Ain Boucherit, Algeria. *Science* **2019**, *362*, 1297–1301. [CrossRef] [PubMed]
8. Bunn, H.T. Meat Eating and Human Evolution: Studies on the Diet and Subsistence Patterns of Plio-Pleistocene Hominids in East Africa. Ph.D. Thesis, University of California, Oakland, CA, USA, 1982.
9. Milton, K. Primate Diets and Gut Morphology: Implications for Hominid Evolution. In *Food and Evolution: Toward a Theory of Human Food Habits*; Harris, M., Ross, E.B., Eds.; Temple University: Philadelphia, PA, USA, 1987; pp. 93–115.
10. Aiello, L.C.; Wheeler, P. The expensive tissue hypothesis. *Curr. Anthropol.* **1995**, *36*, 199–221. [CrossRef]
11. Stanford, C.B.; Bunn, H.T. *Meat Eating and Human Evolution*; Oxford University: Oxford, UK, 2001.
12. Binford, L.R. *Bones: Ancient Men and Modern Myths*; Academic Press: New York, NY, USA, 1981.
13. Blumenschine, R. Percussion marks, tooth marks and experimental determinations of hominid and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania. *J. Hum. Evol.* **1995**, *29*, 21–51. [CrossRef]
14. Domínguez-Rodrigo, M. Meat-eating by early hominids at the FLK-22 Zinjanthropus Site, Olduvai Gorge, Tanzania: An experimental approach using cut mark data. *J. Hum. Evol.* **1997**, *33*, 669–690. [CrossRef]

15. Domínguez-Rodrigo, M.; Barba, R. New estimates of tooth mark and percussion mark frequencies at the FLK-Zinj Site: The carnivore-hominid-carnivore hypothesis falsified. *J. Hum. Evol.* **2006**, *50*, 170–194. [CrossRef] [PubMed]
16. Toth, N.; Schick, K. *The Oldowan: Case Studies into the Earliest Stone Age*; Stone Age Institute Press: Gosport, England, 2006.
17. Key, A.J.M.; Dunmore, C.J. The evolution of the Hominin thumb and the influence exerted by non-dominant hand during stone tool production. *J. Hum. Evol.* **2015**, *78*, 60–69. [CrossRef] [PubMed]
18. Toth, N.; Schick, K. An Overview of the Cognitive Implications of the Oldowan Industrial Complex. *Azania Arch. Res. Afr.* **2018**, *53*, 3–39. [CrossRef]
19. Semaw, S.; Roberts, M.J.; Quade, J.; Renne, P.R.; Butler, R.F.; DOmínguez-Rodrigo, M.; Stout, D.; Hart, W.S.; Pickering, T.; Simpson, S.W. 2.6 million year old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia. *J. Hum. Evol.* **2003**, *45*, 169–177. [CrossRef]
20. Domalain, M.; Bertin, A.; Daver, G. Was *Australopithecus Afarensis* able to make the Lomekwian Stone Tools? Towards a realistic biomechanical simulation of hand force capability in fossil hominins and new insights on the role of the fifth digit. *Comptes Rendus Palevol* **2017**, *16*, 572–584. [CrossRef]
21. Courtenay, L.A.; Yravedra, J.; Maté-González, M.Á.; Aramendi, J.; González-Aguilera, D. 3D analysis of cut marks using a new geometric morphometric methodological approach. *J. Arch. Anthr. Sci.* **2019**, *11*, 651–665. [CrossRef]
22. Maté-González, M.Á.; Palomeque-González, J.F.; Yravedra, J.; González-Aguilera, D.; Domínguez-Rodrigo, M. Micro-photogrammetric and morphometric differentiation of cut marks on bones using metal knives, quartzite and flint flakes. *J. Arch. Anthr. Sci.* **2016**, *10*, 805–816. [CrossRef]
23. Courtenay, L.A.; Yravedra, J.; Aramendi, J.; Maté-González, M.Á.; Martín-Perea, D.M.; Uribelarrea, D.; Baquedano, E.; González-Aguilera, D.; Domínguez-Rodrigo, M. Cut marks and raw material exploitation in the Lower Pleistocene Site of Bell's Korongo (BK, Olduvai Gorge, Tanzania): A geometric morphometric analysis. *Quat. Int.* **2019**, *526*, 155–168. [CrossRef]
24. Aramendi, J.; Maté-González, M.A.; Yravedra, J.; Cruz Ortega, M.; Arriaza, M.C.; González-Aguilera, D.; Baquedano, E.; Domínguez-Rodrigo, M. Discerning carnivore agency through the three-dimensional study of tooth pits: Revisiting crocodile feeding leistoc at FLK-Zinj and FLK NN3 (Olduvai Gorge, Tanzania). *Palaeogeog. Palaeoclimat., Palaeoecol.* **2017**, *488*, 93–102. [CrossRef]
25. Yravedra, J.; García Vargas, E.; Maté González, M.A.; Aramendi, J.; Palomeque-González, J.; Vallés-Iriso, J.; Matasanz-Vicente, J.; González-Aguilera, D.; Domínguez-Rodrigo, M. The use of micro-photogrammetry and geometric morphometrics for identifying carnivore agency in bone assemblage. *J. Arch. Sci. Rep.* **2017**, *14*, 106–115. [CrossRef]
26. Courtenay, L.A.; Yravedra, J.; Huguet, R.; Aramendi, J.; Maté-González, M.Á.; González-Aguilera, D.; Arriaza, M.C. Combining Machine Learning Algorithms and Geometric Morphometrics: A Study of Carnivore Tooth Marks. *Palaeogeo Palaeoclim. Palaeoecol.* **2019**, *522*, 28–29. [CrossRef]
27. Yravedra, J.; Maté-González, M.Á.; Courtenay, L.A.; González-Aguilera, D.; Fernández Fernández, M. The use of canid tooth marks on bone for the identification of livestock predation. *Sci. Rep.* **2019**, *9*, 16301. [CrossRef] [PubMed]
28. Arriaza, M.C.; Domínguez-Rodrigo, M. When Felids and Hominins ruled at Olduvai Gorge: A Machine Learning Analysis of Skeletal Profiles of the Non-Anthropogenic Bed I Sites. *Quat. Sci. Rev.* **2016**, *139*, 43–52. [CrossRef]
29. Domínguez-Rodrigo, M. Successful classification of experimental Bone Surface Modifications (BSM) through Machine Learning algorithms: A solution to the controversial use of BSM in paleoanthropology? *J. Arch. Anthro. Sci.* **2019**, *11*, 2711–2725. [CrossRef]
30. Byeon, W.; Domínguez-Rodrigo, M.; Arampatzis, G.; Baquedano, E.; Yravedra, J.; Maté-González, M.A.; Koumoutsakos, P. Automated identification and deep classification of cut marks on bones and its palaeoanthropological implications. *J. Comp. Sci.* **2019**, *32*, 36–43. [CrossRef]
31. Moclán, A.; Domínguez-Rodrigo, M.; Yravedra, J. Classifying agency in bone breakage: An experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using machine learning (ML) algorithms. *Archae. Anthro. Sci.* **2019**, *11*, 4463–4680. [CrossRef]

32. Domínguez-Rodrigo, M.; Saladié, P.; Cáceres, I.; Huguet, R.; Yravedra, J.; Rodríguez-Hidalgo, A.; Martín, R.; Pineda, A.; Marín, J.; Gené, C.; et al. use and abuse of cut mark analyses: The Rorschach Effect. *J. Arch. Sci.* **2017**, *86*, 14–23. [CrossRef]
33. Domínguez-Rodrigo, M.; Juana, S.; Galán, A.B.; Rodríguez, M. A New Protocol to Differentiate Trampling Marks from Butchery Marks. *J. Archaeol. Sci.* **2009**, *36*, 2643–2654. [CrossRef]
34. Cohen, J. *Statistical Power Analysis for Behavioural Sciences*; Lawrence Erlbaum Assoc.: Mahwah, NJ, USA, 1988.
35. Courtenay, L.A.; Maté-González, M.Á.; Aramendi, J.; Yravedra, J.; González-Aguilera, D.; Domínguez-Rodrigo, M. Testing Accuracy in 2D and 3D Geometric Morphometric methods for cut mark identification and classification. *PeerJ* **2018**, *6*, e5133. [CrossRef]
36. Yravedra, J.; Diez-Martín, F.; Egeland, C.P.; Maté-González, M.Á.; Palomeque-González, J.F.; Arriaza, M.C.; Aramendi, J.; García Vargas, E.; Estaca-Gómez, V.; Sánchez, P.; et al. FLK-West (Lower Bed II, Olduvai Gorge, Tanzania): A newearly Acheulean site with evidence for human exploitation of fauna. *Boreas* **2017**, *46*, 486–502. [CrossRef]
37. Courtenay, L.A.; Yravedra, J.; Huguet, R.; Ollé, A.; Aramendi, J.; Maté-González, M.Á.; González-Aguilera, D. New taphonomic advances in 3D digital microscopy: A morphological characterisation of trampling marks. *Quat Int.* **2019**, *517*, 55–66. [CrossRef]
38. Dryden, I.L.; Mardia, K.V. *Statistical Shape Analysis*; John Wiley & Sons: Chichester, UK, 1998.
39. Bookstein, F. *Morphometric Tools for Landmark Data: Geometry and Biology*; Cambridge University Press: New York, NY, USA, 1991.
40. Klingenberg, C.P.; Monteiro, L.R. Distances and Directions in Multidimensional Shape Spaces: Implications for Morphometric Applications. *Soc. Syst. Biol.* **2005**, *54*, 678–688. [CrossRef] [PubMed]
41. Chollet, F. *Deep Learning with Python*; Manning: New York, NY, USA, 2017.
42. Bishop, C. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, UK, USA, 1995.
43. Bishop, C. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
44. Brownlee, J. *Better Deep Learning: Train. Faster, Reduce Overfitting and Make Better Predictions*; Machine Learning Mastery: Melbourne, Australia, 2019.
45. Zhang, T. Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning – ICML '04, Alberta, Canada, 4–8 July 2004. [CrossRef]
46. Hinton, G. Neural Networks for Machine Learning Online Course. Available online: https://www.coursera.org/learn/neural-netoworks/home/welcome (accessed on 28 November 2019).
47. Duchi, J.; Hazan, E.; Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
48. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 5–7 May 2015; arXiv: 1412.6980.
49. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artificial Intelligence, Lauderdale, FL, USA, 22–24 June 2011.
50. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Perfomanca on ImageNet Classification. *arXiv* **2015**, arXiv:1502.01852v1.
52. Klambauer, G.; Unterthiner, T.; Mayr, A. Self-Normalizing Neural Networks. *Adv. Neural Inf. Process. Syst.* **2017**, arXiv:1706.02515v5.
53. Krogh, A.; Hertz, J.A. A simple weight decay can improve generalization. *Adv. Neural Inf. Process. Syst.* **1991**, *4*, 950–957.
54. Srivastava, N. Improving Neural Networks with Dropout. Master's Thesis, University of Toronto, Toronto, ON, Canada, 2013; pp. 12–13.
55. Kuhn, M.; Johnson, K. *Applied Predictive Modelling*; Springer: New York, NY, USA, 2013.
56. Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: Visualizing classifier performance in R. *Bioinform. Apps. Note* **2005**, *21*, 3940–3941. [CrossRef]
57. Fawcett, T. An introduction to ROC analysis. *Pattern Recog. Lett.* **2006**, *27*, 861–874. [CrossRef]

58. Blumenschine, R.J.; Marean, C.W.; Capaldo, S.D. Blind tests of inter-analyst correspondence and accuracy in the identification of cut marks, percussion marks, and carnivore tooth marks on bone surfaces. *J. Arch. Sci.* **1996**, *23*, 493–507. [CrossRef]
59. Pante, M.C.; Muttart, M.V.; Keevil, T.L.; Blumenschine, R.J.; Njau, J.K.; Merritt, S.R. A new high resolution 3D quantitative method for identifying Bone Surface Modifications with implications for the Early Stone Age archaeological record. *J. Hum. Evol.* **2017**, *102*, 1–11. [CrossRef]
60. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep Learning Applications and Challenges in Big Data Analytics. *J. Big Data* **2015**, *2*, 1–21. [CrossRef]
61. Manusco, S. Elliptic Fourier Analysis (EFA) and Artificial Neural Networks (ANNs) for the identification of grapevine (*Vitis vinifera* L.) genotypes. *Vitis* **1999**, *38*, 73–77.
62. Dobigny, G.; Baylac, M.; Denys, C. Geometric morphometrics, neural networks and diagnosis of sibling *Teterillus* species (Rodentia, Gerbillinae). *Biol. J. Linn. Soc.* **2002**, *77*, 319–327. [CrossRef]
63. Baylac, M.; Villemant, C.; Simbolotti, G. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. *Biol. J. Linn. Soc.* **2003**, *80*, 89–98. [CrossRef]
64. Bocxlaer, B.V.; Schultheiβ, R. Comparison of morphometric techniques for shapes with few homologous landmarks based on machine learning approaches to biological discrimination. *Paleobiology* **2010**, *36*, 497–515. [CrossRef]
65. Lorenz, C.; Ferraudo, A.S.; Suesdek, L. Artificial Neural Network applied as a methodology of mosquito species identification. *Acta Tropica.* **2015**, *152*, 165–169. [CrossRef]
66. Soda, K.J.; Slice, D.E.; Naylor, G.J.P. Artificial neural networks and geometric morphometric methods as a means of classification: A case study using teeth from *Carcharhinus* sp. (Carcharinidae). *J. Morphol.* **2016**, *278*, 131–141. [CrossRef]
67. Richter, A.N.; Khohgoftaar, T.M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artif. Intell. Med.* **2018**, *90*, 1–14. [CrossRef]
68. Ekins, S. The Next Era: Deep Learning in pharmaceutical research. *Pharm. Res.* **2016**, *33*, 259–2603. [CrossRef]
69. Nolle, T.; Luettgen, S.; Seeliger, A.; Mühlhäuser, M. Analyzing business process anomalies using autoencoders. *Mach. Learn.* **2018**, *107*, 1875–1893.
70. Ibrahim, M.R.; Haworth, J.; Cheng, T. Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities* **2019**, *96*, 1–13. [CrossRef]
71. Pineda, A.; Saladié, P.; Vergés, J.M.; Huguet, R.; Cáceres, I.; Vallverdú, J. Trampling versus Cut Marks on chemically altered surfaces: An experimental approach and archaeological application at the barranc de la Boella Site (la Canonja, Tarragona, Spain). *J. Arch. Sci.* **2014**, *50*, 84–93. [CrossRef]
72. Gümrükçu, M.; Pante, M.C. Assessing the Effects of Fluvial Abrasion on Bone Surface Modifications using High Resolution 3D Scanning. *J. Arch. Sci. Rep.* **2018**, *21*, 208–221.
73. Bharadwak, S.; Bhatt, H.S.; Vatsa, M.; Sing, R. Domain specific learning for newborn face recognition. *IEEE Trans. Info. Forens. Sec.* **2016**, *11*, 1630–1641. [CrossRef]
74. Keshari, R.; Vatsa, M.; Singh, R. Learning structure and strength of CNN filters for small sample size training. *arXiv* **2018**, arXiv:1803.11405v1.
75. D'Souza, R.N.; Huang, P.Y.; Yeh, F.C. Small data challenge: Structural analysis and optimization of Convolutional Neural Networks with small sample size. *bioRxiv* **2018**. [CrossRef]
76. Feng, S.; Zhou, H.; Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **2019**, *162*, 300–310. [CrossRef]

*Spanish Translation of Title and Abstract*

# Avances en la ciencia de datos para clasificar las marcas de dientes del tipo depresión de carnívoros

Comprender la competencia por los recursos es una cuestión clave para el estudio de la evolución humana. Desde la presencia de los primeros grupos de homininos, los carnívoros han desempeñado un papel fundamental en el ecosistema. Por ello, el conocimiento de la presión trófica existente entre homininos y carnívoros puede proporcionar una importante fuente de información para entender cómo los humanos sobrevivieron e interactuaron con su entorno y, por lo tanto, evolucionaron. Aunque ya existen numerosas técnicas para detectar la actividad de los carnívoros en yacimientos arqueológicos y paleontológicos, muchas de estas técnicas presentan importantes limitaciones. El presente estudio se basa en una serie de técnicas avanzadas de la ciencia de datos para afrontar estos problemas, definiendo métodos para la identificación de los diferentes agentes implicados en el consumo y manipulación de las presas. Para ello, se presenta una amplia muestra de 620 marcas de dientes tipo depresión, generada por carnívoros, incluyendo muestras de osos, hienas, jaguares, leopardos, leones, lobos, zorros, y licaones. A partir del uso de modelos 3D, técnicas de morfometría geométrica, la modelización de datos mediante estadística robusta, y algoritmos de inteligencia artificial, el presente estudio obtiene una precisión de entre el 88% y el 98%, con unas métricas de evaluación equilibradas en todos los conjuntos de datos. Además, cuando estas técnicas avanzadas de la ciencia de datos se combinan con otras fuentes de datos tafonómicos, los resultados demuestran su valía para futuros estudios tafónomicos, sobre todo para la identificación de la acción de carnívoros a partir de las depresiones de dientes que generan.

*Supplementary Information and Links*

**Supplementary Information available from:**
https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-021-89518-4/MediaObjects/41598_2021_89518_MOESM1_ESM.pdf

**Code available from:**
https://github.com/LACourtenay/Carnivore_Tooth_Pit_Classification

scientific reports

OPEN

# Developments in data science solutions for carnivore tooth pit classification

Lloyd A. Courtenay[1]✉, Darío Herranz-Rodrigo[2,3], Diego González-Aguilera[1] & José Yravedra[2,3]

**Competition for resources is a key question in the study of our early human evolution. From the first hominin groups, carnivores have played a fundamental role in the ecosystem. From this perspective, understanding the trophic pressure between hominins and carnivores can provide valuable insights into the context in which humans survived, interacted with their surroundings, and consequently evolved. While numerous techniques already exist for the detection of carnivore activity in archaeological and palaeontological sites, many of these techniques present important limitations. The present study builds on a number of advanced data science techniques to confront these issues, defining methods for the identification of the precise agents involved in carcass consumption and manipulation. For the purpose of this study, a large sample of 620 carnivore tooth pits is presented, including samples from bears, hyenas, jaguars, leopards, lions, wolves, foxes and African wild dogs. Using 3D modelling, geometric morphometrics, robust data modelling, and artificial intelligence algorithms, the present study obtains between 88 and 98% accuracy, with balanced overall evaluation metrics across all datasets. From this perspective, and when combined with other sources of taphonomic evidence, these results show that advanced data science techniques can be considered a valuable addition to the taphonomist's toolkit for the identification of precise carnivore agents via tooth pit morphology.**

Throughout history, humans and carnivores have been documented to have complex relationships[1–4]. From a more traditional perspective, competition for resources is the most documented[4]. Nevertheless, conflict between these taxonomic orders is also well known, especially in the context of dynamic shifts in who plays the role of predator and who plays the role of prey[1,5–9]. Among the many sites of global importance, interactions of these types have been documented across most continents, including notable cases from the Olduvai Gorge (Tanzania)[4,8], Thomas Quarry (Morocco)[9], Schöningen (Germany)[7,10], Zhoukoudian (China)[11], and the classic sites of Makapansgat (South Africa)[1]. Moreover, in more recent periods collaboration between these two orders have also been recorded[2].

From multiple perspectives, carnivore–hominin interactions have thus been a topic of great interest, in both the study of how humans survived and adapted, as well as the contexts in which this occurred. These types of analyses, however, have not been free of debate. In certain case studies, issues of equifinality have led analysts to propose problematic interpretations. The famous long bone fragment from Divje Babe (Slovenia) was originally interpreted as a 43 Ka Middle Palaeolithic flute. Nevertheless, subsequent analyses have discredited these finds and found the perforations to be product of carnivore bite damage[12,13]. Likewise, the sites of Sima de los Huesos (Atapuerca, Spain) and the Dinaledi Rock Chamber (South Africa), have been interpreted as the deliberate anthropic accumulations of human remains[14,15]. Needless to say, not all researchers agree with these conclusions[5,16].

The discipline of taphonomy has frequently been at the forefront of these debates[4]. Taphonomy employs numerous tools for the detection, documentation, and consequent interpretation of carnivore and human activities involved in the formation of a site[2,3]. Nevertheless, diagnostic tools are frequently subjective, thus requiring a search for more empirical and accurate techniques in the identification and interpretation of Bone Surface

[1]Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003 Ávila, Spain. [2]Department of Prehistory, Complutense University, Prof. Aranguren s/n, 28040 Madrid, Spain. [3]C. A. I. Archaeometry and Archaeological Analysis, Complutense University, Professor Aranguren 2/n, 28040 Madrid, Spain. ✉email: ladc1995@gmail.com

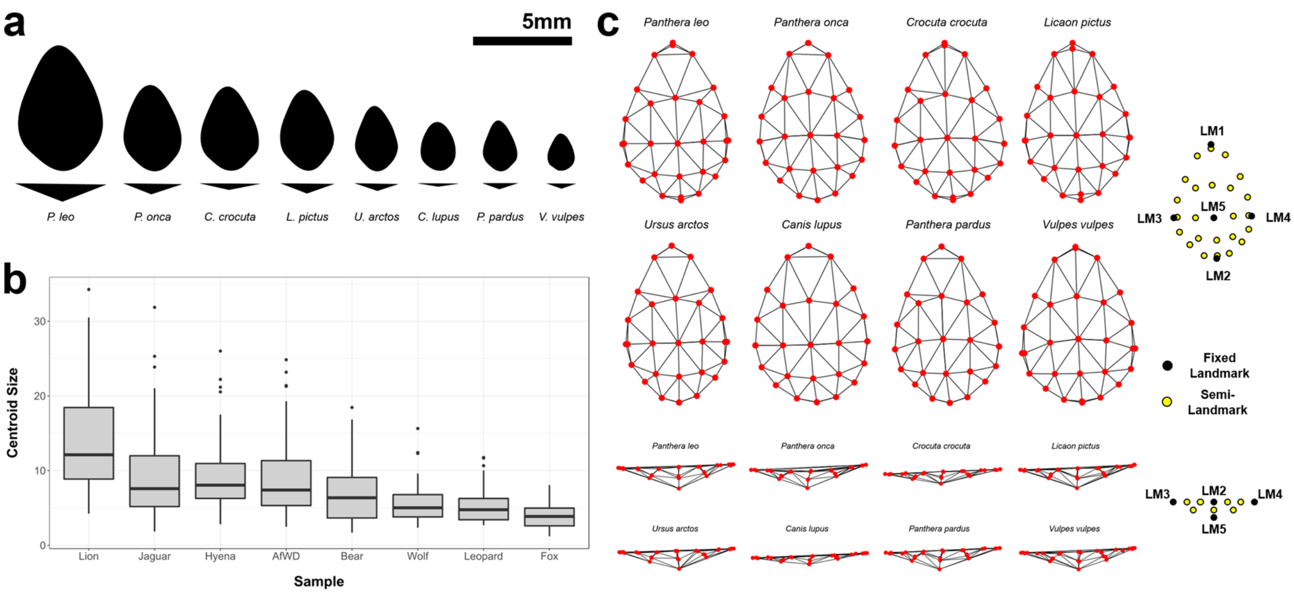


**Figure 1.** Variations in Form and Shape across tooth pits made by different taxa. (**a**) Variations in mean shape-size relationships (top-view used for general morphology and front-view for depth). (**b**) Boxplot diagrams representing centroid size distributions for each species (See Supplementary Table 1). (**c**) Mean landmark configurations for carnivore tooth pits using Delaunay 2.5D Triangulation algorithms for mesh visualisations. AfWD = *Lycaon pictus*. Figures created using the ggplot2 and scikit-learn Python and R libraries.

Modifications (BSM)[17]. This is especially relevant when considering techniques available for discerning of the precise carnivore agencies involved in site formation processes.

Geometric Morphometrics (GM) are a popular multivariate statistical tool for the analysis of morphological variance typically in biological systems[18,19]. Recent years, however, have seen an increase in GM applications outside of anatomy. Applications in taphonomy have yielded impressive results when using GM as a tool for morphological analyses and visualisation. From this perspective, multiple attempts have been made to use GM as a diagnostic tool in carnivore taphonomy[8,20–25]. With the inclusion of Machine Learning (ML) algorithms, data presented by Courtenay et al.[20] present a promising advance for the integration of Artificial Intelligence (AI) and advanced Data Science techniques with GM. Nevertheless, considering the relatively small sample size, these results can also be considered optimistic. Likewise, in a recent study the original landmark model proposed[24] was found to present important margins of error product of landmark quality. These observations infer that analyst experience condition the quality of results[21].

Under this premise, the present study uses an updated version of the landmark model using semi-landmarks[21], and a much larger sample size to expand on the current referential samples available for taphonomic analyses. These efforts aim to provide high quality data that can aid in the understanding of modern carnivore taxa that are frequently found across Eurasia, Africa and the Americas. Samples include three types of felids (*Panthera leo, Panthera onca* & *Panthera pardus*), three types of canids (*Canis lupus, Vulpes vulpes* & *Lycaon pictus*), the spotted hyena (*Crocuta crocuta*), and the brown bear (*Ursus arctos*), that have been frequently subject of study in Pleistocene research[1,26–36]. This larger sample allows us to conclude that > 90% separation of carnivore taxa is still possible, with possibilities for even higher classification rates in the future.

## Results

### Geometric morphometrics.

All samples are described by notable allometric patterns (Squared Residuals = 0.006, F = 4.1, Effect Size = 2.7, $p = 0.005$, Bayes Factor Bound (BFB) = 13.88 against $H_0$), indicating tooth pit size to be an important conditioning factor in morphological variation. This is equally reflected when simply considering Centroid Size values for each of the carnivores (Fig. 1, Table S1), with suggestive to strongly indicative differences detected across most species ($\chi^2 = [5.08, 85.03]$, $p < 0.007$, BFB > 10.59). Exceptions to this include *C. crocuta*, *L. pictus* and *P. onca* when these taxa are compared together ($\chi^2 = [0.14, 1.21]$, $p > 0.27$, BFB < 1.04), as well as *P. pardus* when compared with *C. lupus* ($\chi^2 = 0.42$, $p = 0.51$, BFB = 1.07 against $H_a$).

When considering multivariate morphological tendencies in form, general patterns reveal significant differences throughout comparisons, with each of the taxonomic families being clearly separable ($p \approx 0.001$, BFB ≈ 53.25). While the statistical separation was weakest when comparing Canidae and Ursidae ($p = 0.003$, BFB = 21.11), as well as Ursidae and Hyaenidae ($p = 0.006$, BFB = 11.98), in both these cases differences remain of notable interest ($p < 0.05$). From a similar perspective, species within the families Canidae and Felidae appear easily separable ($p = 0.001$, BFB = 53.25). When describing patterns of variation on a species-specific level, most carnivores present statistical differences ($p \approx 0.001$, BFB ≈ 53.26, Table S2 & S3), nevertheless, exceptions to this can still be found. From this perspective, some degrees of equifinality are therefore still likely to exist when

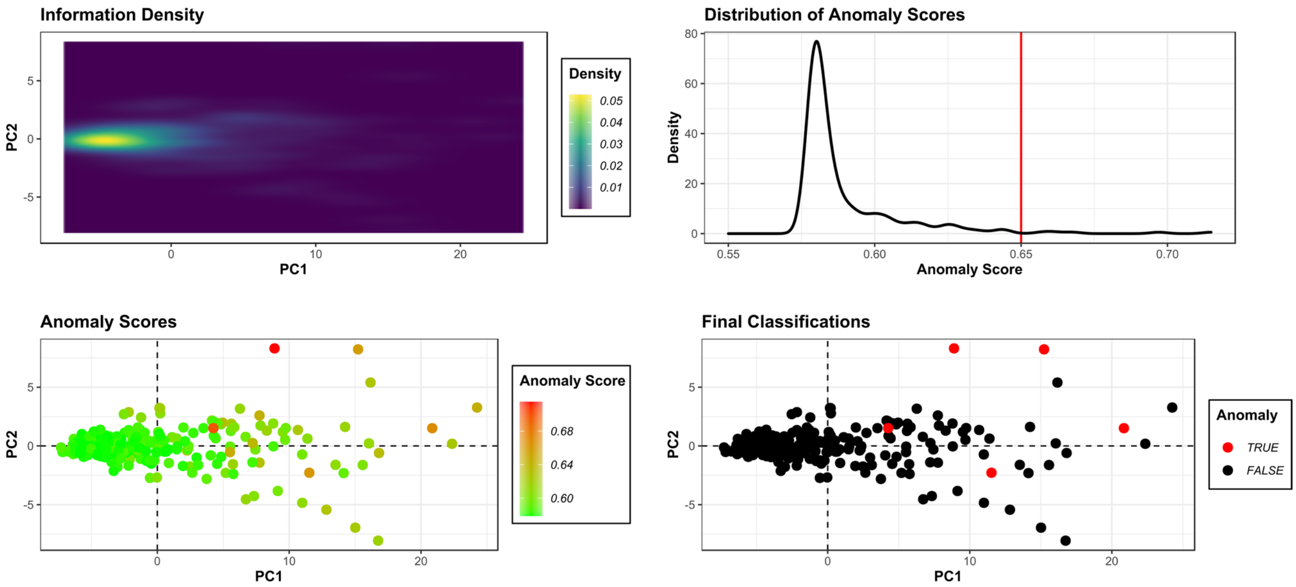


**Figure 2.** Anomaly detection results using Isolation Forests. Top Left Panel: Density of information within Principal Components Analysis. Top Right Panel: Distribution of Anomaly Scores; Vertical red line marks the acceptable threshold. Bottom Left Panel: Scatter plot heat map indicating the anomaly scores for each point. Bottom Right Panel: Final classifications of points as anomalies (True) or not (False). Figure created using the ggplot2 R library.

comparing *L. pictus*, *C. crocuta* and *P. onca* ($p > 0.8$, BFB > 2.06 against $H_a$), as well as when comparing *C. crocuta* and *P. onca* ($p = 0.17$, BFB = 1.22).

Exploring morphological variation through visualisations of mean landmark configurations reveal that the greatest differences appear when considering landmark displacements across the *z*-axis (Fig. 1). From this it can be seen that *C. lupus* tend to leave the most superficial traces, while *P. leo* leave some of the deepest and largest tooth pits of the entire sample. Interestingly, *V. vulpes* and *L. pictus* appear to leave very deep pits in relation to their size. Likewise, when considering variations across a horizontal plane (*x* and *y* axes), slight variations can be seen with some of the canids such as *C. lupus* and *V. vulpes* leaving more circular marks, while felids appear to leave more elongated pits (Fig. 1).

When analysing these central morphological tendencies in accordance with taxonomic groupings, very weak phylogenetic signals are detected, indicating other confounding variables, such as biomechanics, exert a much stronger influence on tooth pit formation than cuspid morphology (Effect Size = − 0.99, $p = 0.81$, BFB = 2.16 against $H_a$. Fig. S1).

## Unsupervised computational learning.

Dimensionality reduction of datasets through Principal Components Analysis (PCA) produced high dimensional, non-homogeneously distributed and noisy datasets on all accounts. General analyses showed PCA in form space to produce a total of 90 Principal Component (PC) Scores, of which the first 6 PC Scores represent over 95% of the total sample variance. Analyses of optimal number of components observed 5 PC scores to be the most representative. Nevertheless high residuals were still noted across a number of these dimensions.

When preparing datasets for further processing, Isolation Forests (IF) proved effective for the elimination of anomalies across all 5 dimensions (Fig. 2). Nevertheless, a relatively high anomalous score threshold was needed for most anomaly detection tasks, considering how species like *P. leo* and *C. lupus* presented very high variability in comparison with other samples. This natural variability consequently produced a global increase of variance across all dimensions, frequently resulting in the adversarial effect of IFs over-classifying entire species as anomalies due to their abnormally large morphological variations. Under this premise, anomaly score distributions were allowed a slight positive tail, with thresholds in the present study defined between 0.625 and 0.700. Using these thresholds, IFs were seen to remove between 3 and 10 pits for each dataset, with the most extreme removal of 10 pits occurring in the European Taxa dataset. Nevertheless, upon inspection of anomaly score distributions (Top right panel; Fig. 2), it can be argued that IFs were still able to preserve the majority of natural variability, only eliminating the most extreme of cases. In light of this, IFs were only seen to remove at most 2.3% of the original sample.

Once datasets had been cleaned, data augmentation proved successful on all accounts with the generation of highly realistic synthetic data by both algorithms. Of the two algorithms tried and tested, Markov Chain Monte Carlo (MCMC, Fig. 3) algorithms appeared the fastest at generating new data with very high equivalency scores (Table 1). Experimentation found MCMCs to produce the most realistic data when sampling from robustly defined gaussian target distributions ($|d| = 0.004$, $p = 1.2e{-}57$, BFB = 2.3e+54), as opposed to the skewed-normal ($|d| = 0.06$, $p = 1.3e{-}05$, BFB = 2515). This was especially evident when considering the skewed-normal had the

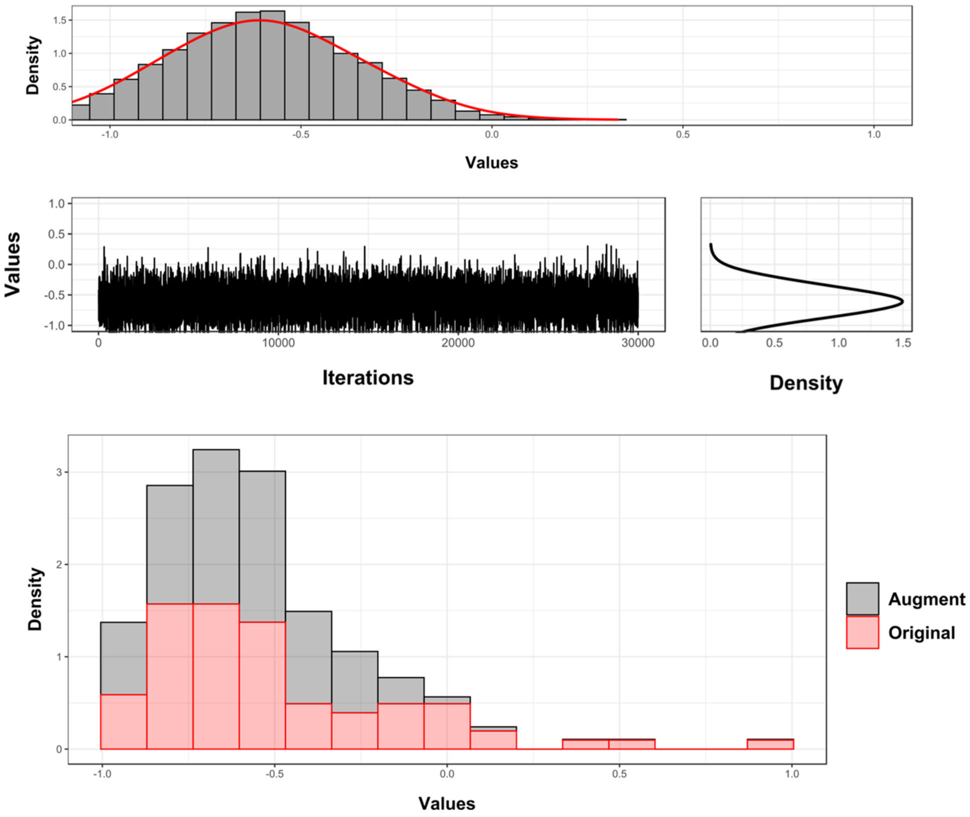


**Figure 3.** Example of trace figures, target density and histograms of the augmented and original datasets as generated using Markov Chain Monte Carlo algorithms. Figure created using the ggplot2 R library.

tendency to exaggerate non-Gaussian elements, which may not be a true reflection of the population distribution (original skew = 0.18, augmented skew = 0.97).

From the perspective of generative neural networks, of the three Generative Adversarial Networks (GANs), Wasserstein Gradient-Penalty loss GANs (WGAN-GP) produced the best results ($|d| = 0.012$, $p = 2.4e{-}13$, BFB = 5.3e+10). Nevertheless, while WGAN-GP proved successful on all datasets, the training of GAN models proved to be computationally expensive, with iterations taking ≈25,000 times longer than MCMC ($\chi^2 = 5.6$, $p = 0.018$, BFB = 5.10).

For final data augmentation tasks both MCMC and WGAN-GP were used, with the best performing algorithm being chosen to augment each dataset prior to supervised training (Tables 1, S4-7).

## Supervised computational learning.

Both supervised models provided high accuracy in the classification of carnivore taxa (Tables 2 & S8-12, Figs. 4, 5 and 6), in most cases producing >90% accuracy (Area Under Curve (AUC) > 0.94, F-Measure > 0.93, $\kappa$ > 0.86). The only exception to this can be found in the case of the Pleistocene European Taxa dataset, which only produced >85% accuracy (AUC ≈ 0.90, F ≈ 0.89, $\kappa \approx 0.85$). Upon analysing the overall performance of each dataset, the greatest results are obtained when differentiating between taxonomic families (Accuracy > 96%, AUC > 0.97, F > 0.97, $\kappa$ > 0.92), as well as the specific species within these families (Table S11 & S12). This can be seen in the cases of the Canidae (Acc. > 97%, AUC > 0.98, F > 0.98, $\kappa$ > 0.95), and the Felidae datasets (Acc. > 96%, AUC > 0.97, F > 0.97, $\kappa$ > 0.95).

When pooling many labels, especially with taxa from different families, overall classification rates tend to drop. Nevertheless, while classification rates may fall below 90% accuracy, miss-classification rates and the frequency of Type I and Type II errors do not rise above 0.2 when considering overall performance (Fig. 4), resulting in very high AUC, Kappa and F scores as well. Under this premise, both Support Vector Machines (SVM) and Neural SVMs (NSVM) can be considered highly efficient classifiers of carnivore tooth marks, yet with greater performance when working with a smaller number of labels. Needless to say, when considering loss values, with the exception of the Pleistocene European dataset, both SVM and NSVM appear to be confident when making new predicitons (Fig. 6).

By considering model performance on individual samples (Tables S8-S12), differentiating between taxa appears to depend on the species being used for comparison. Under this premise, *V. vulpes* (Tables S8) and *P. leo* (Tables S9) appear to be the easiest of the Pleistocene European and African carnivores to identify (SVM Acc. = {95%, 95%}, NSVM Acc. = {94%, 96%}, respectively). On the scale of taxonomic families, *L. pictus* can be considered the easiest canid to identify (SVM Acc. = 98%, NSVM Acc. = 100%), while *P. leo* remains the felid with the highest classification rates (SVM Acc. = 96%, NSVM Acc. = 99%). Each of these observations are especially

| Algorithm | Animal | Measure | PC1 | PC2 | PC3 | PC4 | PC5 | Time (Ms) |
|---|---|---|---|---|---|---|---|---|
| WGAN-GP | *C. crocuta* | \|*d*\| | 0.007 | 0.011 | 0.016 | 0.115 | 0.037 | |
| | | *p* | 4.9e−13 | 2.7e−09 | 6.8e−16 | 1.3e−08 | 8.9e−22 | 1311 |
| | | BFB | 2.6e+10 | 6.9e+06 | 1.5e+13 | 1.6e+06 | 8.5e+18 | |
| | *P. pardus* | \|*d*\| | 0.092 | 0.010 | 0.022 | 0.007 | 0.016 | |
| | | *p* | 3.4e−18 | 3.6e−37 | 2.9e−32 | 1.5e−36 | 1.4e−36 | 1194 |
| | | BFB | 2.7e+15 | 1.2e+34 | 1.7e+29 | 3.0e+33 | 3.2e+33 | |
| | *L. pictus* | \|*d*\| | 0.034 | 0.009 | 0.046 | 0.001 | 0.004 | |
| | | *p* | 1.2e−08 | 1.0e−12 | 1.4e−19 | 4.8e−33 | 4.0e−16 | 1296 |
| | | BFB | 1.7e+06 | 1.3e+10 | 6.1e+16 | 1.0e+30 | 2.6e+13 | |
| | *P. leo* | \|*d*\| | 0.097 | 0.005 | 0.008 | 0.043 | 0.003 | |
| | | *p* | 6.3e−03 | 5.5e−10 | 5.8e−10 | 1.5e−08 | 1.6e−15 | 915 |
| | | BFB | 11.52 | 3.1e+07 | 3.0e+07 | 1.4e+06 | 6.7e+12 | |
| MCMC | *C. crocuta* | \|*d*\| | 0.055 | 0.010 | 0.010 | 0.004 | 0.003 | |
| | | *p* | 5.8e−13 | 4.3e−40 | 8.5e−73 | 6.0e−63 | 3.6e−62 | 0.048 |
| | | BFB | 2.3e+10 | 9.4e+36 | 2.6e+69 | 4.3e+59 | 7.2e+58 | |
| | *P. pardus* | \|*d*\| | 0.004 | 0.007 | 0.004 | 0.003 | 0.007 | |
| | | *p* | 6.9e−29 | 4.8e−70 | 1.5e−104 | 2.8e−75 | 1.0e−85 | 0.048 |
| | | BFB | 8.2e+25 | 4.8e+66 | 1.0e+101 | 7.7e+71 | 1.9e+82 | |
| | *L . pictus* | \|*d*\| | 0.007 | 0.010 | 0.003 | 0.004 | 0.003 | |
| | | *p* | 9.6e−12 | 3.3e−42 | 4.5e−83 | 1.7e−67 | 2.5e−57 | 0.047 |
| | | BFB | 1.5e+09 | 1.2e+39 | 4.3e+79 | 1.4e+64 | 1.1e+54 | |
| | *P. leo* | \|*d*\| | 0.023 | 0.010 | 0.002 | 0.004 | 0.001 | |
| | | *p* | 1.1e−05 | 1.0e−32 | 3.4e−54 | 8.6e−39 | 1.5e−50 | 0.048 |
| | | BFB | 2.9e+03 | 5.0e+29 | 8.8e+50 | 4.9+e35 | 2.1e+47 | |

**Table 1.** Examples of absolute difference (|*d*|), *p*-Values and Bayes Factor Bounds (BFB) obtained when assessing the robust equivalency of synthetic data and real data using Gradient Penalty Wasserstein Loss Generative Adversarial Networks (WGAN-GP) and Markov Chain Monte Carlo (MCMC) Algorithms for data augmentation of the African Taxa dataset. Time values reported represent the number of miliseconds per epoch or iteration of the algorithm.

| Sample | Algorithm | Acc. | Sens. | Spec. | Prec. | Rec. | AUC | F | $\kappa$ | Loss |
|---|---|---|---|---|---|---|---|---|---|---|
| Pleistocene European Taxa | SVM | 0.89 | 0.81 | 0.96 | 1.00 | 0.81 | 0.87 | 0.90 | 0.85 | 0.16 |
| | NSVM | 0.88 | 0.91 | 0.95 | 0.88 | 0.91 | 0.94 | 0.88 | 0.85 | 0.26 |
| African Taxa | SVM | 0.93 | 0.89 | 0.96 | 1.00 | 0.89 | 0.94 | 0.94 | 0.86 | 0.09 |
| | NSVM | 0.93 | 0.93 | 0.98 | 0.93 | 0.93 | 0.97 | 0.93 | 0.91 | 0.10 |
| Taxonomic family | SVM | 0.96 | 0.93 | 0.98 | 1.00 | 0.93 | 0.97 | 0.97 | 0.92 | 0.05 |
| | NSVM | 0.97 | 0.96 | 0.99 | 0.96 | 0.96 | 0.98 | 0.97 | 0.96 | 0.06 |
| Canidae | SVM | 0.97 | 0.97 | 0.98 | 1.00 | 0.97 | 0.98 | 0.98 | 0.95 | 0.05 |
| | NSVM | 0.98 | 0.98 | 0.99 | 0.98 | 0.98 | 0.99 | 0.98 | 0.97 | 0.01 |
| Felidae | SVM | 0.96 | 0.94 | 0.98 | 1.00 | 0.94 | 0.97 | 0.97 | 0.95 | 0.04 |
| | NSVM | 0.97 | 0.97 | 0.99 | 0.97 | 0.97 | 0.98 | 0.97 | 0.96 | 0.01 |

**Table 2.** Overall classification results obtained for all samples using Support Vector Machines (SVM) and Neural Support Vector Machines (NSVM). Reported values include; Accuracy (Acc.), Sensitivity (Sens.), Specificity (Spec.), Precision (Prec.), Recall (Rec.), Area Under Curve (AUC), F-Measure (F), Kappa ($\kappa$) and Loss. All evaluation metrics (with the exception of loss) are recorded as values between 0 and 1, with 1 being the highest obtainable value. Values reported over 0.8 are considered an acceptable threshold for powerful classification models. Loss considers values closer to 0 as the most confident models.

interesting considering these species have been associated with either the largest or the smallest centroid sizes respectively (Fig. 1, Table S1).

While *P. pardus* presents the lowest recorded individual classification rates across all datasets (NSVM Acc. = 0.79, Table S8), this does not have a significant impact on overall model performance (Fig. 4). Even when considering the poorer classification rates presented by *P. pardus*, all algorithms achieve evaluation metrics above

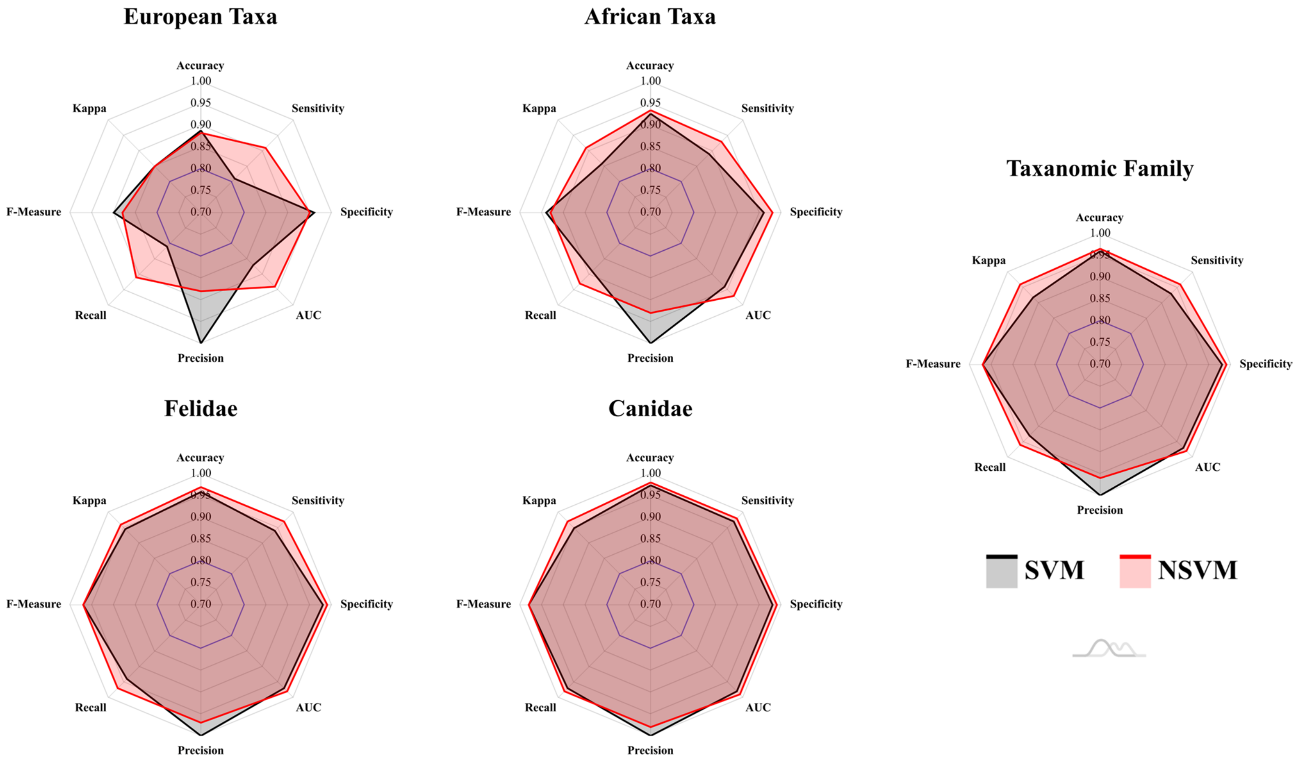


**Figure 4.** Radar plots representing supervised classification results for each of the datasets. Evaluation metrics were calculated on test sets when using both Support Vector Machines (SVM) and Neural Support Vector Machines (NSVM). The blue line marking 0.8 across all graphs represents a standard threshold for the evaluation of good performance for each of the metrics used. Figure created using the amCharts4 JavaScript library.

the acceptable 0.8 threshold. Likewise, a 97% to 99% accuracy has still been obtained when comparing *P. pardus* with other felids (Fig. 5), and a 92% to 93% accuracy when compared with other African species.

Although an element of equifinality is still present, as detected through inconclusive statistical differences in some tooth pit morphologies, both SVM and NSVM are still able to accurately differentiate between *L. pictus* and *C. crocuta* with over 90% success. Nevertheless, algorithm confidence when performing classifications on these species drops, as seen through a large increase in loss values (Table S9). This results in the overall rise in loss and decrease in other performance metrics when these two species are included in a dataset (Table 2, Fig. 6).

Observations comparing SVM with NSVM prove both algorithms to be equally powerful when discerning between carnivore taxa. While NSVM may be seen to have a slight advantage over SVM in some evaluation metrics (Fig. 4), SVM loss rates are generally lower (Fig. 6). Similiarly, NSVM can be seen in some datasets to have very low loss rates for some groups (e.g. Table S10, Canidae loss = 0.001), while especially high loss rates for others (e.g. Table S10, Hyaenidae loss = 0.19). In sum, both SVM and NSVM are valid options for carnivore differentiation, while choice of one or the other must be dependent on the specific case study at hand as well as the analyst's needs.

When observing general performance in model loss (Fig. 6), algorithms produce powerful predictions, with very confident decision boundaries in many cases (Fig. 5).

Finally, when training algorithms without the use of data augmentation (Sup. Appendix 7), it can be seen how the average accuracy slightly drops, with SVM performing 4% worse on non-augmented datasets and NSVM performing 6% worse. While this change is minute, the greatest differences between augmented and non-augmented datasets can be found across loss values, with both SVM and NSVM loosing an average of 10% confidence with each prediction made. As would be expected, algorithms also appear to perform worse on unbalanced datasets, with the Taxonomic Family dataset presenting F-Measure values 0.25 lower, especially in the case of NSVM (Sup. Appendix 7).

## Discussion

In recent years, GM have been applied to a wide array of different applications outside of biology. Among these applications, these tools have shown promising results when applied to the study of BSMs[8,20–25,37]. While subsequent analyses have identified some issues with these techniques for carnivore BSM applications[21], the present study has shown that high quality results are still realistically obtainable (Accuracy > 90 %, AUC > 0.8, $\kappa > 0.8$, Fig. 6). Likewise, the results reported here are supported by considerably larger sample sizes[20,22,24].

Here we have shown how a number of different data science tools can be employed for GM analyses. From one perspective, unsupervised computational learning approaches were able to produce highly realistic augmented

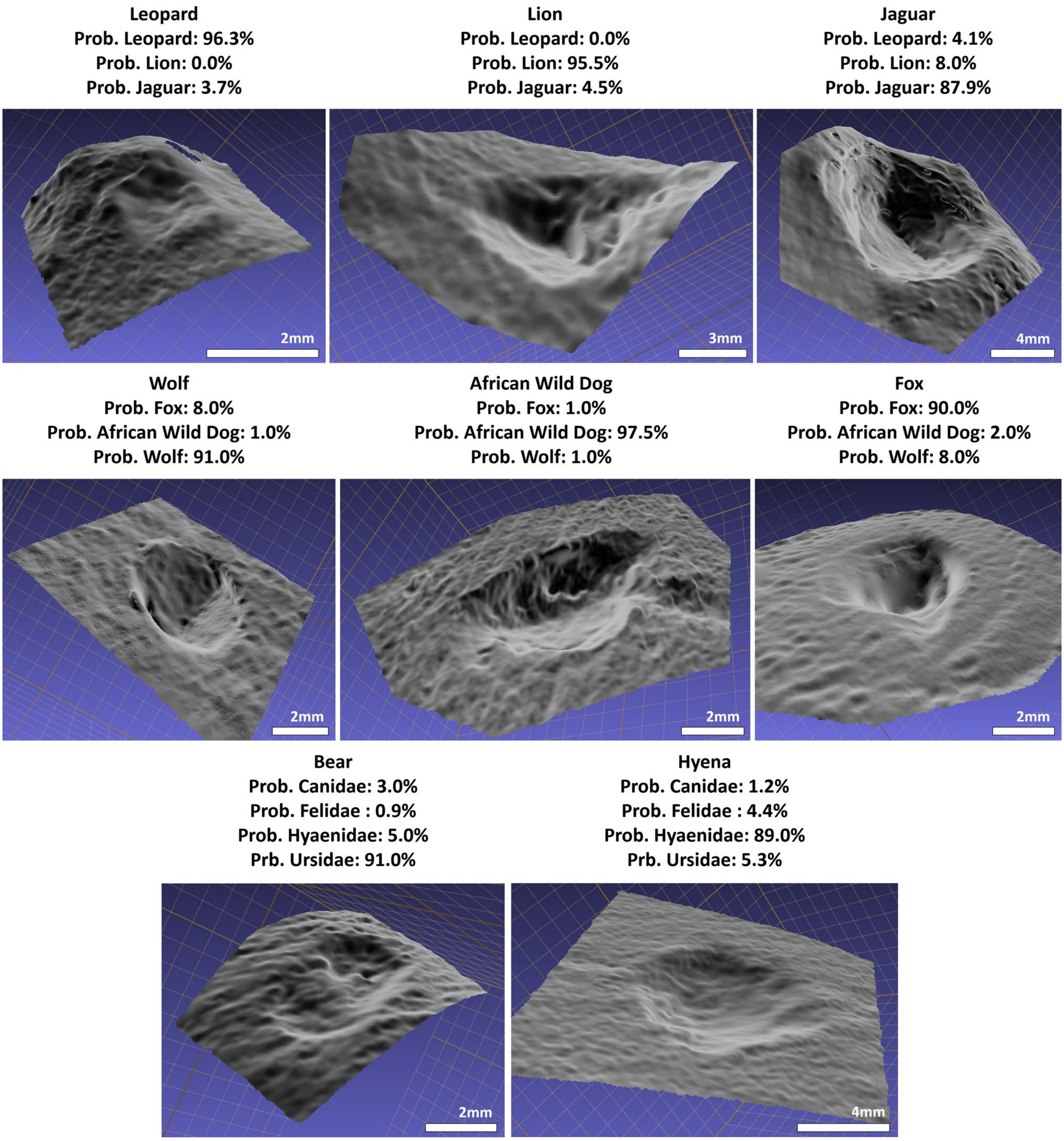


**Figure 5.** Example of tooth pit classifications using Neural Support Vector Machines (NSVM). The select tooth pits were chosen randomly and excluded from the training data so as to avoid bias. NSVMs were then trained on the remaining data and used to classify the present tooth marks, taking note of the algorithms confidence when making predictions. 3D visualisations were created using MeshLab.

datasets, using both neural network based approaches[38–41], as well as Bayesian Inference Engines[42–45]. While the use of Graphics Processing Units (GPUs) are likely to speed up GAN performance, MCMC can still be considered the fastest approach to modelling these datasets with exceptional synthetic-data quality. From a Bayesian perspective, considering how the use of Gaussian distributions is usually seen as a "crude approximation" to the problem solving questions at hand[45], most of the times this also allows models greater generalization capabilities. In addition, this theoretically reduces chance of overfitting supervised models on one particular skewed distribution that may not be a true reflection of the population distribution (Sup. Appendix 4). Moreover, to ensure the present study does not fall into the trap of over-generalising the Gaussian nature of the population, the precise definitions of our target probability distributions were robustly defined[41,46,47].

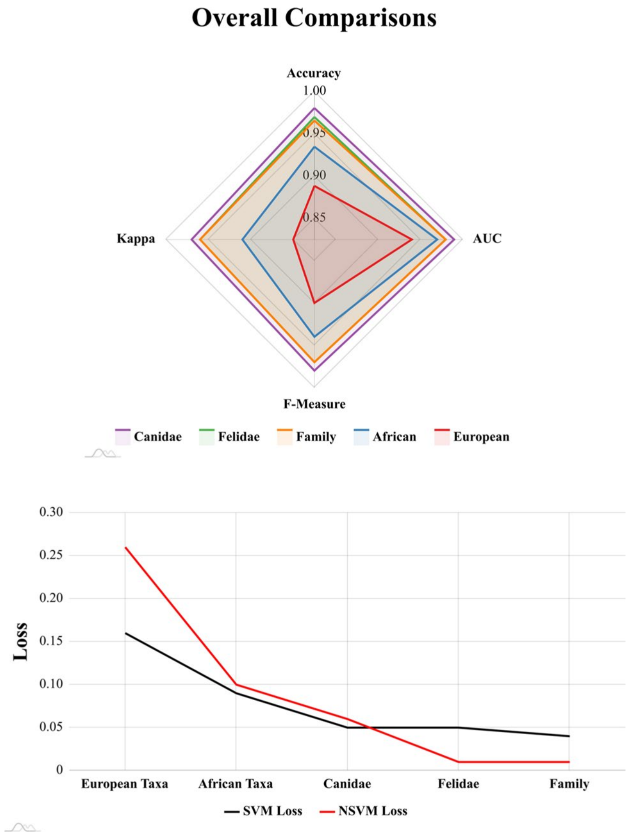


**Figure 6.** Top Panel: Radar plot summarising and comparing performance of the best computational learning models for each of the datasets. Bottom panel: Line graph representing the mean reported loss for both Support Vector Machines (SVM) and Neural Support Vector Machines (NSVM) on each of the datasets. Figure created using the amCharts4 JavaScript library.

From the perspective of supervised learning, the present study reveals the capabilities of computational learning algorithms for the differentiation of carnivore taxa based on the morphology of carnivore tooth pits. Firstly, prior augmentation of each dataset provided both algorithms with enough information to learn from, obtaining above average accuracy when used to classify the original samples. While the present datasets are unable to reach the 100% accuracy reported originally using SVMs[20], this is likely due to the use of bootstrapping in the original study[41]. Here, more robust data augmentation techniques produced completely new synthetic data from which to learn from, providing a more general overview of the target domain. Under this premise, while 100% accuracy was not obtained, our reported >90% can be considered much more reliable. From a similar perspective, while the changes to the original landmark model have shown a reduction in inter-analyst error by 164µm[21], the inclusion of semi-landmark patches has been observed to substantially increase the dimensionality of these GM datasets. In light of this, the new datasets are likely to be harder to model from. Needless to say, considering the increased precision of the landmark model, alongside more robust augmentation techniques, it can be argued that the present results are not only more reliable, but also worth the slight drop in accuracy.

Despite the increase in landmark model complexity, both Radial kernel functions and Laplacian fourier mappings were able to provide SVMs with an appropriate transformed feature space to learn from. Nevertheless, both SVM and NSVM have their advantages and disadvantages. NSVM, for example, can be considered a complex model, with the additional requirement of fine tuning a neural network architecture for feature mappings. NSVM thus presents a large number of parameters and hyperparametrs that have to be adjusted by both the analyst and the model itself. SVM, on the other hand, has the distinct advantage of being easier to tune and train, yet, when using Bayesian algorithms for SVM hyperparameter optimization, training time can increase significantly (Table S13), while NSVMs still perform better on some datasets.

From the perspective of combining supervised and unsupervised learning approaches, the present study can be considered another example of how powerful data augmentation can be for improving classification model performance. Data augmentation is a very popular technique in computer vision, nevertheless, not all of these algorithms are readily applicable to numeric data of this type[48]. Here augmentation has been shown to not only improve the accuracy of most models (Tables 2, S8-S12 & S23-S28, Supplementary Appendix 7), but also improve the generalization capabilities of both SVMs and NSVMs[41,49]. This is mostly seen through the decrease in loss values across taxa (Tables S8-S12 & S23-S28), thus supporting observations made by Courtenay and González-Aguilera[41] when applied to other GM datasets of palaeoanthropological and primatological origin. Similarly, the impact dataset imbalance has on algorithm performance is clearly evident, as seen through great drops in precision, recall and F values (Supplementary Appendix 6, Tables S8-S12 & S23-S28). In light of each of these observations, it can be seen how data augmentation can be a valuable tool for archaeological and palaeontological applications[41], especially in cases where obtaining large sample sizes is difficult.

In the general context of new technologies applied to the field of taphonomy, it can be noted how the inclusion of some carnivore species within the samples have created notable statistical noise. This can be seen through drops in performance when increasing the number of target labels used in classification (e.g. the Pleistocene European Taxa dataset). From this perspective, it is important to point out that highly sophisticated techniques are not the all-encompassing solution that many analysts are looking for. When considering how carnivores can usually be described by the type[2,3,30], ratio[27,34,36,50] and size of bite damage[51–53], alongside the location[54–56] and extent of damage[57], it can be seen how modelling carnivore behaviour should also take into account a wide range of different factors beyond BSMs. While neither one of these techniques can exclusively answer these questions, when combined, taphonomists currently have a very powerful toolkit at their disposal for discerning precise carnivore intervention. From a similar perspective, techniques in remote sensing, photogrammetry and microscopy also provide distinct advantages for the collection of different types of data, supported in many cases by the use of high resolution metric data[37,58–60]. Likewise, the use of computational learning has also proven a useful diagnostic tool for the analysis of fracture plane patterns[61], obtaining high classification rates when applied to archaeological samples as well[62]. From another perspective, computer vision applications can also be considered an interesting development in the field of taphonomy[63]. In sum, and wherever possible, rather than commingling multiple species together into one large group, prior processes of elimination based on general taphonomic evidence should be performed in order to remove the least likely animals to have intervened. Algorithms will then be much easier to train, obtaining state of the art classification rates.

"Occam's Razor" suggests that a more complex model is not always a better one. As seen here, without the use of large kernel machines, SVMs are equally likely to produce high level results. Likewise, while GANs are powerful non-parametric generative models, Bayesian inference is still a valuable tool for distribution modelling, as seen through better and faster performance in some of the reported cases. From this perspective, data science applications and AI can be considered both a very promising field of research, as well as a complex and challenging "pandoras box" of algorithms which analysts must take into consideration before planning a study. Nevertheless, and in combination with multiple other sources of data, advanced data science techniques can be considered a significant contribution to a taphonomist's arsenal.

## Material and methods

**Sample.** A total of 620 carnivore tooth pits were included in the present study. These samples included tooth marks produced by;

- Brown Bears (*Ursus arctos*, Ursidae, 69 pits)
- Spotted Hyenas (*Crocuta crocuta*, Hyaenidae, 86 pits)
- Wolves (*Canis lupus*, Canidae, 80 pits)
- African Wild Dogs (*Lycaon pictus*, Canidae, 89 pits)
- Foxes (*Vulpes vulpes*, Canidae, 53 pits)
- Jaguars (*Panthera onca*, Felidae, 77 pits)
- Leopards (*Panthera pardus*, Felidae, 84 pits)
- Lions (*Panthera leo*, Felidae, 82 pits)

Samples originated from a number of different sources, including animals kept in parks as well as wild animals. Samples obtained from wild animals included those produced by foxes as well as wolves. The only sample containing both wild and captive animals was the wolf sample. Preliminary data from these tooth pits revealed animals in captivity to have highly equivalent tooth pit morphologies to wild animals ($|d| = 0.125$, $p = 9.0e{-}14$, BFB = 1.4e+11), while tooth scores revealed otherwise ($|d| = 0.152$, $p = 0.99$, BFB = 3.7e+01 against $H_a$). Under this premise, and so as to avoid the influence of confounding variables that go beyond the scope of the present study, tooth scores were excluded from the present samples and are under current investigation (*data in preperation*). Nevertheless, other research have shown tooth pits to be more informative than tooth scores when considering morphology[20,23].

When working with tooth mark morphologies, preference is usually given to marks found on long bone diaphyses. This is preferred considering how diaphyses are denser than epiphyses, and are thus more likely to survive during carnivore feeding. Nevertheless, when working with captive or semi-captive animals, controlling the bones that carnivores are fed is not always possible. This is due to the rules and regulations established by the institution where these animals are kept[64]. While this was not an issue for the majority of the animals used within the present study, in the case of *P. pardus*, animals were only fed ribs in articulation with other axial elements. In light of this, a careful evaluation on the effects this may have on the analogy of our samples was performed (Supplementary Appendix 2). These reflections concluded that in order to maintain a plausible analogy with tooth marks produced by other animals on diaphyses, tooth marks could only be used if found on the shaft of bovine ribs closest to the tuburcle, coinciding with the posterior and posterior-lateral portions of the rib, and farthest away from the costochondral junction[65]. This area of the rib corresponds to label RI3 described by Lam et al.[65]. Moreover, with a reported average cortical thickness of 2.3mm (± 0.13 mm) and Bone Mineral Density of $4490 kg/m^3 [213.5, 334.6]$[66], bovine ribs are frequently employed in most bone simulation experiments used in agricultural as well as general surgical sciences. Finally, considering the grease, muscle and fat content of typical domestic bovine individuals[67], alongside the general size of *P. pardus* teeth, it was concluded that the use of rib elements for this sample was the closest possible analogy to the tooth marks collected from other animals.

Carnivores were fed a number of different sized animals, also dependent in most cases on the regulations established by the institution where these animals are kept[64]. Nevertheless, recent research has found statistical similarities between tooth marks found on different animals[25], with the greatest differences occurring between large and small sized animals. Needless to say, considering the typical size of prey some of these carnivores typically consume, this factor was not considered of notable importance for the present study[25] (Supplementary Appendix 1).

For the purpose of comparisons, animals were split into 5 groups according to ecosystem as well as taxonomic family. From an ecological perspective, two datasets were defined; (1) the Pleistocene European Taxa dataset containing *U. arctos*, *V. vulpes*, *C. crocuta*, *P. pardus*, *P. leo* and *C. lupus*; and (2) the African Taxa dataset containing *C. crocuta*, *P. pardus*, *L. pictus* and *P. leo*. When considering taxonomic groupings, animals were separated into 3 groups, including; (1) the Canidae dataset, including *V. vulpes*, *L. pictus* and *C. lupus*; (2) the Felidae dataset, including *P. pardus*, *P. onca* and *P. leo*; and (3) a general Taxonomic Family dataset, including all Canidae in the same group, all Felidae in the same group, followed by Hyaenidae and Ursidae. Some complementary details on each of these carnivores have been included in Supplementary Appendix 1.

All experiments involving carnivores were performed in accordance with the relevant ethical guidelines as set forth by park keepers and general park regulations. No animals were sacrificed specifically for the purpose of these experiments. Likewise, carnivores were not manipulated or handled at any point during the collection of samples. Collection of chewed bones were performed directly by park staff and assisted by one of the authors (JY). The present study followed the guidelines set forth by ARRIVE (https://arriveguidelines.org/) wherever necessary. No licenses or permits were required in order to perform these experiments. Finally, in the case of animals in parks, bone samples were provided by the park according to normal feeding protocols. More details can be consulted in the Extended Samples section of the supplementary files.

**3D modelling and landmark digitisation.** Digital reconstructions of tooth marks were performed using Structured Light Surface Scanning (SLSS)[68]. The equipment used in the present study was the DAVID SLS-2 Structured Light Surface Scanner located in the C.A.I. Archaeometry and Archaeological Analysis lab of the Complutense University of Madrid (Spain). This equipment consists of a DAVID USB CMOS Monochrome 2-Megapixel camera and ACER K11 LED projector. Both the camera and the projector were connected to a portable ASUS X550VX personal laptop (8 GB RAM, Intel® Core$^{\mathrm{TM}}$ i5 6300HQ CPU (2.3 GHz), NVIDIA GTX 950 GPU) via USB and HDMI respectively. The DAVID's Laser Scanner Professional Edition software is stored in a USB Flash Drive. Equipment were calibrated using a 15 mm markerboard, using additional macro lenses attached to both the projector and the camera in order to obtain optimal resolution at this scale. Once calibrated the DAVID SLS-2 produces a point cloud density of up to 1.2 million points which can be exported for further processing via external software.

The landmark configuration used for this study consists of a total of 30 landmarks (LMs)[21]; 5 fixed Type II landmarks[18] and a 5 × 5 patch of semilandmarks[69] (Fig. S2). Of the 5 fixed landmarks, LM1 and LM2 mark the maximal length ($l$) of each pit. For the correct orientation of the pit, LM1 can be considered to be the point along the maximum length furthest away from the perpendicular axis marking the maximum width ($w$). LM2 would therefore be the point closest to said perpendicular axis (see variables $d_1$ and $d_2$ in Fig. S2 for clarification). LM3 and LM4 mark the extremities of the perpendicular axis ($w$) with LM3 being the left-most extremity and LM4 being the right-most extremity. LM5 is the deepest point of the pit. The semilandmark patch is then positioned over the entirety of the pit, so as to capture the internal morphology of the mark.

Landmark collection was performed using the free Landmark Editor software (v.3.0.0.6.) by a single experienced analyst. Inter-analyst experiments prior to landmark collection revealed the landmark model to have a robustly defined human-induced margin of error of 0.14 ± 0.09 mm (Median ± Square Root of the Biweight Midvariance). Detailed explanations as well as an instructional video on how to place both landmarks and semilandmarks can be consulted in the Supplementary Appendix and main text of Courtenay et al.[21].

**Geometric morphometrics.** Once collected, landmarks were formatted as morphologika files and imported into the R free software environment (v.3.5.3, https://www.r-project.org/). Initial processing of these files consisted in the orthogonal tangent projection into a new normalized feature space. This process, frequently referred to as Generalized Procrustes Analysis (GPA), is a valuable tool that allows for the direct comparison of landmark configurations[18,19,70]. GPA utilises different superimposition procedures (translation, rotation and scaling) to quantify minute displacements of individual landmarks in space[71]. This in turn facilitates the comparison of landmark configurations, as well as hypothesis testing, using multivariate statistical analyses. Nevertheless, considering observations made by Courtenay et al.[20,21,25] revealed tooth mark size to be an important conditioning factor in their morphology, prior analyses in allometry were also performed[72]. From this perspective, allometric analyses first considered the calculation of centroid sizes across all individuals; the square root of the sum of squared distances of all landmarks of an object from their centroid[18]. These calculations were then followed by multiple regressions to assess the significance of shape-size relationships. For regression, the logarithm of centroid sizes were used. In cases where shape-size relationships proved significant, final superimposition procedures were performed excluding the scaling step of GPA (*form*).

In addition to these analyses, preliminary tests were performed to check for the strength of phylogenetic signals[73]. This was used as a means of testing whether groups of carnivores produced similar tooth pits to other members of the same taxonomic family. For details on the phylogenies used during these tests, consult Fig. S1 and Supplementary Appendix 1.

For the visualisation of morphological trends and variations, Thin Plate Splines (TPS) and central morphological tendencies were calculated[19,71]. From each of these mean landmark configurations, for ease of pattern visualisation across so many landmarks, final calculations were performed using Delaunay 2.5D Triangulation algorithms[74] creating visual meshes of these configurations in Python (v.3.7.4, https://www.python.org/).

Once normalised, landmark coordinates were processed using dimensionality reduction via Principal Components Analyses (PCA). In order to identify the optimal number of Principal Component Scores (PC Scores) that best represented morphological variance, permutation tests were performed calculating the observed variance explained by each PC with the permuted variance over 50 randomized iterations[75]. Multivariate Analysis of Variance (MANOVA) tests were then performed on these select PCs to assess the significance of multivariate morphological variance among samples.

Geometric Morphometric applications were programmed in the R programming language (Sup. Appendix 8).

**Robust statistics.** While GPA is known to normalize data[76], this does not always hold true. Under this premise, caution must be taken when performing statistical analyses on these datasets. Taking this into consideration, prior to all hypothesis testing, normality tests were also performed. These included Shapiro tests and the inspection of Quantile–Quantile graphs. In cases where normality was detected, univariate hypothesis tests were performed using traditional parametric Analysis of Variance (ANOVA). For multivariate tests, such as MANOVA, calculations were derived using the Hotelling-Lawley test-statistic. When normality was rejected, robust alternatives to each of these tests were chosen. In the case of univariate testing, the Kruskal–Wallis non-parametric rank test was prefered, while for MANOVA calculations, Wilk's Lambda was used.

Finally, in light of some of the recommendations presented by The American Statistical Association (ASA), as debated in Volume 73, Issue Sup1 of *The American Statistician*[77,78], the present study considers $p$-values of $> 2\sigma$ from the mean to indicate only suggestive support for the alternative hypothesis ($H_a$). $p > 0.005$, or where possible, $3\sigma$ was therefore used as a threshold to conclude that $H_a$ is "significant". In addition, Bayes Factor Bound (BFB) values (Eq. 1) have also been included alongside all corresponding $p$-Values[79]. Unless stated otherwise, BFBs are reported as the odds in favor of the alternative hypothesis (BFB:1). More details on BFB, Bayes Factors and the $p > 3\sigma$ threshold have been included in Supplementary Appendix 3. General BFB calibrations in accordance with Benjamin and Berger's Recommendation 0.3[79], as well as False Positive Risk values according to Colquhoun's proposals[80], have also been included in Table S20 of Supplementary Appendix 3.

$$BFB = \frac{1}{-e\;p\;\log(p)} \tag{1}$$

All statistical applications were programmed in the R programming language (Sup. Appendix 8).

**Computational learning.** Computational Learning employed in this study consisted of two main types of algorithm; Unsupervised and Supervised algorithms. The concept of "learning" in AI refers primarily to the creation of algorithms that are able to extract patterns from raw data (i.e. "learn"), based on their "experience" through the construction of mathematical functions[38,81]. The basis of all AI learning activities include the combination of multiple components, including; linear algebra, calculus, probability theory and statistics. From this, algorithms can create complex mathematical functions using many simpler concepts as building blocks[38]. Here we use the term "Computational Learning" to refer to a very large group of sub-disciplines and sub-sub-disciplines within AI. Deep Learning and Machine Learning are terms frequently used (and often debated), however, many more branches and types of learning exist. Under this premise, and so as to avoid complication, the present study has chosen to summarise these algorithms using the term "Computational".

Similar to the concepts of Deep and Machine Learning, many different types of supervision exist. The terms supervised and unsupervised refer to the way raw data is fed into the algorithm. In most literature, data will be referred to via the algebraic symbol $x$, whether this be a vector, scalar or matrix. The objective of algorithms are to find patterns among a group of $x$. In an unsupervised context, $x$ is directly fed into the algorithm without further explanation. Algorithms are then forced to search for patterns that best explain the data. In the case of supervised contexts, $x$ is associated with a label or target usually denominated as $y$. Here the algorithm will try and find the best means of mapping $x$ to $y$. From a statistical perspective, this can be explained as $p\left(y|x\right)$. In sum, unsupervised algorithms are typically used for clustering tasks, dimensionality reduction or anomaly detection, while supervised learning is typically associated with classification tasks or regression.

The workflow used in the present study begins with dimensionality reduction, as explained earlier with the use of PCA. While preliminary experiments were performed using non-linear dimensionality reduction algorithms, such as t-distributed Stochastic Neighbor Embedding (t-SNE)[82] and Uniform Manifold Approximation and Projection (UMAP)[83], PCA was found to be the most consistent across all datasets, a point which should be developed in detailed further research. Once dimensionality reduction had been performed, and prior to any advanced computational modelling, datasets were cleaned using unsupervised Isolation Forests (IFs)[84]. Once anomalies had been removed, data augmentation was performed using two different unsupervised approaches; Generative Adversarial Networks (GANs)[38–41] and Markov Chain Monte Carlo (MCMC) sampling[44]. Data augmentation was performed for two primary reasons; (1) the simulation of larger datasets to ensure supervised algorithms have enough information to train from, and (2) to balance datasets so each sample has the same size. Both MCMCs and GANs were trialed and tested using robust statistics to evaluate quality of augmented data[41]. Once the best model had been determined, each of the datasets were augmented so they had a total sample size of $n = 100$. In the case of the Taxonomic Family dataset, augmentation was performed until all samples had the same size as the largest sample.

Once augmented, samples were used for the training of supervised classification models. Two classification models were tried and tested; Support Vector Machines (SVM)[85] and Neural Support Vector Machines (NSVM)[86,87]. NSVMs are an extension of SVM using Neural Networks (NNs)[38] as feature extractors, in substituting the kernel functions typically used in SVMs. Hyperparameter optimization for both SVMs and NSVMs were performed using Bayesian Optimization Algorithms (BOAs)[88].

Supervised computational applications were performed in both the R and Python programming languages (Sup. Appendix 8). For full details on both unsupervised and supervised computational algorithms, consult the Extended Methods section of the Supplementary Materials.

Evaluation of supervised learning algorithms took into account a wide array of different popular evaluation metrics in machine and deep learning. These included; Accuracy, Sensitivity, Specificity, Precision, Recall, Area Under the receiver operator characteristic Curve (AUC), the F-Measure (also known as the F1 Score), Cohen's Kappa ($\kappa$) statistic, and model Loss. Each of these metrics, with the exception of loss, are calculated using confusion matrices, measuring the ratio of correctly classified individuals (True Positive & True Negative) as well as miss-classified individuals (False Positive & False Negative). For more details see Supplementary Appendix 6.

Accuracy is simply reported as either a decimal [0, 1] or a percentage. Accuracy is a metric often misinterpreted, as explained in Supplementary Appendix 6, and should always be considered in combination with other values, such as Sensitivity or Specificity. Both Sensitivity and Specificity are values reported as decimals [0, 1], and are used to evaluate the proportion of correct classifications and miss-classifications. AUC values are derived from receiver operator characteristic curves, a method used to balance and graphically represent the rate of correctly and incorrectly classified individuals. The closest the curve gets to reaching the top left corner of the graph, the better the classifier, while diagonal lines in the graph represent a random classifier (poor model). In order to quantify the curvature of the graph, the area under the curve can be calculated (AUC), with $AUC = 1$ being a perfect classifier and $AUC = 0.5$ being a random classifier. The $\kappa$ statistic is a measure of observer reliability, usually employed to test the agreement between two systems. When applied to confusion matrix evaluations, $\kappa$ can be used to assess the probability that a model will produce an output $\hat{y}$ that coincides with the real output $y$. $\kappa$ values typically range between [0, 1], with $\kappa = 1$ meaning perfect agreement, $\kappa = 0$ being random agreement, and $\kappa = 0.8$ typically used as a threshold to define a near-perfect or perfect algorithm.

While in the authors' opinion, AUC, Sensitivity and Specificity values are the most reliable evaluation metrics for studies of this type (Supp. Appendix 6), for ease of comparison with other papers or authors who choose to use other metrics, we have also included Precision, Recall and F-Measure values. Precision and Recall values play a similar role to sensitivity and specificity, with recall being equivalent to sensitivity, and precision being the calculation of the number of correct positive predictions made. Precision and Recall, however, differ from their counterparts in being more robust to imbalance in datasets. F-Measures are a combined evaluation of these two measures. For more details consult Supplementary Appendix 6.

Loss metrics were reported using the Mean Squared Error (Eq. 2);

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{2}$$

Loss values are interpreted considering values closest to 0 as an indicator of greater confidence when using the model to make new predictions.

Final evaluation metrics were reported when using algorithms to classify only the original samples, without augmented data. Augmented data was, therefore, solely used for training and validation. Finally, so as to assess the impact data augmentation has on supervised learning algorithms, algorithms were also trained on the raw data. This was performed using 70% of the raw data for training, while the remaining 30% was used as a test set.

## Data availability

All the relevant data and code used for the present study have been made readily available online via the corresponding author's GitHub page: https://github.com/LACourtenay/Carnivore_Tooth_Pit_Classification. Any queries or issues regarding data or code should be directed to L.A. Courtenay (ladc1995@gmail.com).

Received: 16 February 2021; Accepted: 27 April 2021
Published online: 13 May 2021

## References

1. Brain, C. K. *Hunters or the Hunted? An introduction to African cave taphonomy* (University of Chicago Press, 1981).
2. Binford, L. R. *Bones: Ancient Men and Modern Myths* (Academic Press Inc., 1981).
3. Blumenschine, R. Percussion marks, tooth marks and experimental determinations of the timing of hominid and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania. *J. Hum. Evol.* **29**(1), 21–51 (1995).
4. Domínguez-Rodrigo, M., Barba, R. & Egeland, C. P. *Deconstructing Olduvai* (Springer, 2007).
5. Andrews, P. & Fernandez-Jalvo, Y. Surface modifications of the Sima de los Huesos fossil humans. *J. Hum. Evol.* **33**, 191–217 (1997).
6. Cueto, M., Camarós, E., Castaños, P., Ontañón, R. & Arias, P. Under the skin of a lion: unique evidence of Upper Paleolithic exploitation and use of cave lion (*Panthera spelaea*) from the Lower Gallery of La Garma (Spain). *PLoS ONE* **11**(10), e0163591. https://doi.org/10.1371/journal.pone.0163591 (2016).
7. Serangeli, J., Kolfschoten, T. V., Starkovich, B. M. & Conard, N. J. The European saber-tooth cat (*Homotehrium latidens*) found in the "Spear Horizon" at Schöningen (Germany). *J. Hum. Evol.* **89**, 172–180. https://doi.org/10.1016/j.jhevol.2015.08.005 (2015).
8. Aramendi, J. *et al.* Who ate OH80 (Olduvai Gorge, Tanzania)? A geometric morphometric analysis of surface bone modifications of a Paranthropus boisei skeleton. *Quatern. Int.* **517**, 118–130. https://doi.org/10.1016/j.quaint.2019.05.029 (2019).
9. Daujeard, C. *et al.* Plesitocene hominins as a resource for carnivores: a c. 500,000-year-old human femur bearing tooth-marks in North Africa (Thomas Quarry I, Morocco). *PLoS ONE* **11**(4), e0152284. https://doi.org/10.1371/journal.pone.0152284 (2016).
10. Starkovich, B. M. & Conard, N. J. Bone taphonomy of the Schöningen "Spear Horizon South" and its implications for site formation and hominin meat provisioning. *J. Hum. Evol.* **89**, 154–171. https://doi.org/10.1016/j.jhevol.2015.09.015 (2015).
11. Boaz, N. T., Ciochon, R. L., Xu, Q. & Liu, J. Mapping and taphonomic analysis of the Homo erectus loci at Locality 1 Zhoukoudian, China. *J. Hum. Evol.* **46**, 519–549. https://doi.org/10.1016/j.jhevol.2004.01.007 (2004).

12. D'Errico, F., Villa, P., Pinto Llona, A. . C. & Idarraga, R. . R. A middle palaeolithic origin of music? Using cave-bear bone accumulations to assess the Divje Babe I bone "flute". *Antiquity* **72**(275), 65–79. https://doi.org/10.1017/s0003598x00086282 (1998).
13. Diedrich, C. G. "Neanderthal bone flutes": simply products of Ice Age spotted hyena scavenging activities on cave bear cubs in European cave bear dens. *R. Soc. Open Sci.* **2**(4), 140022. https://doi.org/10.1098/rsos.140022 (2015).
14. Arsuaga, J. L. *et al.* Sima de los Huesos (Sierra de Atapuerca, Spain). The site. *J. Hum. Evol.* **2–3**, 109–127. https://doi.org/10.1006/jhev.1997.0132 (1997).
15. Dirks, P. N. *et al.* Geological and taphonomic context from the new hominin species Homo naledi from the Dinaledi Chamber, South Africa. *eLife* **4**, e09561. https://doi.org/10.7554/eLife.09561 (2015).
16. Egeland, C. P., Domínguez-Rodrigo, M., Pickering, T. R., Menter, C. G. & Heaton, J. L. Hominin skeletal part abundances and claims of deliberate disposal of corpses in the Middle Pleistocene. *Proc. Natl. Acad. Sci.* **115**(18), 4601–4606. https://doi.org/10.1073/pnas.1718678115 (2018).
17. Domínguez-Rodrigo, M. *et al.* Use and abuse of cut mark analyses: the Rorsach effect. *J. Archaeol. Sci.* **86**, 14–23. https://doi.org/10.1016/j.jas.2017.08.001 (2017).
18. Dryden, I. & Mardia, K. *Statistical Shape Analysis* (Wiley, 1998).
19. Bookstein, F. L. *Morphometric Tools for Landmark Data* (Cambridge University Press, 1991).
20. Courtenay, L. A. *et al.* Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **522**, 28–29. https://doi.org/10.1016/j.palaeo.2019.03.007 (2019).
21. Courtenay, L. A. *et al.* Obtaining new resolutions in carnivore tooth pit morphological analyses: a methodological update for digital taphonomy. *PLoS ONE* **15**(10), e0240328. https://doi.org/10.1371/journal.pone.0240328 (2020).
22. Yravedra, J. *et al.* The use of micro-photogrammetry and geometric morphometrics for identifying carnivore agency in bone assemblages. *J. Archaeol. Sci. Rep.* **14**, 106–115. https://doi.org/10.1016/j.jasrep.2017.05.043 (2017).
23. Yravedra, J., Maté-González, M. Á., Courtenay, L. A., González-Aguilera, D. & Fernández-Fernández, M. The use of canid tooth marks on bone for the identification of livestock predation. *Sci. Rep.*https://doi.org/10.1038/s41598-019-52807-0 *(2019).*
24. Aramendi, J. *et al.* Discerning carnivore agency through the three-dimensional study of tooth pits: Revisiting crocodile feeding behaviour at FLK-Zinj and FLK NN3 (Olduvai Gorge, Tanzania). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **488**, 93–102. https://doi.org/10.1016/j.palaeo.2017.05.021 (2017).
25. Courtenay, L. A. *et al.* The effects of prey size on carnivore tooth mark morphologies on bone; the case study of Canis lupus signatus. *Hist. Biol.*https://doi.org/10.1080/08912963.2020.1827239 *(2020).*
26. Marean, C. W. & Kim, S. Y. Mousterian large-mammal remains from Kobeh Cave. *Curr. Anthropol.* **39**, S79–S113. https://doi.org/10.1086/204691 (1998).
27. Arriaza, M. C., Domínguez-Rodrigo, M., Yravedra, J. & Baquedano, E. Lions as bone accumulators? Palaeontological and ecological implications of a modern bone assemblage from Olduvai Gorge. *PLoS ONE* **11**(5), e0153797. https://doi.org/10.1371/journal.pone.0153797 (2016).
28. Gidna, A. O., Kusui, B., Mabulla, A., Musiba, C. & Domínguez-Rodrigo, M. An ecological neo-taphonomic study of carcass consumption by lions in Tarangire National Park (Tanzania) and its relevance for human evolutionary biology. *Quatern. Int.* **322–323**, 167–180. https://doi.org/10.1016/j.quaint.2013.08.059 (2014).
29. Pickering, T. R., Heaton, J. L., Zwodeski, S. E. & Kuman, K. Taphonomy of bones from baboons killed and eaten by wild leaopards in Mapungubwe National Park, South Africa. *J. Taphon.* **9**(2), 117–159 (2011).
30. Haynes, G. A guide for differentiating mammalian carnivore taxa responsible for gnaw damage to herbivore limb bones. *Paleobiology* **9**(2), 164–172 (1983).
31. Yravedra, J., Lagos, L. & Bárcena, F. A taphonomic study of wild wolf Canis lupus modifications of horse bones in Northwestern Spain. *J. Taphon.* **9**(1), 37–65 (2011).
32. Yravedra, J., Andrés, M. & Domínguez-Rodrigo, M. A taphonomic study of the African wild dog (*Lycaon pictus*). *Archaeol. Anthropol. Sci.* **6**, 113–124. https://doi.org/10.1007/s12520-013-0164-1 (2014).
33. Yravedra, J., Andrés, M., Fosse, P. & Besson, J. P. Taphonomic analysis of small ungulates modified by fox (*Vulpes vulpes*) in Southwestern Europe. *J. Taphom.* **12**(1), 37–67 (2014).
34. Rodríguez-Alba, J. J., Linares-Matás, G. & Yravedra, J. First assessments of the taphonomic behaviour of jaguar (*Panthera onca*). *Quatern. Int.* **517**, 88–96. https://doi.org/10.1016/j.quaint.2019.05.004 (2019).
35. Saladié, P., Huguet, R., Díez, C., Rodríguez-Hidalgo, A. & Carbonell, E. Taphonomic modifications produced by modern brown bears (*Ursus arctos*). *Int. J. Osteoarchaeol.* **23**(1), 13–33. https://doi.org/10.1002/oa.1237 (2013).
36. Gidna, A., Yravedra, J. & Domínguez-Rodrigo, M. A cautionary note on the use of captive carnivores to model wild predator behavior: a comparison of bone modification patterns on long bones by captive and wild lions. *J. Archaeol. Sci.* **40**, 1903–1910 (2013).
37. Courtenay, L. A., Huguet, R., González-Aguilera, D. & Yravedra, J. A hybrid geometric morphometric deep learning approach for cut and trampling mark classification. *Appl. Sci.*https://doi.org/10.3390/app10010150 *(2020).*
38. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
39. Goodfellow, I. *et al.* Generative adversarial nets. In *Proc. Int. Conf. Neur. Inf. Process. Syst.* 2672–2680. arXiv:1406.2661v1 (2014).
40. Lucic, M., Kurasch, K., Michalski, M., Bousquet, O. & Gelly, S. Are GANs created equal? A large scale study. In *Proc. Int. Conf. Neur. Inf. Process. Syst.* 698–707. arXiv:1406.2661v1 (2018).
41. Courtenay, L. A. & González-Aguilera, D. Geometric morphometric data augmentation using generative computational learning algorithms. *Appl. Sci.*https://doi.org/10.3390/app10249133 *(2020).*
42. Metropolis, N., Rosenbluth, A., Teller, A. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
43. Hastings, W. Monte Carlo sampling methods using Markov chains and their application. *Biometrika* **57**, 97–109 (1970).
44. Gamerman, D. & Lopes, H. F. *Markov Chain Monte Carlo* (Chapman & Hall, 2006).
45. Martin, O. *Bayesian Analysis with Python* (Packt, 2018).
46. Höhle, J. & Höhle, M. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS J. Photogram. Remote Sens.* **64**, 398–406 (2009).
47. Rodríguez-Martín, M., Rodríguez-Gonzálvez, P., Ruiz de Oña Crespo, E. & González-Aguilera, D. Validation of portable mobile mapping system for inspection tasks in thermal and fluid-mechanical facilities. *Remote Sens.* **11**(19), 2205. https://doi.org/10.3390/rs11192205 (2019).
48. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 60. https://doi.org/10.1186/s40537-019-0197-0 (2019).
49. Such, F. P., Rawal, A., Lehman, J., Stanley, K. O. & Clune, J. Generative teaching networks: accelerating neural architecture search by learning to generate synthetic training data. *Uber AI Labs.* arXiv:1912.07768v1 (2019).
50. Domínguez-Rodrigo, M., Gidna, A. O., Yravedra, J. & Musiba, C. A comparative neo-taphonomic study of felids, hyaenids and canids: an analogical framework based on long bone modification patterns. *J. Taphon.* **10**(3), 147–164 (2012).
51. Andrés, M., Gidna, A. O., Yravedra, J. & Domínguez-Rodrigo, M. A study of dimensional differences of tooth marks (pits and scores) on bones modified by small and large carnivores. *Archaeol. Anthropol. Sci.* **4**(3), 209–219. https://doi.org/10.1007/s12520-012-0093-4 (2012).

52. Domínguez-Rodrigo, M. & Piqueras, A. The use of tooth pits to identify carnivore taxa in tooth-marked archaeofaunas and their relevance to reconstruct hominid carcass processing behaviours. *J. Archaeol. Sci.* **30**(11), 1385–1391. https://doi.org/10.1016/S0305-4403(03)00027-X (2003).
53. Selvaggio, M. M. & Wilder, J. Identifying the involvement of multiple carnivore taxa with archaeological bone assemblages. *J. Archaeol. Sci.* **28**, 465–470. https://doi.org/10.1006/jasc.2000.0557 (2001).
54. Parkinson, J., Plummer, T. & Hartstone-Rose, A. Characterizing felid tooth marking and gross bone damage patterns using GIS image analysis: an experimental feeding study with large felids. *J. Hum. Evol.* **80**, 114–134. https://doi.org/10.1016/j.jhevol.2014.10.011 (2015).
55. Pobiner, B., Dumouchel, L. & Parkinson, J. A new semi-quantitative method for coding carnivore chewing damage with an application to modern African lion-damaged bones. *Palaios* **35**(7), 302–315. https://doi.org/10.2110/palo.2019.095 (2020).
56. Domínguez-Rodrigo, M. *et al.* A 3D taphonomic model of long bone modification by lions in medium-sized ungulate carcasses. *Sci. Rep.* **11**, 4944. https://doi.org/10.1038/s41598-021-84246-1 (2021).
57. Domínguez-Rodrigo, M. *et al.* A new methodological approach to the taphonomic study of paleontological and archaeological faunal assemblages: a preliminary case study from Olduvai Gorge (Tanzania). *J. Archaeol. Sci.* **59**, 35–53. https://doi.org/10.1016/j.jas.2015.04.007 (2015).
58. Pante, M. *et al.* A new high-resolution 3-D quantitative method for identifying bone surface modifications with implications for the Early Stone Age archaeological record. *J. Hum. Evol.* **102**, 1–11. https://doi.org/10.1016/j.jhevol.2016.10.002 (2017).
59. Bello, S. M. & Soligo, C. A new method for the quantitative analysis of cutmark micromorphology. *J. Archaeol. Sci.* **35**(6), 1542–1552 (2008).
60. Duches, R. *et al.* Experimental and archaeological data for the identification of projectile impact marks on small-sized mammals. *Sci. Rep.* **10**(1), 9092. https://doi.org/10.1038/s41598-020-66044-3 (2020).
61. Moclán, A., Domínguez-Rodrigo, M. & Yravedra, J. Classifying agency in bone breakage: an experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using machine learning (ML) algorithms. *Archaeol. Anthropol. Sci.* **11**, 4663–4680. https://doi.org/10.1007/s12520-019-00815-6 (2019).
62. Moclán, A. *et al.* Identifying the bone-breaker at the Navalmaíllo Rock Shelter (Pinilla del Valle, Madrid) using machine learning algorithms. *Archaeol. Anthropol. Sci.* **12**(2), 1–17. https://doi.org/10.1007/s12520-020-01017-1 (2020).
63. Jiménez-García, B., Abellán, N., Baquedano, E., Cifuentes-Alcobendas, G. & Domínguez-Rodrigo, M. Corrigendum to "deep learning improves taphonomic resolution: high accuracy in differentiating tooth marks made by lions and jaguars". *J. R. Soc. Interface* **17**, 20200782. https://doi.org/10.1098/rsif.2020.0782 (2020).
64. Fidgett, A. L. & Plowman, A. Nutrition and diet evaluation. In Bishop, J., Hosey, G. & Plowman, A. (eds.) *Handbook of Zoo & Aquarium Research*, 154–175 (BIAZA, 2013).
65. Lam, Y. M., Chen, X. & Pearson, O. M. Intertaxonomic variability in patterns of bone density and the differential representation of Bovid, Cervid and Equid elements in the archaeological record. *Am. Antiq.* **64**, 343–362 (1999).
66. Szalma, J. *et al.* The influence of the chosen in vitro bone simulation model on intraosseous temperatures and drilling times. *Sci. Rep.* **9**, 11817. https://doi.org/10.1038/s41598-019-48416-6 (2019).
67. Johnson, E. R. & Chant, D. C. Use of carcass density for determining carcass composition in beef cattle. *N. Zeal. J. Agric. Res.* **41**(3), 325–333. https://doi.org/10.1080/00288233.1998.9513317 (1998).
68. Maté-González, M. Á., Aramendi, J., Yravedra, J. & González-Aguilera, D. Statistical comparison between low-cost methods for 3D characterization of cut-marks on bones. *Remote Sens.* **9**(9), 873. https://doi.org/10.3390/rs9090873 (2017).
69. Gunz, P., Mitteroecker, P. & Bookstein, F. L. Semilandmarks in three dimensions. In *Modern Morphometrics in Physical Anthropology* (ed. Slice, D. E.) 73–98 (Plenum Publishers, 2005).
70. Klingenberg, C. & Monteiro, L. Distances and directions in multidimensional shape spaces: implications for morphometric applications. *Soc. Syst. Biol.* **54**, 678–688. https://doi.org/10.1080/10635150590947258 (2005).
71. Bookstein, F. Principal warps: thin plate spline and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(6), 567–585. https://doi.org/10.1109/34.24792 (1989).
72. Adams, D. C., Rohlf, F. J. & Slice, D. E. A field comes of age: geometric morphometrics in the 21st century. *Hystrix* **24**(1), 7–14. https://doi.org/10.4404/hystrix-24.1-6283 (2013).
73. Klingenberg, C. P. & Gidaszewski, N. A. Testing and quantifying phylogenetic signals and homoplasy in morphometric data. *Syst. Biol.* **59**(3), 245–261. https://doi.org/10.1093/sysbio/syp106 (2010).
74. Delaunay, B. Sur la sphère vide. *Bull. l'Acad. Sci. l'URSS Classe des Sci. Math. Nat.* **6**, 793–800 (1934).
75. Viñuela, A. *et al.* Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat. Commun.* **11**, 4912. https://doi.org/10.1038/s41467-020-18581-8 (2020).
76. Diaconsis, P. & Freedman, D. Asymptotics of graphical projection of pursuit. *Ann. Stat.* **12**, 798–815 (1984).
77. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a world beyond "$p< 0.05$". *Am. Stat.* **73**(Sup1), 1–19 (2019).
78. Wasserstein, R. L. & Lazar, N. A. The ASA statement on p-values: context, process, and purpose. *Am. Stat.* **70**(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108 (2016).
79. Benjamin, D. J. & Berger, J. O. Three recommendations for improving the use of p-values. *Am. Stat.* **73**(Sup1), 186–191. https://doi.org/10.1080/0031305.2018.1543135 (2019).
80. Colquhoun, D. The false positive risk: a proposal concerning what to do about p-values. *Am. Stat.* **73**(Sup1), 192–201. https://doi.org/10.1080/00031305.2018.1529622 (2019).
81. Bishop, C. *Pattern Recognition and Machine Learning* (Springer, 2006).
82. Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems.* **857–864** (2003).
83. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.*https://doi.org/10.21105/joss.00861 *(2018).*
84. Liu, F. T., Ting, K. M. & Zhou, Z. H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17 (2008).
85. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn* **20**, 273–297. https://doi.org/10.1007/BF00994018 (1995).
86. Wiering, M. A. *et al.* The neural support vector machine. In *The 25th Benelux Artificial Intelligence Conference*, 257–254 (2013).
87. Rahimi, A. & Recht, B. Random features for large-scale kernel machines. *Proc. Int. Conf. Neural Inf. Process. Syst.* **20**, 1–8. https://doi.org/10.5555/2981562.2981710 (2007).
88. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. *Proc. Int. Conf. Neural Inf. Process. Syst.* **24**, 2546–2554. https://doi.org/10.5555/2986459.2986743 (2011).

## Acknowledgements

We would like to thank the C.A.I. Archaeometry and Archaeological Analysis group from the Complutense University, for their support and access to facilities. We would also like to extend our gratitude to the TIDOP research group of the University of Salamanca, for their comments and support. Among the researchers of TIDOP, we are particularly grateful for the useful comments and suggestions made by Roberto García, Susana del Pozo, and

Susana Lagüela. L.A.C. is also very grateful to Nikolay Oskolkov for his ideas and suggestions on dimensionality reduction techniques. L.A.C. would also like to thank Rosa Huguet for her support and advice, and Guillermo Blanco for his help and suggestions. Finally we would like to thank two anonymous reviewers for their very helpful suggestions. With regards to samples, we cannot forget the help provided by Sergio Ripoll in collecting fox samples in Ayllón; Santiago Borragán for providing access to lions, bears, jaguars, African wild dogs and wolves in Cabárceno; Isabel Barja for her great help in obtaining wolf samples from Villardeciervos and Flechas in the Sierra de la Culebra; and finally Jesus Recuero and Antonio Garrucho for providing access to the leopards in the Biopark of Fuengirola. Finally we thank Jordan Courtenay for her help with some of the English editing.

## Author contributions

L.A.C.: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing - Original draft, reviewing and editing. D.H.R.: Formal analysis, investigation, data curation, writing - Original draft, reviewing and editing. D.G.A.: Resources, supervision, project administration, funding acquisition, writing - reviewing and editing. J.Y.: Conceptualization, formal analysis, investigation, resources (bone samples), supervision, project administration, writing - original draft, reviewing and editing. All authors have reviewed and aproved this manuscript for publication.

## Funding

L.A.C. is supported by the Spanish Ministry of Science, Innovation and Universities with a FPI Predoctoral Grant (Ref. PRE2019-089411) associated to project RTI2018-099850-B-I00 and the University of Salamanca. D.H.R. is supported by the Ministry of Science, Innovation and Universities, as part of the i+D+I Garantía Juvenil programme (Ref. PEJ2019-005420-A). The present research was additionally supported by the Palarq foundation (Ref. PR2004_19/01), under project "Dilucidando la acción de carnívoros en los yacimientos del Pleistoceno Inferior Ibérico del Ponton de la Oliva (Madrid), y de Fuente Nueva III y Venta Micena 3 y 4 (Granada)", 2019-2020 call for funding. Funders had no role in study design, data collection and analysis, decision to publish or preperation of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-89518-4.

**Correspondence** and requests for materials should be addressed to L.A.C.

*Publications: Section 4*

# Transdisciplinary Applications in Science

## *Spanish Translation of Title and Abstract*

# La competencia de los carnívoros para los recursos animales en el yacimiento del Pleistoceno Temprano de Barranco León (1,46 Ma, Orce, España)

Barranco León (Orce, Guadix Baza, España) es uno de los yacimientos con mayor evidencia de actividad humana en el Pleistoceno Inferior del suroeste de Europa. En este yacimiento se han encontrado restos humanos asociados a fauna y artefactos líticos, vinculados a la presencia de marcas de corte y percusión antropogénicas en los restos osteológicos. Sin embargo, aunque este yacimiento es un claro ejemplo del acceso de los primeros homíninos europeos a los cadáveres, las acumulaciones se han identificado como un palimpsesto, en los que múltiples agentes, incluidos los carnívoros, interactuaron y desempeñaron un papel en la modificación de los procesos de formación del yacimiento. Desde esta perspectiva, la interpretación y estudio del yacimiento de Barranco León es de gran dificultad. Tradicionalmente, las interpretaciones han presentado a Barranco León como una zona en la que tanto los homíninos, como la hiena gigante (*Pachycrocuta brevirostris*), competían por el acceso a los cadáveres dejados por félidos machairodontinos, como el *Homotehrium latidens*. Sin embargo, como se presentará en este estudio, la complejidad y la presión trófica de Barranco León es mucho mayor de lo que se había planteado en un principio. Este estudio presenta un detallado análisis tafonómico de las actividades de los carnívoros en el nivel D1 del conjunto de Barrano León. Mediante la modelización 3D, la morfometría geométrica, y el aprendizaje computacional, proporcionamos nuevos conocimientos sobre las marcas de dientes tipo depresión observadas en los materiales faunísticos. Aquí mostramos que mientras la *Pachycrocuta* y *Homotherium* fueron agentes activos en la formación del yacimiento, otros carnívoros como *Ursus etruscus*, *Xenocyon (Lycaon) lycaonoides*, y, en particular, *Canis mosbachensis*, son también agentes importantes a tener en cuenta a la hora de investigar la región de Guadix Baza.

## *Supplementary Information and Links*

**Data available from:**

https://doi.org/10.1016/j.quascirev.2022.107912

Quaternary Science Reviews 300 (2023) 107912

Contents lists available at ScienceDirect

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

# Deciphering carnivoran competition for animal resources at the 1.46 Ma early Pleistocene site of Barranco León (Orce, Granada, Spain)

Lloyd A. Courtenay [a, b, *], José Yravedra [b, c, d, e], Darío Herranz-Rodrigo [b, c], Juan José Rodríguez-Alba [b], Alexia Serrano-Ramos [f], Verónica Estaca-Gómez [b], Diego González-Aguilera [a], José Antonio Solano [f, g], Juan Manuel Jiménez-Arenas [f, g, h]

[a] *Department of Cartographic and Land Engineering, Higher Polytechnic School of Avila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain*
[b] *Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren S/n, 28040, Madrid, Spain*
[c] *C.A.I. Archaeometry and Archaeological Analysis, Complutense University, Prof. Aranguren S/n, 28040, Madrid, Spain*
[d] *Grupo de Investigación Ecosistemas Cuaternarios. Complutense University, Prof. Aranguren S/n, 28040, Madrid, Spain*
[e] *Grupo de Investigación Arqueología Prehistórica. Complutense University, Prof. Aranguren S/n, 28040, Madrid, Spain*
[f] *Department of Prehistory and Archaeology, University of Granada, Campus Universitario de Cartuja, 18071, Granada, Spain*
[g] *Museum Primeros Pobladores de Europa 'Josep Gibert', Cam. San Simon, 18858, Orce, Granada, Spain*
[h] *Institute of Peace and Conflict Research, University of Granada, C/ Rector López Argüeta S/n, 18001, Granada, Spain*

ARTICLE INFO

*Article history:*
Received 12 June 2022
Received in revised form
9 November 2022
Accepted 4 December 2022
Available online xxx

Handling Editor: Danielle Schreve

*Keywords:*
*Canis mosbachensis*
Tooth marks
Trophic pressure
Geometric morphometrics
3D modelling
Computational learning
Archaeological data science
Digital taphonomy

ABSTRACT

Barranco León (Orce, Guadix Baza, Spain) is one of the sites with the oldest evidence of human activity in south-western Europe. This site has yielded human remains in association with both fauna and lithic artefacts, linked through the presence of anthropogenic cut and percussion marks. Nevertheless, while this site is a clear example of early hominin access to carcasses, the accumulations have been identified as a palimpsest, where multiple agents including carnivorans played a role in modifying and interacting in site formation processes. From this perspective, the interpretation and study of the Barranco León site is of great difficulty. Traditionally, interpretations have presented Barranco León as an area where hominins as well as the giant hyena, Pachycrocuta brevirostris, competed for access to carcasses left by machairodontine felids, such as the saber-toothed Homotherium latidens. Nevertheless, as will be presented in this study, the complexity and trophic pressure of Barranco León is much more complicated than originally hypothesized. This study presents a detailed taphonomic analysis of carnivoran activities in the level D1 of the Barranco León assemblage. 3D modelling, geometric morphometrics, and computational learning are used to provide new insights into the tooth pits observed on faunal materials. Here we show that Canis mosbachensis plays a pivotal role in the formation of the site, followed by Pachycrocuta, Homotherium, Ursus etruscus, and Xenocyon (Lycaon) lycaonoides. From this perspective, it can be seen that while Pachycrocuta and Homotherium were active agents in the formation of the site, other carnivorans are also important agents to consider when investigating the Guadix Baza region.

## 1. Introduction

The Lower Pleistocene of Europe is a key time period that has provided insights into the activity and biology of some of the first hominin populations outside of Africa. Throughout this moment in human evolution, hominins and carnivorans have been documented to have complex relationships, often competing for many of the same resources (Binford, 1981; Brain, 1981; Domínguez-Rodrigo et al., 2007; Rodríguez et al., 2012; Lozano et al., 2016; Rodríguez-Gómez et al., 2016; *inter alia*). In light of the importance of meat consumption in early human evolution, this topic is especially interesting when considering the intensity of this competition for some early European sites and the influence this may have on population dynamics (Périquet et al., 2015), or the basic survival of early *Homo* (Turner, 1992).

One of the most important sites that presents the earliest evidence of this type of competition outside of Africa is the Eurasian

[*] Corresponding author. Department of Cartographic and Land Engineering, Higher Polytechnic School of Avila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain.
*E-mail address:* ladc1995@gmail.com (L.A. Courtenay).

https://doi.org/10.1016/j.quascirev.2022.107912
0277-3791/© 2022 Elsevier Ltd. All rights reserved.

site of Dmanisi (Tappen et al., 2007, 2022). While sites predating 1 Ma are scarce in Europe, iconic sites such as Pirro Nord (Cheheb et al., 2019), Bois de Riquet (Bourguignon et al., 2016), le Vallonet (Echassoux, 2004), and Sima del Elefante (Huguet et al., 2013), are also fundamental in the study of this topic in early human evolution. Likewise, later Early Pleistocene sites such as Barranc de la Boella (Pineda et al., 2015, 2017), and level TD6 from Gran Dolina (Saladié et al., 2014), are also key case-studies presenting carnivoran-hominin competition. Each of these cases present interesting insights into carnivoran-hominin interactions, ranging from examples of low (e.g. Dmanisi, Bois de Riquet & Barranc de la Boella; *ibid*) to high (e.g. TD6; *ibid*) anthropogenic activity. Other sites such as Le Vallonet and Sima del Elefante (*ibid*), on the other hand, present bones with both carnivoran and anthropogenic cut marks. In either case, both carnivorans and hominins clearly coincided in these sites with the intent of obtaining nutritional resources of some form.

Barranco León (BL, 1.46 Ma, Orce, Guadix Baza, Spain) is another example of an emblematic open-air site from this time period, presenting a clear association between lithic and faunal remains, while yielding one of the oldest hominin fossils in south-western Europe (Toro-Moyano et al., 2013). BL is additionally characterised by an intense exploitation of local flint and limestone raw materials, attributed to the Oldowan techno-complex (Titton et al., 2018, 2020, 2021). Technological analyses have also identified stone raw material exploitation to be concentrated on the production of small flakes in the case of flint, and pounding/percussive activities in the case of limestone (Barsky et al., 2015a; Titton et al., 2018, 2021).

The faunal assemblage of BL presents a diverse range of species in association with human activity, including anthropogenically processed herbivorous large ungulates, such as hippopotamids, equids and cervids, alongside smaller reptiles such as chelonians (Espigares et al., 2019; Yravedra et al., 2022a). Nevertheless, while BL is a site rich with fossil material of archaeological interest, the nature of hominin intervention in this site is still ambiguous, due to the large array of carnivoran activity documented. This palimpsest, alongside other key localities such as Venta Micena (Palmqvist et al., 2002; Luzón et al., 2021), and Fuente Nueva 3 (Espigares et al., 2013; Rodríguez-Goméz et al., 2016; Yravedra et al., 2021), has revealed a great degree of trophic pressure and competition for resources in the region of Guadix Baza. These observations have led authors to construct models with carnivores, such as *Pachycrocuta brevirostris,* as a protagonist in the modification of sites, feasting on the remains of carrion left by other interveners, such as hominins (Espigares et al., 2013), or machairodontine (saber-tooth) felids (Palmqvist et al., 2007a,b, 2011; Rodríguez-Goméz et al., 2016).

Nevertheless, recent taphonomic analyses of the faunal assemblages in this area present contradictory views (Espigares et al., 2019; Yravedra et al., 2021, 2022a), arguing that, at present, insufficient data is available to support these views. These authors argue that the frequency and intensity of carnivoran bite damage is not necessarily analogous with the activity of hyaenids, while the location of cut marks on anatomical elements of higher nutritional value is not an indication of secondary access. In contrast, the frequency of these cut marks does not fit in with models where hominins have primary access to prey (Domínguez-Rodrigo, 1997; Domínguez-Rodrigo, 1999; Domínguez-Rodrigo and Barba, 2006; Domínguez-Rodrigo et al., 2007). From this perspective, the interpretation of BL is still open to debate, while more information is needed in order to truly understand the competition among the multiple agents present in this region.

This study has the objective of performing an in depth analysis on the taphonomic evidence of carnivoran activity in the site of Barranco León, in particular the tooth pits found on bone. The estimation of trophic pressure will play a fundamental role on the interpretation of this site, as data of this nature can provide key insights into the adaptability of early European hominins. Here we support taphonomic finds with the use of 3D modelling, geometric morphometrics, robust statistics, and computational learning in order to present new insights into the bone surface modifications observed on faunal materials. As a result, this study will demonstrate the advantages of using new technological advances to support the interpretation of archaeological and palaeontological sites.

## 2. The site of Barranco León

The Early Pleistocene site of Barranco León (BL) is found in the northeastern part of the Cenozoic Guadix-Baza Basin (Fig. 1). This area, located in proximity with the town of Orce (Granada, Spain), is rich with archaeological and palaeontological deposits, all of which are crucial for the study of early human evolution. BL is an open air site, consisting of 9 stratigraphic levels (Fig. 1; Anadón et al., 2003; Anadón and Gabás, 2009; Oms et al., 2011), of which levels D1 and D2 are the most relevant in archaeological terms. The present study focuses on the taphonomic analysis of the richest of these layers, level D1 (BL-D1), dated at 1.46 ± 0.19 Ma using Electron Spin Resonance (Toro-Moyano et al., 2013), and in accordance with palaeomagnetic and biostratigraphic data situating this level in the upper Matuyama chron (Oms et al., 2000).

The site of BL is located on the palaeoshoreline of the Guadix-Baza lake. This particular locality has thus been characterised by bordering marginal freshwaters, sourced from the adjacent highlands and mixed with surface and hydrothermal waters from the main saline lake (Anadón et al., 2015). Geologically, BL-D1 is characterised by sandy gravels, product of a sudden event producing high-energy currents (Oms et al., 2011). Palaeoenvironmental data reveal the BL-D1 accumulation to have occurred in a Mediterranean woodland or shrubland environment, yet without an important grassy component (Saarinen et al., 2021), while precipitation and temperature levels would have been higher than in present (Blain et al., 2011, 2016; Sánchez-Bandera et al., 2020; Martínez-Monzón et al., 2021).

The lithic assemblage of BL is mostly comprised of local flint and micritic limestone tools, typical of the Oldowan technocomplex (Turq et al., 1996; Gibert et al., 1998; Toro-Moyano et al., 2009, 2010, 2011, 2013; Barsky et al., 2010, 2015b; Titton et al., 2018, 2020, 2021, *inter alia*), with some quartzite tools as well (Toro-Moyano et al., 2011). These include cores, flakes, flake fragments, debris, retouched pieces, angular fragments, hammers, unmodified cobbles, and subspheroids (Titton et al., 2020). Specialised use of raw materials reveal flint to be preferable for the production of small sharp implements, while limestone is found in the form of percussive devices (Titton et al., 2018). The integrity of the assemblage is highlighted by the presence of refitted artefacts (Toro-Moyano et al., 2013; Titton et al., 2021), additionally highlighting *in situ* knapping processes.

Faunal remains consist mostly of equids, followed by cervids, hippopotamids and bovids (Fig. 2, Supplementary Materials; Espigares et al., 2019; Yravedra et al., 2022a). As for the carnivoran species identified, the BL assemblage contains remains of Hyaenidae such as *Pachycrocuta brevirostris*; Canidae including *Xenocyon (Lycaon) lycaonoides, Canis mosbachensis* and *Vulpes* cf. *alopecoides*; Ursidae including *Ursus etruscus*; Mustelidae including *Meles meles* and *Martellictis ardea* (Ros-Montoya et al., 2021); and finally, the felid cf. *Homotherium* sp. (Martínez-Navarro et al., 2010; Espigares et al., 2019; Yravedra et al., 2022a). While not directly identified in the BL assemblage, BL is also contemporaneous with other large Felidae, including *Megantereron cultridens/whitei*, *Panthera*

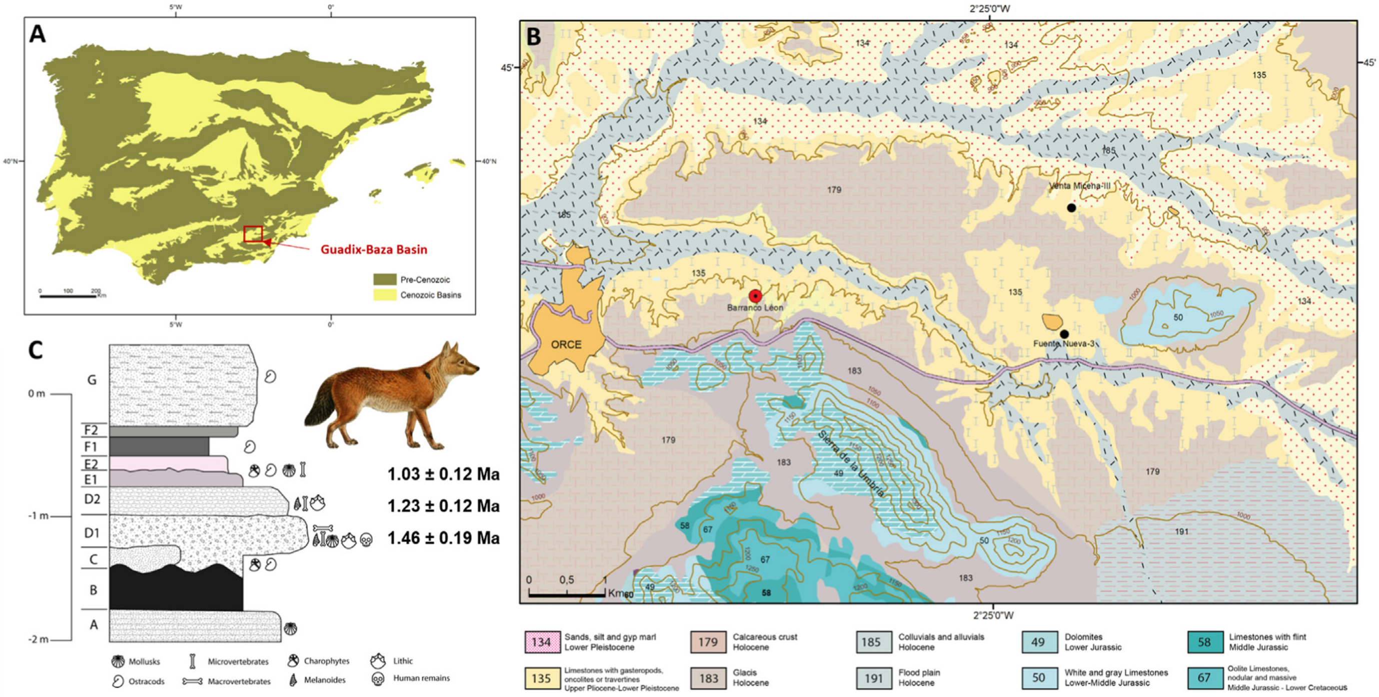

A
5°W
0°
40°N
Guadix-Baza Basin
Pre-Cenozoic
Cenozoic Basins
0 100 200 Km
B
2°25'0"W
45'
134
135
179
183
185
191
49
50
58
67
Venta Micena-III
Barranco León
Fuente Nueva-3
ORCE
Sierra de la Umbría
0 0,5 1 Km
134 Sands, silt and gyp marl Lower Pleistocene
135 Limestones with gasteropods, oncolites or travertines Upper Pliocene-Lower Pleistocene
179 Calcareous crust Holocene
183 Glacis Holocene
185 Colluvials and alluvials Holocene
191 Flood plain Holocene
49 Dolomites Lower Jurassic
50 White and gray Limestones Lower-Middle Jurassic
58 Limestones with flint Middle Jurassic
67 Oolite Limestones, nodular and massive Middle Jurassic - Lower Cretaceous
C
0 m
-1 m
-2 m
G
F2
F1
E2
E1
D2
D1
C
B
A
1.03 ± 0.12 Ma
1.23 ± 0.12 Ma
1.46 ± 0.19 Ma
Mollusks
Ostracods
Microvertebrates
Macrovertebrates
Charophytes
Melanoides
Lithic
Human remains

**Fig. 1.** Location of Barranco León (Guadix-Baza Basin). A. General location of the Guadix-Baza Basin in the Iberian penisula, B: Regional location of Barranco León in the area around the town of Orce. C: Stratigraphy for Barranco León.

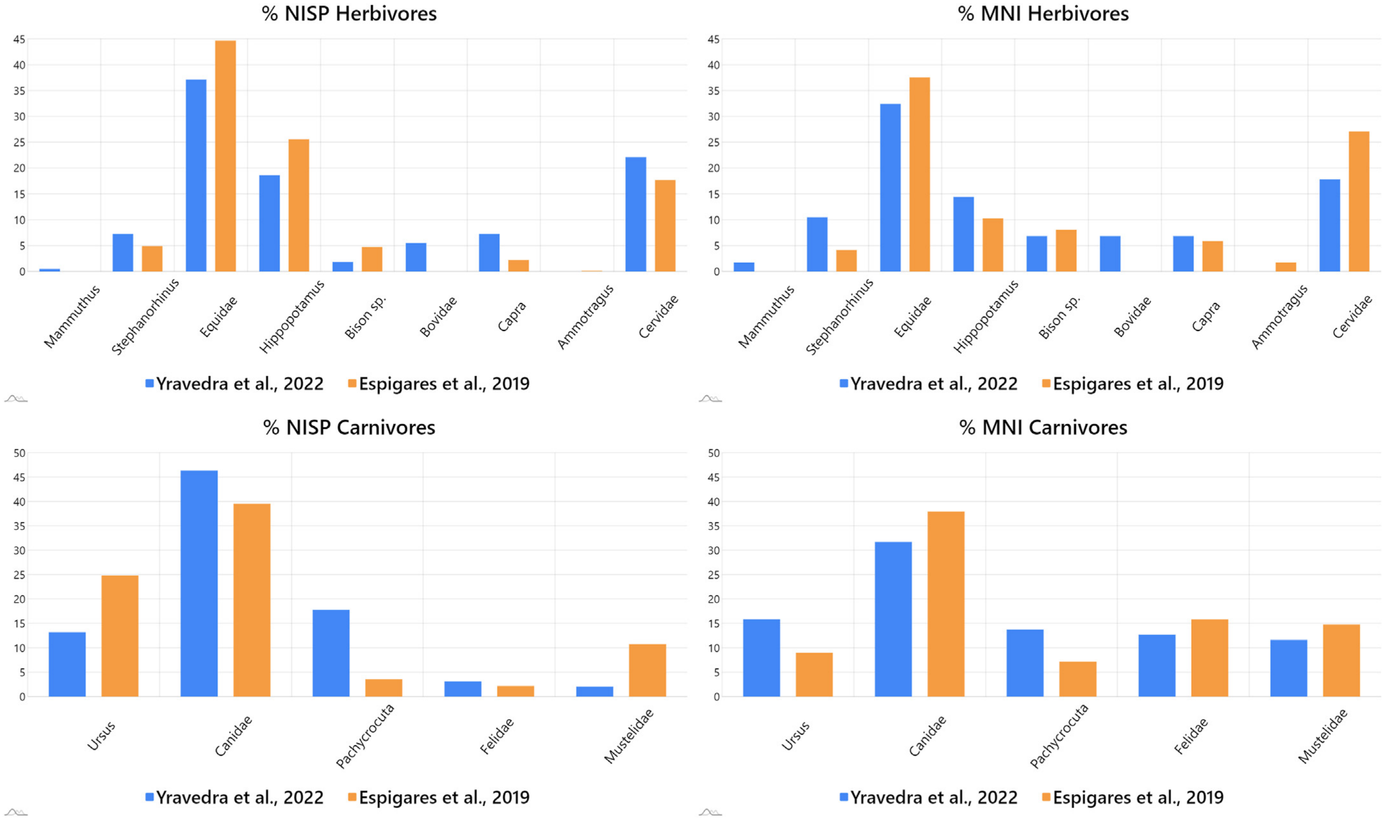


**Fig. 2.** Taxonomic profiles of BL-D1 calculated by Yravedra et al. (2022a) and Espigares et al. (2019), according to Number of Identifiable Specimens (NISP) and Minimum Number of Individuals (MNI) (See Supplementary Files 1 and 2).

*gombaszoegensis* and *Acinonyx pardinensis* (Turner, 1992, 1995; Turner and Antón, 1997; Antón, 2013).

Zooarchaeological analyses reveal the BL-D1 assemblage to be mostly composed of adult individuals, with the exception of larger animals (Size 5: 500–1000 kg; Bunn, 1982), who are predominantly represented by non-adult individuals (Supplementary File 2). Skeletal profiles are fully represented, with a predominance of appendicular and cranial elements (Supplementary File 3; Espigares et al., 2019; Yravedra et al., 2022a). Taphonomic data from these remains highlight an intense degree of fragmentation and fractured bones. Cut and percussion marks appear on a wide array of taxa, of all sizes, while carnivoran alterations are intermixed alongside anthropogenic alterations.

Nevertheless, the interpretation of this data is debated by authors, with Yravedra et al. (2022a) proposing carnivorans to have had a more secondary access to faunal remains, as seen through low tooth mark frequencies, as opposed to the primary access proposed by Espigares et al. (2019).

## 3. Materials and methods

### 3.1. Sample

The osteological samples obtained from BL-D1 that have been analysed here include 10,848 specimens, excavated between the years 2016 and 2020. The present study thus complements prior zooarchaeological and taphonomic research associated with the levels BL-D1 and BL-D2 (Yravedra et al., 2022a). From this perspective, here we develop a more detailed perspective on the carnivorans who have interacted with the fossil accumulations of BL-D1, and how they may have affected human activities. For the purpose of this study, only bones presenting good cortical surface preservation rates were included, reducing the original sample size to 3559 specimens. Bones were considered to present good cortical preservation if overlying taphonomic processes, such as abrasion or weathering, hindered the ability to inspect cortical surfaces and identify anthropogenic or carnivoran traces (Yravedra, 2005). These specimens were then inspected for the presence of tooth marks.

Tooth marks can be typically categorised into 4 main groups, including (Haynes, 1980, 1983; Binford, 1981; Brain, 1981; Blumenschine, 1995); rounded circular depressions (*pits*), elongated depressions or linear marks with a rounded base (*scores*), circular holes (punctures), and the progressive deletion of bones seen by damage to edges (*furrowing*). The present study has focused the in-depth analysis of tooth marks to only consider tooth pits. This criteria was chosen considering tooth scores to be problematic when including captive carnivorans as a reference sample (Courtenay et al., 2021b), while the use of tooth pits have in general been found to produce higher quality results (*ibid*; Courtenay et al., 2019, 2020a).

Once localised, only entire tooth pits found on the diaphyseal portion of long bones were then separated for digital modelling. This is preferred considering how diaphyses are denser than epiphyses, and are thus more likely to survive both carnivoran feeding as well as taphonomic processes. Carnivorans were fed a number of different sized animals, dependent on the regulations established by the institution where each animal is kept (where applicable). Nevertheless, considering the typical size of prey some of these carnivorans are known to consume, as well as additional data regarding the statistical equivalency of marks on different sized animals, these variables were also considered to be unimportant when selecting tooth pits (Courtenay et al., 2020a, 2021b). Once modified, bones were collected and cleaned in boiling water for 12 h, without the use of additional chemicals.

Fossil materials were cleaned with great care, employing the use of a brush and water solvent, while considering recommendations described by Valtierra et al. (2020).

For the analysis of tooth pit morphologies, the selected fossil tooth marks were then analysed alongside modern comparative materials, consisting of 613 tooth pits originally described and studied by Courtenay et al. (2021a). These include tooth pits by brown bears (*Ursus arctos*, Ursidae, $n = 69$), spotted hyenas (*Crocuta crocuta*, Hyaenidae, $n = 86$), foxes (*Vulpes vulpes*, Canidae, $n = 53$), wolves (*Canis lupus*, Canidae, $n = 80$), African wild dogs (*Lycaon pictus*, Canidae, $n = 89$), leopards (*Panthera pardus*, Felidae, $n = 77$), jaguars (*Panthera onca,* Felidae, $n = 77$), and lions (*Panthera leo,* Felidae, $n = 82$). Samples include tooth pits produced by a mixture of both wild and captive carnivorans. Nevertheless, considering observations made by Courtenay et al. (2020a), captivity is not likely to be a major conditioning factor in tooth pit morphology.

Alongside these modern samples, fossil tooth mark samples attributed to *Pachycrocuta brevirostris,* originating from the Early Pleistocene site of Venta Micena 3 (VM3), were also used for comparative purposes (Yravedra et al., 2022b).

For more details on the modern comparative samples, consult Courtenay et al. (2021a) and their corresponding supplementary materials. For more details on the comparative fossil samples, consult Yravedra et al. (2022a, 2022b).

### 3.2. Methods

The objectives of the present study are to characterise the interaction of carnivorans with the osteological accumulations found at BL-D1. For this purpose, the present study will be limited only to the description and analysis of carnivoran alterations, including bite damage, fracture planes, as well as digestive alterations. For details on other non-carnivoran related taphonomic processes, consult Yravedra et al. (2022a).

#### 3.2.1. Taphonomic analyses of Carnivoran activity

To characterise the activity of carnivorans in BL-D1, osteological remains were classified and grouped into bones that are taxonomically determinable, or bones that are indeterminable. Wherever possible, indeterminable bones were grouped according to size, following the categories described by Bunn (1982) and Bunn and Pickering (2010). From this perspective, and following the same criteria described by Yravedra et al. (2021, 2022a), fauna were divided into six groups; Microfauna (Size 0), including species weighing less than 25 kg; Very Small Size (1), including macrovertebrates species weighing 25–50 kg; Small Size (2), including species weighing 50–125 kg; Intermediate Size (3), including species weighing 125–500 kg, with an additional division between 3a (125–250 kg) and 3 b (250–500 kg); Large Size (4), including species weighing 500–1000 kg; and Very Large Size (5) for species weighing >1000 kg. Carnivorans were classified according to three separate size classes: small carnivorans (e.g., foxes, lynxes, and mustelids); intermediate carnivorans (e.g., *Canis mosbachensis*); and large carnivorans (e.g., *Homotherium, Megantereon, Ursus* and *Pachycrocuta*), following Espigares et al. (2019).

Bone cortical surfaces were then inspected using 10–40x handheld lenses. Tooth marks were classified either as pits, scores or punctures, while furrowing damage was also analysed (Binford, 1981; Blumenschine, 1995; Blumenschine et al., 1996). Modifications were quantified for specimens with well-preserved bone surfaces, in terms of NISP values.

Once identified, marks were quantified considering their distribution according to anatomic element, while also inspecting the intensity of modifications. This included the calculation of pit-score ratios on long bone diaphyses, which could then be compared with modern day comparative samples described by Arriaza et al. (2019).

Furrowing was also considered following the “taphotype” classes described by Domínguez-Rodrigo et al. (2015), however as this alteration has only been observed on 0.5% of the sample ($n = 18$, Yravedra et al., 2022a), insufficient information is available for an in-depth characterization.

Bone breakage was assessed following the suggestions of multiple authors (Villa and Mahieu, 1991; Alcántara-García et al., 2006; Pickering and Egeland, 2006; Moclán et al., 2019), taking into consideration fracture plane type, metric properties, as well as the presence, absence and type of notch. From this perspective, all green fracture planes were measured using a goniometer (measurement error ≈ 5°), as described by Villa and Mahieu (1991). Notches were then analysed through both descriptive and metric approaches. Considering the current available reference samples for this type of data, only appendicular bones could be used for this analysis, excluding metapodials. Likewise, equids, as well as small (0–2) and large (4 & 5) sized animals had to be excluded. Once data had been collected, these were compared with experimental samples provided by Moclán et al. (2019), which include bones broken by anthropogenic, hyaenid, and canid activities. For this, qualitative and quantitative data were assessed using a Factor Analysis of Mixed Data (FAMD) (Pagès, 2004), followed by the classification approaches described by Moclán et al. (2019). The best performing algorithm obtained for this purpose was the Random Forest (RF; 88.3% Accuracy, Kappa = 0.80).

Finally, when considered necessary, conclusions drawn from zooarchaeological and taphonomic data were supported by complimentary statistical tests. These included tests for equal proportions according to Pearson's $\chi^2$ test statistic, as well as $\chi^2$ contingency table tests. Where tests on contingency tables were found to compute an unreliable test statistic, this test was replaced by the *G* correction of the test statistic (Sokal and Rohlf, 1981).

All statistics were computed using the R v4.0.4 programming language.

### 3.2.2. 3D modelling and landmark digitization procedures

3D modelling procedures of the selected tooth pits were performed using Structured Light Surface Scanning (Fig. 3). The equipment used was the DAVID SLS-2 located in the C.A.I. Archaeometry and Archaeological lab of the Complutense University of Madrid (Spain). This is a low-cost, powerful, and portable piece of equipment (Maté-González et al., 2017), which could be easily transported to the Museum where fossil materials are located.

Once models had been constructed, tooth pits were digitized using a landmark configuration consisting of 25 landmarks (Fig. 4); five fixed Type II landmarks located on the exterior and interior of each pit, and a 5 × 5 sliding semilandmark patch, removing semilandmarks that overlap with the 5 fixed landmarks (Courtenay et al., 2020b). The 5 fixed landmark are used to mark the maximal length (LM1 & LM2), width (LM3 & LM4), and depth (LM5) of each pit. For the correct orientation of the pit, LM1 is placed farthest away from the perpendicular axis marking the maximal width, and LM2 is thus placed on the opposite extremity. LM3 and LM4 are then placed along this perpendicular axis marking the left (LM3) and right (LM4) maximum extremities. The semilandmark patch is then positioned over the entirety of the pit, so as to capture the internal morphology of the mark and its walls (Fig. 4). Sliding of semilandmarks was performed by minimising

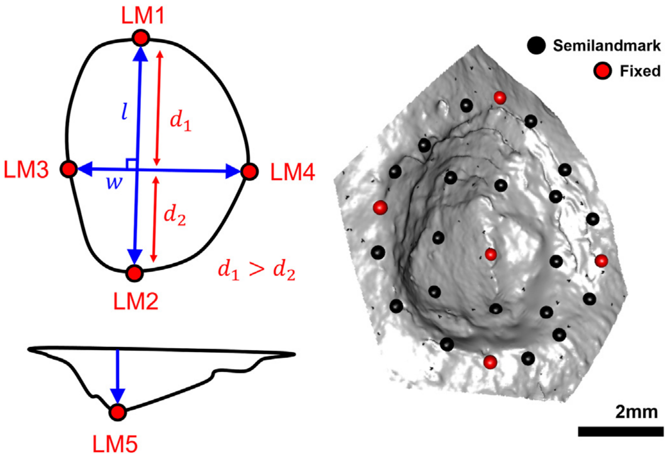


**Fig. 4.** Detailed graphical description of the landmark model employed. LM = Landmark. w = maximum width. d = distance.

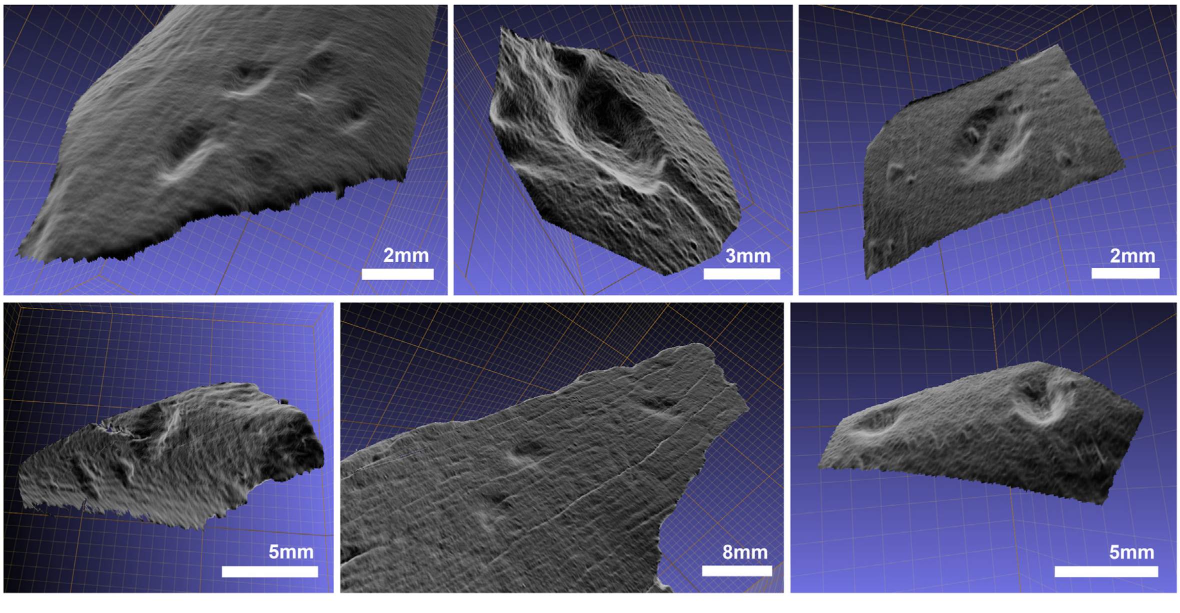


**Fig. 3.** D models of bite damage produced by carnivorans in the site of Barranco León. 3D models were obtained using Structured Light Surface Scanning.

bending energy based on the Thin Plate Spline (TPS) approach (Bookstein, 1991, 1997; Gunz and Mitteroecker, 2013).

The repeatability of this landmark model was robustly defined as 0.139 ± 0.092 $\in$ {0.002:0.586} mm (Courtenay et al., 2020b). These human-induced errors are product of analyst experience, and the time taken to perform the study. All landmarking procedures should thus follow the detailed instructions provided in the main paper and supplementary materials of Courtenay et al. (2020b), while digitization sessions should be performed with as much care and metric accuracy (for defining LM1 to LM5) as possible.

### 3.2.3. *Geometric morphometrics*

Once digitized, landmarks were formatted into morphologika files and imported into the R environment (v.4.0.4). Landmarks were first subjected to a Generalized Procrustes Analysis (GPA), so as to normalize data and project landmarks into a new superimposed feature space (Bookstein, 1991; Rohlf, 1999). In order to take into account observations on the weight of tooth pit size (Aramendi et al., 2017; Courtenay et al., 2019, 2021a, b), GPA was performed excluding the scaling procedure, so as to analyse pits in form space. Once superimposed, landmark configurations were analysed in terms of the Procrustes distances between each other, and Centroid Size (CS) distributions.

For Procrustes distances and CS analysis, distributions were first analysed for homogeneity using Shapiro-Wilk tests. All following statistical tests were then conditioned by these results, using traditional statistical approaches where homogeneity was found to be present, and robust statistical approaches otherwise (Höhle and Höhle, 2009; Rodríguez-Martín, 2019; Courtenay et al., 2020b). Descriptive statistics are then calculated considering the mean or median measurement of central tendency (for Gaussian and non-Gaussian data respectively), while distribution variability is measured in terms of the standard deviation or the Median Absolute Deviation (MAD). From a different perspective, univariate statistical tests were also performed, using a linear ANOVA model, or the Kruskal-Wallis test.

Procrustes distance calculations are additionally used to calculate the reference sample with the closest morphological affinity to each of the fossil tooth pits. From this perspective, the original statistical analysis can be used as a first approximation to the classification of each of the tooth pits, which can later be confirmed or fine-tuned using computational learning.

For multivariate analyses, dimensionality reduction via Principal Component Analysis (PCA) was performed. The PC scores representing up to 99% of morphological variance were then selected and used for further statistical processing. Multivariate Analyses of Variance (MANOVA) were used to assess for differences in form feature space, using either the Hotelling-Lawley or Wilk's Lambda test statistic. Finally, Thin Plate Splines (TPS) were also calculated, using a Delaunay 2.5D Triangulation algorithm to facilitate the visualization of landmark configuration patterns (Bookstein, 1989; Lopez-Fernandez et al., 2017).

All statistics were performed in the R (v.4.0.4) programming language.

### 3.2.4. *Computational learning*

Once analysed statistically, the observations made calculating morphological affinity with Procrustes distances were then supported by classification algorithms that could provide a final more concrete label to each of the tooth pits.

For the classification of each of the fossil tooth marks, computational learning algorithms were trained, following the procedure recommended by Courtenay et al. (2021a). This methodological workflow consists in (1) the augmentation (x100) of each dataset via an unsupervised approach, followed by (2) the training of supervised classification algorithms (Courtenay and González-Aguilera, 2020; Courtenay et al., 2021a). For data augmentation, a multivariate Monte Carlo Markov Chain was used to simulate the morphological characteristics of 100 tooth marks per sample. This was performed for the balancing of data set sizes, while also preventing over/underfitting in later supervised analyses (Courtenay and González-Aguilera, 2020). Quality of augmented data was then evaluated following the suggestions of Courtenay and González-Aguilera (2020). The final augmented datasets were found to be highly equivalent to the original data ($|d| = 0.004$, $p = 6.9e\text{-}29$).

Once augmented, Support Vector Machines (SVM) and Neural Support Vector Machines (NSVM) were trained (Courtenay et al., 2021a). SVMs were trained using a $k$-fold cross-validated approach ($k = 10$), and a *Radial Basis Function* kernel. Optimal configuration of the kernel was computed using Bayesian Optimisation algorithms (Snoek et al., 2012; Shahriari et al., 2016). NSVMs were trained using typical deep learning approaches (Goodfellow et al., 2016), first employing the use of a Laplacian Random Fourier Function (RFF) -based neural network (Rahimi and Recht, 2007; TancikSrinivasan et al., 2020), and then replacing the final activation layer with a linear SVM (Wieringvan der Ree et al., 2013; Courtenay et al., 2021a). NSVMs were trained in batches of 32 pits for 1000 epochs. The Adam optimizer and a triangular cyclic learning rate were employed. Additional tuning of the SVM activation layer was also performed using Bayesian approaches (Snoek et al., 2012; Shahriari et al., 2016).

Both SVM and NSVM were trained on 80:20% train:test sets, using only augmented data for training. Evaluation was performed on the original dataset. Once trained, algorithms were then used to predict labels and label probabilities for each of the fossil individuals. The summary of the two trained algorithm performance on test sets is provided in Table 1.

SVMs were programmed in the R programming language (v.4.0.4), while NSVMs were programmed in Python (v.3.7.4). For more details see Courtenay et al. (2021a).

Once marks had been classified, fossil tooth pits were separated into their corresponding groups for a more detailed assessment and characterization of the fossil species present. This was performed using the same methodological procedure as the geometric morphometric analyses described above, while additionally testing for morphological equivalence using a Two One-Sided Equivalency tests (TOST), according to Cohen's $d$ (Lakens, 2017). For homogeneous distributions, Welch's t-statistic was used, while non-parametric approaches employed the use of Yuen's trimmed robust $t$-statistic.

Finally, TPS were used to calculate mean configurations of different samples, convert these configurations into meshes, and calculate the distance between the faces of each mesh so as to quantify differences between the mean configurations. Distance calculations were computed using the nearest neighbour distance

**Table 1**
Evaluation results on the trained computational learning algorithms that will be used for the classification of Barranco León tooth marks. SVM = Support Vector Machines. NSVM = Neural SVM. Accuracy, Sensitivity and Specificity are values between 0 and 1, with 1 being the highest obtainab1e value. Kappa values are between 0 and 1, with values above 0.8 being considered a powerful model. Loss considers values closer to 0 as the most accurate models.

| | SVM | NSVM |
|---|---|---|
| Accuracy | 0.93 | 0.89 |
| Kappa | 0.86 | 0.85 |
| Sensitivity | 0.89 | 0.81 |
| Specificity | 0.96 | 0.96 |
| Loss | 0.09 | 0.10 |

from a reference mesh to a warped mesh, using as a reference mesh the 3D model corresponding to the median individual of one of the groups (Rohlf, 1998).

### 3.2.5. Hypothesis testing

In accordance with the recommendations set forth by the editors and contributors of the *American Statistician*, *p*-values were not evaluated using $p < 0.05$ as a threshold for defining statistical significance (Wasserstein et al., 2019), while the term "significant" has also been avoided. In its place, all hypothesis testing was performed using Bayesian calibrations for the evaluation of *p*-values. Under this premise, the False Positive Risk (FPR) was calculated for each *p*-value (Colquhoun, 2019), using the Selle-Berger approach (Benjamin and Berger, 2019) for the definition of Null Hypothesis ($H_0$) and Alternative Hypothesis ($H_a$) ratios. Where necessary, FPR was also used to derive Probability of $H_0$ values ($p(H_0)$), providing a means to calibrate *p* values over 0.3681 (Courtenay et al., 2021a, c). Unless specified otherwise, prior probabilities in support of $H_a$ were set at 0.5, indicating complete randomness, as recommended by Colquhoun (2019).

In light of these calibrations, *p*-values were thus evaluated using a robust value of 0.003 ($3\sigma$) as a threshold for more conclusive results. This *p*-value can be considered to have and a FPR of 4.5 +/− [1.2, 15.9] %, using priors of 0.5 +/− [0.2, 0.8] (Courtenay et al., 2021c).

## 4. Results

### 4.1. Taphonomic analyses of carnivore activity

Among the 3559 fossils analysed in this study, 368 tooth marks were identified on 167 bones (4.7%) from BL-D1 (Supplementary File 4 & 5). Tooth marks were found on all taxonomic groups and anatomical elements, nevertheless, appendicular long bones were found to present the highest number of tooth marks (Sup. File 5).

The frequency of tooth marks identified are relatively low, with <5% of the osteological sample presenting carnivoran modifications ($\chi^2$ = 79.2, $p$ = 2.2e-16, FPR = 2.2e-12%; Sup. File 4). When considering only appendicular elements, only 74 bones have been observed to present carnivoran bite damage, which is still <20% of the total sample of appendicular elements from BL-D1 ($\chi^2$ = 34.8, $p$ = 3.6e-09, FPR = 1.9e-05%; Sup. File 5 & 6). The intensity of carnivoran damage can additionally be considered low ($\chi^2$ = 157.5, $p$ = 2.2e-16, FPR = 2.2e-12%), when observing 90% of bite marked bones to present less than 5 marks per bone, and no specimen has been observed to present over 10 marks (Sup. File 7). Finally, only 1% of specimens present digestive alterations.

When testing for trends according to animal size, no notable patterns can be observed for neither the presence of bite damage ($G$ = 2.9, $p$ = 0.57, $p(H_0)$ = 53.4%), nor the frequency of tooth marks per bone ($G$ = 0.23, $p$ = 0.99, $p(H_0)$ = 97.4%).

Analysing the type of bite damage, punctures are rarely found in BL-D1 ($n$ = 13), while pits ($n$ = 199) and scores ($n$ = 156) are the most common type of tooth mark ($\chi^2$ = 231.9, $p$ < 2.2e-16, FPR <2.2e-12%; Sup. File 7 & 8). When analysing these frequencies in more detail, it can be observed that pits dominate on long bones of animals Size 1, 2 and 3 (pit: score ≈ 59.9 : 40.1%), while larger animals present a predominance of scores (pit: score = 20.0 : 80.0%).

The relationship of pit/score ratios according to animal size has also tested to be of notable importance ($\chi^2$ = 16.1, $p$ = 0.0003, FPR = 0.69%). Nevertheless, pit/score ratios have only been found to be important in the case of animal Sizes 4–5 ($\chi^2$ = 19.3, $p$ = 1.1e-05, FPR = 0.03%), while Sizes 1–2 can only be considered to be slightly conclusive ($\chi^2$ = 8.3, $p$ = 0.004, FPR = 5.5%). Animals of Size 3 do not present sufficient differences to be considered conclusive ($\chi^2$ = 1.1, $p$ = 0.30, FPR = 49.5%). When considering the possible effects that small sample sizes may have on these results, corrected prior probabilities for FPR (priors = 0.2; Courtenay et al., 2021c) still reveal a 2.8% probability that this observation is a false positive in the case of animals of size 4–5. Nevertheless, these corrected priors put into question the reliability of observations made for animals of Sizes 1–2 (FPR = 18.97%). In light of these results, tooth mark frequencies observed on animals of Sizes 4 and 5 can be considered to be similar to those produced by modern day large felids (Domínguez-Rodrigo et al., 2012), nevertheless, score:pit ratios alone are not a diagnostic trait for carnivoran activity, while insufficient data is available to provide a conclusion with regards to smaller animals.

Only a very small sample size of 10 bones (0.3%) were found suitable for fracture pattern analysis. These include a mixture of indeterminable long bone shafts from a mixture of Size 3a and 3 b animals. Among this sample, no epiphyseal regions were found to be present, while the average fragment length was measured at 58.4 mm (Interval 2, according to Moclán et al., 2019). The average number of fracture planes was calculated to be 1, while 70% of this sample present longitudinal fracture planes, followed by oblique fracture planes (30%). Plane angles were mostly measured to be acute ($n$ = 5, Circular Mean = 71.6°), followed by obtuse ($n$ = 3, Circular Mean = 107.3°). Notches are present on all accounts, varying in number between 1 and 9 notches (average = 3.7), while notch typologies for this sample are mostly incomplete ($n$ = 6), followed by simple notches ($n$ = 2) and micro-notches ($n$ = 2). When performing classifications on these samples, calculations reveal 6 of the bones to have been broken by canids (87.0% confidence), 2 bones to have been broken by anthropogenic agents (74.9% confidence), while 2 bones remain indeterminable (<65% confidence). None of the bones in this sample were found to be product of hyaenid activity. When analysing this data statistically (Fig. 5), the BL-D1 sample is described mostly by low number of fracture planes, small fragment sizes, and mostly acute angles, similar to patterns described multivariately by modern day *Canis lupus*. Nevertheless, this archaeological sample must be increased in future.

### 4.2. The classification of tooth pits from BL-D1

The BL-D1 tooth mark sample analysed in the present study consist of 64 pits observed on 29 specimens. The majority of tooth pits originate from indeterminable fragments ($n$ = 11, 37.9%; Table 2), followed by Size 3 ($n$ = 9, 31.0%) and Size 2 animals ($n$ = 7, 24.1%). In addition, a single fragment originates from a Size 4 animal (3.4%). Of the 29 bones, only three have been classed as identifiable on a species level, including *Capra alba* (BL19-J48-D1-1), *Equus* sp. (BL20-F47-D1-66), and a single carnivoran individual; *Ursus etruscus* (BL19-F47-D1-29).

The majority of these pits are small (CS Median = 4.4 mm, MAD = 2.6). Robustly calculated 95% confidence intervals ([1.6, 11.8] mm) additionally approximate these samples more in the range of modern wolves (Median = 5.0 mm, MAD = 2.4, 95% CI = [2.8, 9.5]), than any other of the modern comparative samples (Courtenay et al., 2021a). When analysing statistical differences, the BL-D1 sample appears to be more similar to canids of the *Canis* genus ($\chi^2$ = 0.2, $p$ = 0.7, $p(H_0)$ = 57.2%), while also revealing some proximities with ursids ($\chi^2$ = 2.7, $p$ = 0.1, $p(H_0)$ = 38.7%). Similarly, the majority of the sample is found to be different in size to the tooth pits left by large felids ($\chi^2$ = 61.7, $p$ = 4.1e-15, FPR = 3.7e-13%), followed by Hyaenids ($\chi^2$ = 30.19, $p$ = 3.9e-08, FPR = 1.8e-06%). Nevertheless, the BL-D1 sample presents an inhomogeneous distribution ($w$ = 0.9, $p$ = 0.0002, FPR = 0.005%), with large variability,

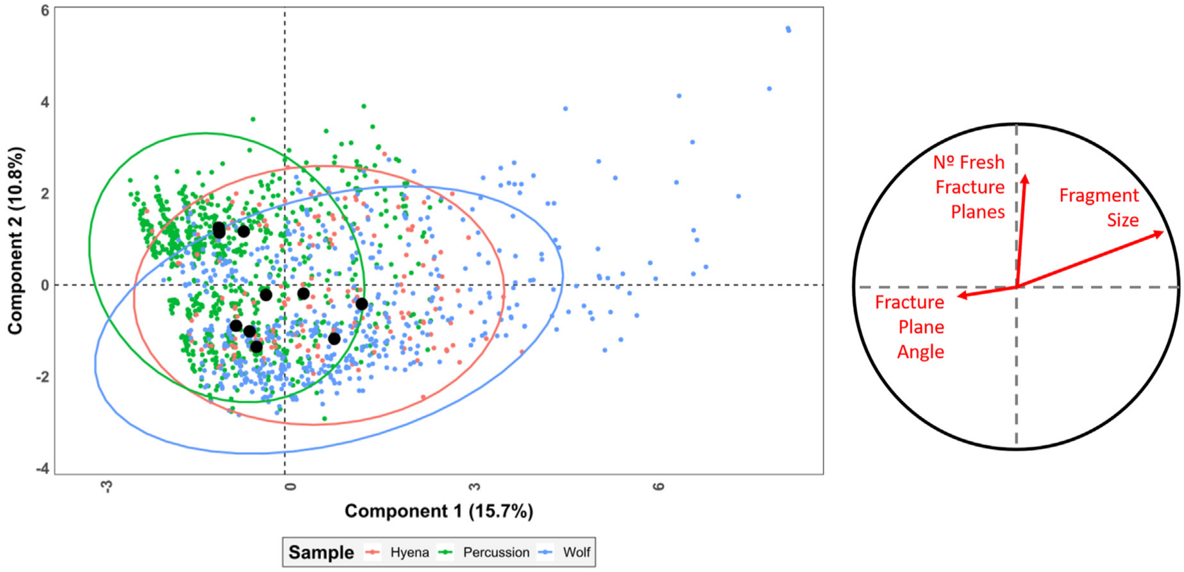


**Fig. 5.** Factor Analysis of Mixed Data plot of fracture plane variables from modern day carnivore reference samples (Moclán et al., 2019), and the preliminary Barranco León sample (black circles).

indicating the possible intervention of multiple sized carnivorans, with a predominance of smaller animals.

In terms of Procrustes distances, a strong morphological signal can be detected when comparing the entirety of the BL-D1 sample with wolves ($d = 2.5$), with some affinities with ursids ($d = 2.7$). Procrustes distances are greatest when comparing with hyaenids ($d = 3.5$), followed by large felids ($d = 3.3$), while the genus *Lycaon* is also noted to be different ($d = 3.1$).

When combining both morphological and metric variables in form space, the BL-D1 sample exclusively approximates the morphology of modern day *Canis lupus* ($p = 0.45$, $p(H_0) = 50.6\%$), while presenting notable differences with all other samples ($p < 0.001$, FPR <1.8%).

Classifications of the BL-D1 sample using computational learning confirm these observations, with the majority of samples being classed as morphologically similar to modern day *Canis lupus* ($n = 33$, 51.6%). Nevertheless, as indicated by the large mixture of different sized tooth pits, other carnivorans have also been detected, including large members of the Felidae family ($n = 8$, 12.5%), Ursidae ($n = 7$, 10.9%), and Hyaenidae ($n = 6$, 9.4%). Alongside the genus *Canis*, algorithms were able to detect 5 additional canid tooth marks (7.8%), morphologically similar with modern day *Lycaon pictus*.

Finally, a handful of tooth marks were found to be indeterminable ($n = 5$, 7.8%), with both algorithms unable to reach a consensus (probabilities of <70%). Nevertheless, evaluation of these traces reveal them to be much smaller in size than any of the comparative samples (CS mean = 1.9 mm, sd = 0.4, 95% CI = [1.5, 2.3]), with the smallest tooth marks used for comparisons originating from *Vulpes vulpes* (CS mean = 4.1 mm, sd = 1.8, 95% CI = [1.8, 7.8]). From this perspective, it can be hypothesized that these marks originate from a much smaller carnivoran, or an animal capable of producing small tooth pits, not included within the present comparative sample.

When assessing the carnivorans identified according to each fossil specimen, interesting possible interactions between multiple carnivorans can be observed through tooth marks from different species found on the same specimen. This possible competition for resources can be observed in a number of cases, with the tooth marks of both ursids and canids (specimen BL19-K48-D1-172 and BL-14-I52-D1-152), large felids and canids (BL15-F51-D1-45 and BL16-F53-D1-24), large felids and ursids (BL-16-I52-D1-sn and BL17-H56-D1-16), *Lycaon* and Canis (BL18-H47-D1-21), and *Lycaon* and hyaenids (BL19-J48-D1-1), being predicted to have been found on the same bone. Possible competition between canids and hyaenids is also frequent (BL18-L48-D1-71, DL14-I51-D1-26 and DL15-I52-D1-1). The two additional cases of overlap between two species is observed in the presence of smaller tooth pits in association with large carnivorans (BL-15-F50-D1-4 and BL-19-K48-D1-12).

### 4.3. Characterizing the BL-D1 tooth pit sample

Visualizing morphological variation in accordance with the final classified samples confirms a separation between the predicted species (Fig. 6a), with PC1 (86.53% variance) primarily representing differences in size, and PC2 (1.95% variance) representing variations in shape. Similarly, when comparing the classified fossil samples with the central tendency of their modern day equivalents (Fig. 6b), distributions in form feature space effectively confirm these morphological affinities. From this perspective, each of the detected fossil tooth marks appears to present strong similarities with their modern day relatives (Table 3). In light of this, the suggested *Canis* tooth marks can be associated with *Canis mosbachensis*, the Hyaenidae tooth marks with *Pachycrocuta brevirostris, Lycaon* tooth marks with *Lycaon lycaonoides*, and Ursidae tooth marks with *Ursus etruscus*. While some overlap still exists between samples seen through some equifinality in form Procrustes distances, general trends in multivariate feature space seem to separate *Canis* individuals from both Ursidae, Hyaenidae, and *Licaon*, while Felidae samples are restricted to one extreme of feature space. The smaller size of the BL-D1 *Canis* tooth marks also create some morphological affinities with modern day *Vulpes vulpes*, however this point will be explored in further detail in the following section.

In the case of Felidae, the tooth pits from the present sample are slightly smaller than that of the modern day *Panthera leo* (Table 4; Fig. 7), however larger than both that of *Panthera pardus* and *Panthera onca*. Similarly, morphological data reveals the felid tooth pits at BL-D1 to be closer in form to *Panthera leo* (Table 3), while

**Table 2**
Classification results, Procrustes Form Distances (Proc. D), and Centroid Sizes, for each of the Barranco León tooth marks analysed in the present study. Sp. ID indicates the Specimen's ID (Site-Year-Square-Level-Number). Where possible, animal size classes have been included. Procrustes distances are calculated from each pit to their corresponding classified label. Computational Learning (CL) classification percentages are obtained from the most confident classification algorithm (Support Vector Machines (SVM) or Neural Support Vector Machines (NSVM).

| Sp. ID | Size | Class Label | Proc. D | CL Probability | Algorithm | Centroid Size |
|---|---|---|---|---|---|---|
| BL-14-G49-D1-105 | Indet | Felidae | 2.47 | 83.07% | SVM | 11.77 |
| | | Felidae | 1.92 | 83.05% | SVM | 10.14 |
| Sn | 3 b | *Lycaon* | 0.9 | 93.12% | SVM | 6.72 |
| | | *Lycaon* | 1.32 | 86.54% | SVM | 6.26 |
| BL-14-F52-D1-116 | Indet | *Canis* | 1.08 | 75.96% | SVM | 3.47 |
| | | *Canis* | 1.54 | 95.61% | SVM | 2.98 |
| | | *Canis* | 1.14 | 82.30% | SVM | 5.01 |
| BL-14-I52-D1-152 | Indet | *Canis* | 0.29 | 100.00% | SVM | 4.46 |
| | | Ursidae | 1.06 | 79.06% | SVM | 5.89 |
| | | Ursidae | 1.24 | 89.83% | SVM | 5.49 |
| Sn | 4 | Felidae | 2.85 | 85.37% | SVM | 12.46 |
| BL-16-F53-D1-24 | 3 | *Canis* | 2.98 | 96.05% | SVM | 1.48 |
| | | *Canis* | 2.86 | 97.39% | SVM | 1.61 |
| | | *Canis* | 1.44 | 99.16% | NSVM | 5.27 |
| | | Felidae | 1.95 | 88.31% | SVM | 11.77 |
| BL-14-I51-D1-26 | 3a | Hyaenidae | 0.56 | 82.65% | NSVM | 5.1 |
| | | *Canis* | 0.96 | 74.64% | NSVM | 4.33 |
| BL-15-I52-D1-1 | 3 | Hyaenidae | 1.8 | 85.24% | SVM | 7.96 |
| | | *Canis* | 1.25 | 81.07% | SVM | 3.73 |
| BL-15-F50-D1-4 | Indet | *Canis* | 1.68 | 89.67% | SVM | 2.7 |
| | | Indet | | | | 1.9 |
| BL-15-G51-D1-22 | Indet | Indet | | | | 1.63 |
| BL-15-F50-D1-46 | Indet | *Canis* | 1.37 | 85.01% | NSVM | 3.69 |
| BL-15-F51-D1-45 | Indet | Felidae | 2.37 | 90.45% | SVM | 8.99 |
| | | *Canis* | 1.45 | 90.78% | NSVM | 3.05 |
| | | *Canis* | 2.86 | 75.95% | SVM | 5.9 |
| | | Felidae | 1.3 | 81.69% | SVM | 8.32 |
| BL-17-H56-D1-16 | 2 | Ursidae | 0.70 | 98.96% | NSVM | 7.34 |
| | | Ursidae | 1 | 89.55% | NSVM | 4.97 |
| | | Felidae | 2.66 | 85.37% | SVM | 10.51 |
| BL-17-F48-D1-22 | Indet | *Canis* | 2.59 | 87.05% | SVM | 1.88 |
| | | *Canis* | 2.08 | 77.96% | NSVM | 2.39 |
| BL-17-I48-D1-200 | Indet | *Canis* | 1.68 | 97.39% | SVM | 2.83 |
| BL-18-J48-D1-10 | 3 b | *Canis* | 0.79 | 94.51% | NSVM | 4.12 |
| | | *Canis* | 1.1 | 91.44% | NSVM | 2.3 |
| | | *Canis* | 2.16 | 90.28% | NSVM | 2.1 |
| BL-18-L48-D1-71 | 2 | Hyaenidae | 1.68 | 75.15% | NSVM | 6.97 |
| | | Hyaenidae | 2.17 | 98.29% | NSVM | 9.2 |
| | | *Canis* | 1.16 | 80.56% | SVM | 3.47 |
| | | *Canis* | 0.94 | 100.00% | SVM | 3.63 |
| BL-18-J48-D1-144 | 2 | *Canis* | 0.98 | 97.07% | SVM | 3.64 |
| BL-18-H47-D1-21 | 3 | *Canis* | 0.8 | 100.00% | SVM | 4.44 |
| | | *Canis* | 1.63 | 80.67% | NSVM | 2.92 |
| | | *Lycaon* | 1.16 | 85.65% | NSVM | 6.01 |
| BL-18-K48-D1-223 | 2 | *Canis* | 1.04 | 98.79% | SVM | 3.49 |
| | | *Canis* | 0.88 | 80.84% | SVM | 3.77 |
| BL-19-J48-D1-1 | 2 | *Lycaon* | 1.10 | 83.44% | SVM | 7.12 |
| | | *Lycaon* | 1 | 90.48% | NSVM | 5.62 |
| | | Hyaenidae | 1.58 | 87.81% | NSVM | 8.06 |
| BL-19-L48-D1-15 | 2 | *Canis* | 1.16 | 84.22% | NSVM | 4.24 |
| BL-19-K48-D1-12 | Indet | Indet | | | | 2.33 |
| | | Indet | | | | 2.15 |
| | | Indet | | | | 1.47 |
| | | *Canis* | 1.04 | 84.27% | SVM | 3.6 |
| BL-19-K48-D1-172 | 2 | *Canis* | 2.66 | 86.30% | NSVM | 1.8 |
| | | *Canis* | 1.67 | 85.50% | NSVM | 2.82 |
| | | *Canis* | 2.52 | 85.17% | NSVM | 1.98 |
| | | Ursidae | 0.84 | 83.92% | NSVM | 6.85 |
| BL-19-F47-D1-29 | Carn3 | Hyaenidae | 1.55 | 79.55% | NSVM | 6.17 |
| BL-20-F47-D1-11 | 3 | Ursidae | 1.40 | 90.30% | SVM | 5.2 |
| BL-20-F47-D1-26 | 3 b | *Canis* | 0.43 | 91.81% | SVM | 4.37 |
| BL-20-L48-D1-99 | 3 b | *Canis* | 0.49 | 93.01% | SVM | 4.21 |
| BL-16-I52-D1-sn | Indet | Felidae | 1.62 | 87.08% | SVM | 12.09 |
| | | Ursidae | 1.14 | 89.40% | NSVM | 6.51 |

Procrustes distances are much greater when compared with both other species of felid. From this perspective, and using *Homotherium latidens* as the closest analogy with *Panthera leo* in size (Turner and Antón, 1997; Antón et al., 2005, 2014; Antón, 2013), the BL-D1 sample can therefore be approximated to *H. latidens*, as opposed to the smaller *Megantereon cultridens*, and *Panthera gombaszoegensis*, which are more similar to the modern day *Panthera onca*.

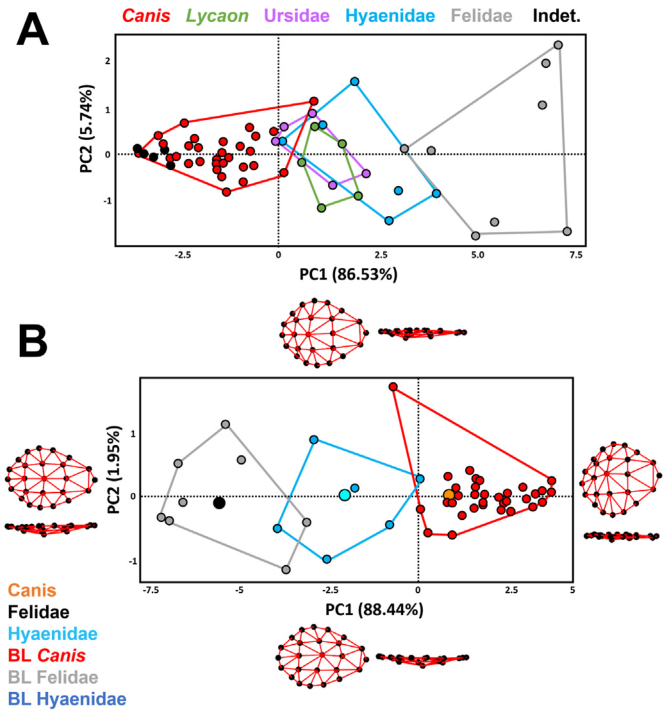


**Fig. 6.** — Principal Component Analyses (PCA) in Form feature space characterizing (A) the morphological variation of each of the BL-D1 samples, and (B) the mean PCA comparing modern day carnivoran central configurations with each of the main BL-D1 samples. For the purpose of visual clarity, ursids, Lycaon and any indeterminable marks have been excluded from panel B. Predicted form deformations via thin plate splines are depicted on each extremity of their corresponding PC score in panel B, employing the use of a 2.5D Triangulation algorithm.

### 4.4. Characterising fossil carnivorans from orce

Due to the smaller sample sizes of fossil *Homotherium latidens* and *Ursus etruscus* tooth pits, the detailed characterization of the fossil tooth pits will focus solely on the analysis of suggested Canidae from BL-D1, as well as predicted Hyaenidae tooth marks combining data from BL-D1 and VM3 (Yravedra et al., 2022b).

#### 4.4.1. Canis mosbachensis

The sample classed as *Canis mosbachensis* is very similar to the comparative samples produced by modern day *Canis lupus*, presenting noticeable overlap in form feature space (Fig. 8). Based on this data, morphologically speaking *Canis mosbachensis* can be described by presenting mostly superficial pits. This links strongly with observations made on *Canis lupus* specimens by multiple authors (Yravedra et al., 2019; Courtenay et al., 2020a, 2021a, b). When compared in general with other large canids, both species of *Canis* are restricted to a portion of feature space described by more asymmetrical pits, with the point of maximal depth shifting closer to one edge of the pit than the other (PC1 = 87.02% variance). The predicted *Lycaon* samples, on the other hand, are observed to be much deeper, with a greater variance across feature space, while their tooth pits are observed to be more elongated, with the point of maximal depth shifting along the LM1-LM2 axis.

When considering size, allometry is present between *Lycaon* and *Canis* samples (F = 22.3, Residuals$^2$ = 0.027, Effect Size (ES) = 5.7, $p = 0.001$, FPR = 1.8%), as well as between *Canis lupus* and *Canis mosbachensis* (F = 19.0, Res.$^2$ = 0.030, ES = 0.3, $p = 0.001$, FPR = 1.8%). Allometry between *Canis mosbachensis* and *Vulpes vulpes* is much less evident (F = 2.72, Res$^2$ = 0.008, ES = 2.1, $p = 0.02$, FPR = 19.1%). These observations indicate size to be a conditioning factor in the morphological variation between most

**Table 3**
Procrustes distances calculated in form associating the final classified fossil species from BL-D1 with their extant relatives. *Including only BL-D1 fossil specimens. **Including both BL-D1 and VM3 fossil specimens.

| | *C. lupus* | L. pictus | C. crocuta | U. arctos | *P. leo* | P. onca | P. pardus | V. vulpes |
|---|---|---|---|---|---|---|---|---|
| *C. mosbachensis* | 1.233 | 3.641 | 4.387 | 2.595 | 8.074 | 3.943 | 0.993 | 0.255 |
| *L. lycaonoides* | 2.302 | 0.799 | 1.321 | 1.312 | 4.749 | 1.149 | 2.606 | 3.470 |
| *P. brevirostris** | 1.282 | 1.270 | 2.433 | 0.641 | 5.676 | 1.572 | 1.523 | 2.433 |
| *P. brevirostris*** | 1.005 | 1.592 | 2.329 | 0.692 | 5.992 | 1.869 | 1.203 | 2.124 |
| *U. etruscus* | 2.173 | 0.790 | 1.418 | 0.908 | 4.884 | 0.948 | 2.378 | 3.319 |
| *H. latidens* | 7.160 | 4.806 | 4.054 | 5.847 | 1.125 | 4.514 | 7.415 | 8.336 |

**Table 4**
Descriptive statistics comparing centroid sizes of each of the samples analysed within this study. *Samples from Barranco León. **Samples from Barranco León and Venta Micena 3 (Yravedra et al., 2022b). Lower and Upper Confidence Intervals (CI) are calculated using robust 95% quantile intervals.

| | | Min | Lower CI | Central | Deviation | Upper CI | Max |
|---|---|---|---|---|---|---|---|
| Canidae | *Lycaon pictus* | 2.23 | 3.13 | 6.65 | 3.06 | 17.24 | 22.09 |
| | *Lycaon lycaonoides** | 5.76 | 5.76 | 6.44 | 0.56 | 7.19 | 7.19 |
| | *Canis lupus* | 2.14 | 2.52 | 4.56 | 2.18 | 8.59 | 14.18 |
| | *Canis mosbachensis** | 1.49 | 1.62 | 3.45 | 1.15 | 5.42 | 6.49 |
| | *Vulpes vulpes* | 1.10 | 1.73 | 3.74 | 1.63 | 7.02 | 7.49 |
| Felidae | *Panthera leo* | 3.90 | 4.52 | 11.00 | 6.00 | 24.67 | 31.51 |
| | *Homotherium latidens** | 8.34 | 8.34 | 10.92 | 1.56 | 12.64 | 12.64 |
| | *Panthera onca* | 1.68 | 2.55 | 6.94 | 3.86 | 19.42 | 28.90 |
| | *Panthera pardus* | 2.44 | 2.67 | 4.40 | 1.92 | 8.53 | 10.65 |
| Hyaenidae | *Crocuta crocuta* | 2.62 | 3.50 | 7.38 | 3.21 | 15.93 | 23.64 |
| | *Pachycrocuta brevirostris*** | 2.49 | 2.57 | 5.40 | 2.54 | 12.16 | 14.93 |
| Ursidae | *Ursus arctos* | 1.54 | 1.63 | 5.87 | 3.67 | 12.69 | 16.90 |
| | *Ursus etruscus** | 4.99 | 4.99 | 6.11 | 0.86 | 7.37 | 7.37 |
| BL-D1 Indeterminable | | 1.48 | 1.48 | 1.92 | 0.36 | 2.36 | 2.36 |

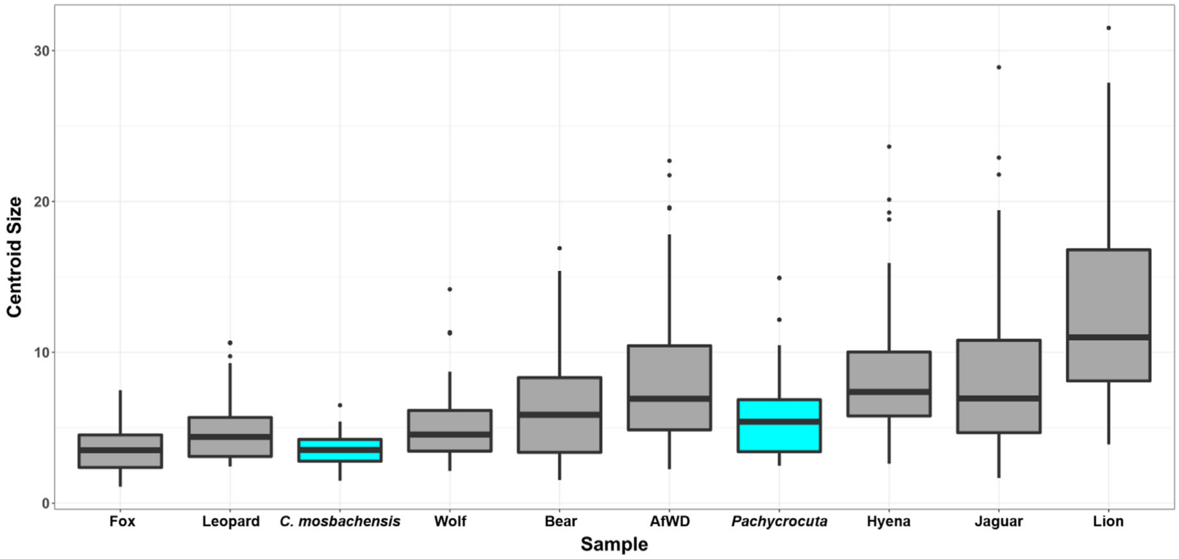


**Fig. 7.** Graphical representation of the centroid sizes described in Table 4. Highlighted boxplots refer to fossil carnivoran species.

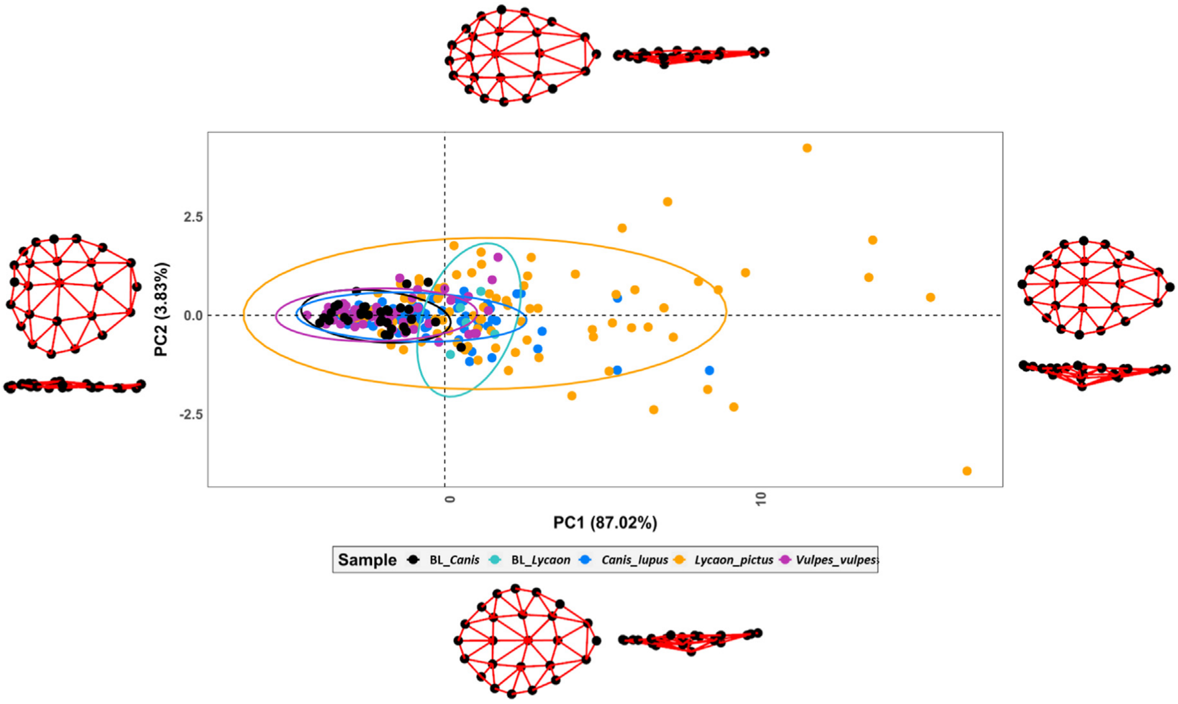


**Fig. 8.** Principal Component Analysis in form feature space characterizing the tooth pit morphologies of the Barranco León and modern day Canidae samples. Predicted form deformations via thin plate splines are depicted on each extremity of the graph. Form deformations are visualized using a 2.5D Triangulation algorithm.

species, while *Canis mosbachensis* and *Vulpes vulpes* present the least amount of morphological variation dependent on size. CS between both *C. lupus* and *C. mosbachensis* are not the same (TOST $t = -0.8$, $p = 0.98$, $p$ ($H_0$) = 94.9%), with *Canis lupus* producing larger tooth marks than that of *Canis mosbachensis* (Table 4), while CS between *Vulpes vulpes* and *C. mosbachensis* are similar (TOST $t = -3.3$, $p = 0.40$, $p$ ($H_0$) = 50.1%). In this context, *Lycaon pictus* proves to be the member of the Canidae family with the largest and deepest tooth pits, while the five *Lycaon lycaonoides* pits detected in BL-D1 fall well within the 95% confidence interval (TOST $t = -3.8$, $p = 0.86$, $p$ ($H_0$) = 73.9%).

Nevertheless, despite the morphological affinities presented between *Canis lupus* and *Canis mosbachensis*, some multivariate differences in form space can still be noted to some extent (MANOVA $p = 0.009$, FPR = 10.3%). *Lycaon pictus* can be observed to be very different in form space ($p = 0.001$, FPR = 1.8%). This indicates both species of *Canis* to be clearly separable from the *Lycaon* genus, while intra-genus differentiation may be possible.

While the proposed *C. mosbachensis* samples also present morphological affinities with *Vulpes vulpes* as well, the closest early relative of this species, *Vulpes* cf. *alopecoides*, was known to be much smaller than modern day *Vulpes* (Garrido, 2008; Lucenti and

Madurell-Malapeira, 2020). Based on this observation, the original classification results, and the fact that *C. mosbachensis* samples fall between both modern day *C. lupus* and *V. vulpes*, we can confirm our original proposal of assigning these pits to *C. mosbachensis*, defining the morphological variability of these pits to be much closer to *C. lupus*, with a size more similar to modern day *V. vulpes*.

Finally, when computing the difference in calculated meshes obtained from central morphological tendencies, the proposed *C. mosbachensis* sample can be found to present the lowest overall distance with *C. lupus* and *V. vulpes* (Table 5), while appearing very different from *Lycaon pictus*. When observing heat maps that display these differences (Fig. 9), the majority of changes appear across the upper extremities of the pit, likely indicating differences in the circular-like nature described in Fig. 6, as well as around LM5 (the deepest point). It can be seen how *C. mosbachensis* is characterised by more superficial pits, as opposed to the deeper pits produced by *V. vulpes*; a characteristic of which supports the prediction that these pits are closer to *Canis* as opposed to *Vulpes* (Courtenay et al., 2021a). When observing differences with the *Lycaon,* both samples appear to be much deeper than *Canis*, while also presenting stronger deformations in the LM1-LM2 axis of the pit. This indicates a greater elongation in the tooth pit. Finally, when considering *Lycaon pictus* as a reference, it can be observed that the two possible *Lycaon lycaonoides* pits are very similar to *Lycaon pictus* (Table 5), with slight differences in some features of the pit between LM5 and L4.

### 4.4.2. *Pachycrocuta brevirostris*

The present sample associated to *Pachycrocuta brevirostris* is described by 31 tooth pits in total; 20 tooth pits originally detected from VM3, alongside 11 tooth pits from BL-D1. Similar to the observations made by Yravedra et al. (2022b), the present tooth marks are observed to present greater morphological affinity to modern day *Crocuta crocuta* than any other carnivoran (Table 3). When combining the two samples, these tendencies increase, with those pits associated with *Pachycrocuta brevirostris* appearing slightly closer to *Crocuta crocuta*, while the degree of separation from other samples increases (Table 3). When analysing the differences between samples, VM3 and BL-D1 appear to present strong morphological affinities between each other (Proc. $d = 2.13$, $p = 0.29$). Nevertheless, the tooth marks from BL-D1 appear to be slightly larger (CS mean = 7.3 mm, sd = 1.5, 95% CI = [5.1, 9.3]) than those from VM3 (median = 5.01 mm, MAD = 2.6, 95% CI = [2.5, 12.2]). The combination of these two samples thus captures a broader spectrum of the proposed *Pachycrocuta* tooth mark morphological variability.

As previously observed, allometry is still an important component of Hyaenidae tooth pit morphological variation (F = 7.6, Res.$^2$ = 0.02, ES = 5.80, $p < 0.001$, FPR <1.8%) (Aramendi et al., 2017; Arriaza et al., 2019, 2021; Courtenay et al., 2021a; Yravedra et al., 2022b), nevertheless, allometric differences between *Crocuta crocuta* and *Pachycrocuta brevirostris* are unimportant (F = 7.6, Res.$^2$ = 0.016, ES = 3.6, $p = 0.225$, FPR = 47.7%). While the current *Pachycrocuta* pits can be observed to be slightly smaller than *Crocuta crocuta* (Table 4), the magnitude of similarities or differences is not of notable value based on the present data (t = −2.8, $p = 0.65$, $p$ ($H_0$) = 56.7%).

Analysis of morphological traits in form space (Fig. 10) reveal both BL-D1 and VM3 samples to greatly overlap with modern day *Crocuta crocuta*. All three Hyaenidae samples can be described by mostly elongated pits, different to those traits previously described for *Canis*, and more similar to the *Lycaon* samples described above. The main morphological patterns that can be described are associated with the position of LM5, and the sliding semilandmarks that shift around it. Similarly, both species of hyaenid present great morphological variance, with the ability to produce a combination of small, large, deep and superficial pits, all within the same sample.

When computing deformations in central configurations (Fig. 9), results only differ slightly from original observations by Yravedra et al. (2022b), in that the change in size is lower than in the present sample, while the average deformation in mesh faces slightly increases from those originally published (0.06 ± 0.04 mm).

## 5. Discussion

This study presents a detailed analysis and description of tooth pits obtained from level D1 of the Lower Pleistocene site of Barranco León. As has been seen throughout this study, the taphonomic story of BL-D1 is much more complex than originally perceived. In previous research, authors propose *Pachycrocuta brevirostris* as the main carnivoran to have intervened in the formation of BL-D1 (Rodríguez-Gómez et al., 2016; Espigares et al., 2019). Nevertheless, the data obtained here reveal the presence of a larger number of carnivorans, with a notable contribution by a relatively small canid which we have associated with *Canis mosbachensis*, thus shedding new light on the interpretation of this site.

Although the present study has been able to detect the activity of *Pachycrocuta brevirostris*, the impact this carnivoran has had on the formation of BL-D1 site is likely smaller than previously considered (Espigares et al., 2019). First of all, the low overall frequency of tooth marks in BL-D1 is not a likely indicator of hyaenid activity (Espigares et al., 2019; Yravedra et al., 2022b: Supplementary File 6), while only 6 of the tooth pits (9.4%) analysed here have been detected to present morphological affinities with hyaenid species (Table 2). Overall fracture patterns, frequency of furrowing, and a <10 number of tooth marks per specimen (Yravedra et al., 2022a), also contradict patterns produced by hyaenid species (Kuhn et al., 2009).

From a different perspective, the role machairodontine felids had in the Orce sites is frequently proposed as being the primary predator, with hominins and hyaenids opportunistically scavenging the remaining carrion (Palmqvist et al., 2007a,b, 2011; Rodríguez-Gómez et al., 2016). From a taphonomic perspective, however, little evidence has been presented that directly detects the action of these carnivorans in the Orce sites. This is due to the proposal that machairodontine felids are unlikely to mark bone during feeding (Marean, 1989; Palmqvist et al., 2007a,b). While this hypothesis is supported by compelling biomechanical data (Palmqvist et al.,

**Table 5**
Description of distances (mm) from the reference mesh to the compared mesh, visualized graphically in Fig. 9.

| | Lycaon pictus | | | | Canis lupus | | | | Vulpes vulpes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Min | Central | Dev. | Max | Min | Central | Dev. | Max | Min | Central | Dev. | Max |
| *L. pictus* | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.34 | 0.20 | 0.64 | 0.09 | 0.53 | 0.27 | 0.98 |
| *L. lycaonoides* | 0.00 | 0.04 | 0.04 | 0.17 | 0.04 | 0.33 | 0.19 | 0.56 | 0.11 | 0.52 | 0.25 | 0.88 |
| *C. lupus* | 0.03 | 0.07 | 0.02 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.11 | 0.34 |
| *C. mosbachensis* | 0.09 | 0.11 | 0.02 | 0.15 | 0.00 | 0.03 | 0.02 | 0.06 | 0.00 | 0.01 | 0.01 | 0.03 |
| *V. vulpes* | 0.08 | 0.10 | 0.02 | 0.14 | 0.00 | 0.02 | 0.02 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 |

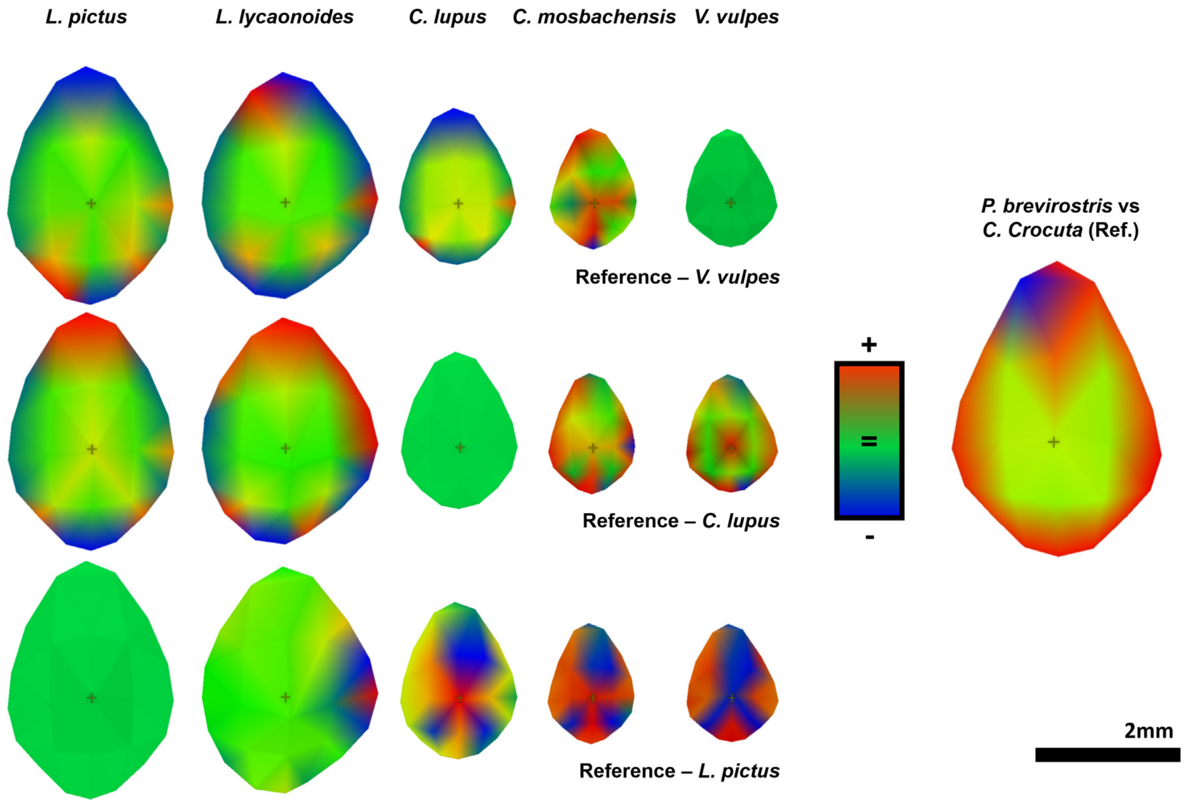


**Fig. 9.** Heatmaps obtained from 3D meshes, computed using the central configuration of each carnivoran sample. Changes in colour indicate how similar one mesh is when compared with the reference mesh, with shades of blue indicating negative deformations and shades of red indicating positive deformations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

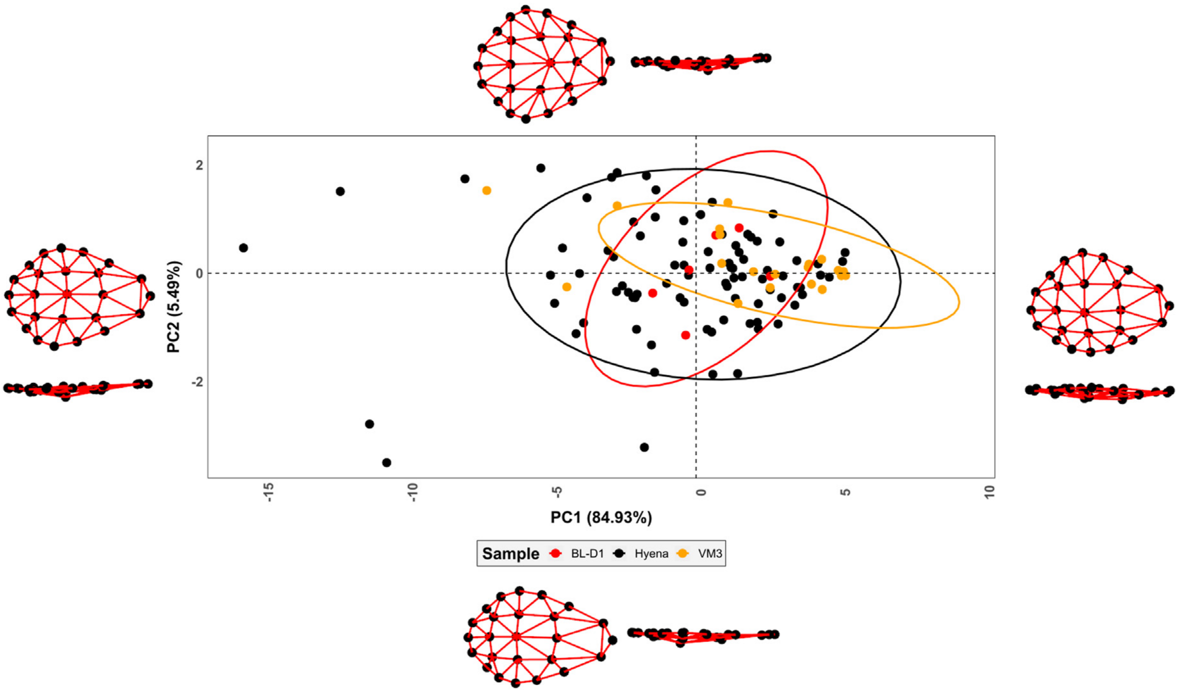


**Fig. 10.** Principal Component Analysis in form feature space characterizing the tooth pit morphologies of the Barranco León, Venta Micena 3, and modern day Hyaenidae samples. Predicted form deformations via thin plate splines are depicted on each extremity of the graph. Form deformations are visualized using a 2.5D Triangulation algorithm.

2007a,b), it is important to note that tooth marks are often accidental. Furthermore, data from other sites have still been able to attribute the presence of tooth marks to machairodontine feeding habits (Marean and Ehrhardt, 1995; Domínguez-Rodrigo et al., 2022). In light of this, bite damage cannot be exclusively related to the chewing or consumption of bone-related nutrients, highlighting a lack of evidence to say that machairodontines were not able to produce tooth marks, just less likely.

While both *Pachycrocuta* and *Homotherium* have been identified here as being likely contributors to both the BL-D1 and VM3 assemblages (Yravedra et al., 2022b), the majority of data from the present sample has been found to be product of *Canis mosbachensis* activity ($n = 33$; 51.6%). *Canis mosbachensis* is mostly known as an omnivorous species (Palmqvist et al., 2008), of smaller size than modern day *Canis lupus*. Modern day *Canis lupus* are social hunters, enabling them to hunt prey larger than themselves (Gittleman, 1985; Vezina, 1985; Yravedra et al., 2011). Nevertheless, considering the physiology of *Canis mosbachensis*, as well as the significantly larger trophic pressure observed in the Early Pleistocene carnivoran guild of Guadix Baza (Rodríguez et al., 2012; Lozano et al., 2016), it is unlikely that *Canis mosbachensis* would have been the predator responsible for the large ungulate carcasses recovered from BL-D1. From this perspective, and also considering the adult-rich age profiles (Yravedra et al., 2022a), *Canis mosbachensis* can be interpreted to have had a secondary role in the modification of fossils. This is especially relevant when considering tooth marks made by different carnivorans on the same specimen (Table 2), implying *Canis mosbachensis* to have scavenged the kills of other carnivorans, such as *Homotherium latidens*.

From the perspective of the other taphonomic agents, remains of both *Lycaon lycaonoides* and *Ursus etruscus* have been identified in BL-D1, represented by a single adult individual each (Espigares et al., 2019; Yravedra et al., 2022a). Here we have been even able to infer the interaction of these species with the assemblage, as seen through the presence of 5 and 7 tooth marks respectively. While *Ursus etruscus* is believed to be an omnivorous species, primarily feeding on plants (Palmqvist et al., 2008; Medin et al., 2017), most bears, such as the brown bear (*Ursus arctos*), are known to scavenge and sporadically hunt (Mattson, 1997). Likewise, ursids from the Venta Micena localities have also been interpreted to present an increased intake of animal based foods (Medin et al., 2017). *Lycaon,* on the other hand, are hypercarnivorous hunters (Estes and Goddard, 1967; Malcolm and Van-Lawick, 1975; Rhodes and Rhodes, 2004). Nevertheless, modern-day African Wild Dogs are not known for generating intense bone surface modifications (Yravedra et al., 2013; Fourvel et al., 2018). In light of these observations, the importance of the present study can be found in the ability to detect these carnivorans, despite their lack of general prevalence as taphonomic agents in many sites. Likewise, while the present study has been unable to specify the precise agents responsible for the smaller marks in BL-D1, the presence of smaller carnivorans such as mustelids in the Guadix Baza region (Madurell-Malapeira et al., 2011; Martínez-Navarro et al., 2010; Ros-Montoya et al., 2021), suggest that they could be a plausible candidate for these marks. Nevertheless, the possibility that other larger carnivorans, external to and not yet detected in this region, could have entered the site and left marks. Finally, the presence of anthropogenic modifications in BL-D1 cannot be ignored (Yravedra et al., 2022a), especially considering the 9 specimens observed to have both cut and tooth marks. While, in general, the percentages of cut and percussion marked bones are relatively low, their location across the appendicular skeleton indicate early access of hominins to the animal resources at this site. Among these evidence, authors have identified both filleting and evisceration activities by human populations (Espigares et al., 2019; Yravedra et al., 2022a). Nevertheless, a more detailed investigation must be carried out into the nature of these alterations.

Finding direct analogies with fossil carnivorans is a complicated process, conditioned by the fragile and incomplete nature of the fossil record. While it is impossible to state with 100% confidence the precise agents intervening in a site (Marean and Ehrhardt, 1995), the proposed methodological approaches can be considered a more empirical advance in identifying these agents (Courtenay et al., 2019, 2021a; Courtenay and González-Aguilera, 2020). The use of computational learning algorithms has been able to make predictions with 88.77 ± 7.49% confidence, assigning each tooth pit to their closest modern day analogy. In light of these calculations, as well as the assessment of Procrustes distances and centroid sizes, this can be considered a valuable new perspective on the interpretation of archaeological and palaeontological sites.

The original hypotheses surrounding the interpretation of the Orce sites primarily identifies *Pachycrocuta* and hominins as the sole modifiers of bones. From the perspective of carnivorans, this theory is based on the “large size” of tooth pits (Espigares et al., 2019). The assumption that large tooth pits are indicatory of *Pachycrocuta* excludes the fact that tooth pits are often produced accidently. Therefore, while large felids may not have been consumers of bone-type nutrients, this does not imply that machairodontines could not have left tooth pits. Likewise, modern day lions are not osteo or durophagic, yet still leave tooth marks (Gidna et al., 2013). Here we have shown that the larger tooth pits of BL-D1 present a greater morphological affinity to Felidae (Proc. $D = 1.125$), than Hyaenidae (4.054). While it could be argued that allometry is conditioning these results, if size is eliminated from analyses, Procrustes distances of shape still reveal these tooth marks to be analogous with a species of felid, over *Crocuta crocuta* (Table 6). Moreover, over 50% of the present sample have been associated with *Canis mosbachensis*, which are closer in morphology to smaller canids than *Pachycrocuta* (Table 6). While Procrustes distances in form space approximate these marks to *Vulpes* as well, shape space highlights the affinities with *Canis* (Table 6), thus enforcing our predictions. Finally, as seen in the present sample, the majority of tooth pits analysed would fall into the range of smaller carnivorans (Andrés et al., 2012, Fig. 5), which contradicts the *Pachycrocuta* based hypothesis.

From the perspective of inter-carnivoran competition for resources, the observation that multiple carnivorans fed on the same bone could be of great interest to taphonomic research. While a possibility remains that some of these examples of “competition” are misclassifications by the algorithms, the morphological

**Table 6**
— Procrustes distances of shape between each of the fossil carnivore species detected, and their modern day analogues. Both the Pachycrocuta and Homotherium samples include tooth pits detected from Barranco León and Venta Micena 3 (Yravedra et al. 2022b).

| | *C. lupus* | L. pictus | C. crocuta | U. arctos | *P. leo* | P. onca | P. pardus | V. vulpes |
|---|---|---|---|---|---|---|---|---|
| *C. mosbachensis* | 0.043 | 0.060 | 0.060 | 0.087 | 0.069 | 0.078 | 0.082 | 0.050 |
| *L. lycaonoides* | 0.082 | 0.092 | 0.101 | 0.109 | 0.100 | 0.112 | 0.120 | 0.116 |
| *P. brevirostris** | 0.036 | 0.039 | 0.032 | 0.053 | 0.051 | 0.048 | 0.046 | 0.042 |
| *P. brevirostris*** | 0.035 | 0.041 | 0.032 | 0.053 | 0.054 | 0.050 | 0.048 | 0.043 |
| *U. etruscus* | 0.062 | 0.075 | 0.086 | 0.062 | 0.090 | 0.079 | 0.079 | 0.112 |
| *H. latidens* | 0.062 | 0.051 | 0.054 | 0.059 | 0.062 | 0.046 | 0.046 | 0.067 |

affinities described by Procrustes distances, and differences in tooth mark sizes, seem to support a differential classification. Future research into the spatial distribution of these traces may help develop our understanding of the order in which different species intervened on carrion (Parkinson et al., 2014, 2015, 2022; Parkinson, 2018; Mora et al., 2022).

In summary, the trophic pressure of the Orce region is of increasing complexity, with evidence of multiple agents competing for resources in the same region. While the most logical hypothesis would place large felids at the top of the food chain, it cannot be denied that both *Pachycrocuta* (Kruuk, 1972; Bearder, 1977; Tilson and Henschel, 1986; Mills, 1984a & b; Henshel, 1986; Cooper, 1990; Turner and Antón, 1996) and hominins (Domínguez-Rodrigo, 1999; Domínguez-Rodrigo et al., 2007), could have played a major role in this ecosystem. From this perspective, BL-D1 can be considered a palimpsest where both carnivorans and hominins would have had to frequently adapt to survive.

## 6. Conclusion

In this paper, 3D modelling, geometric morphometrics, and computational learning were used to provide new insights into the tooth pits observed on faunal materials at Barranco León (Orce, Granada, Spain). Regardless of the scenario, here we present empirical data that strongly implicate the morphological affinities between the tooth marks of BL-D1 with the genus *Canis*. While errors may exist, the present study has based the identification of fossil carnivoran species on an extensive collection of modern analogues. Likewise, this study has been able to detect a number of different carnivorans, including; canids, large felids, hyaenids and ursids, sometimes interacting on the same bones. The most notable carnivore to have been estimated to have intervened in BL-D1 being *Canis* mosbachensis. These observations thus add to the ecological complexity of the Guadix Baza region and highlight a more prominent role of large canids in the southern European Early Pleistocene.

From this perspective, the data collected can be considered a valuable and novel reference collection for the study of extinct carnivoran species. Nevertheless, the site of Barranco León is still under excavation, and a larger sample could be considered fundamental so as to provide an in depth characterization and interpretation of the site.

The nature of the BL-D1 palimpsest is thus gradually appearing to be much more complex than originally perceived, shedding new light on the hominin populations ≈ 1.4 Ma in the Guadix-Baza region. As can be seen, a number of carnivores contributed to this accumulation, mixing the fossil remains with those that may have already been present. From this perspective, computational learning and robust statistics can be considered a valuable contribution to deciphering hominin activities in the taphonomic register.

## Funding

This research was funded by the Junta de Andalucía, Consejería de Cultura: Orce Research Project “Primeras ocupaciones humanas y contexto plaeoecológico a partir de los depósitos pliopleitocenos de la cuenca de Guadix-Baza: zona arqueológica de la Cuenca de Orce (Granada, España), 2017–2020” (Ref: BC.03.032/17). We also received support from the PALARQ Foundation with the convocatory of Analitics 2019): “Identificando Carnívoros a partir de análisis Tafonómicos de última generación aplicando Fotogrametría y Morfometría Geométrica de las Marcas de Diente. Aplicación a Yacimientos del Pleistoceno Inferior Ibérico: FN3, Venta Micena 3y 4 (Granada), Pontón de la Oliva (Patones, Madrid)”. This work is also funded by the Spanish Ministry of Science and Innovation with the call for funding “Proyectos Generación de Conocimiento 2021: Challenges and research opportunities in the Pleistocene Archaeological Sites of the Guadix-Baza Basin (Granada, Spain). Cuenca de Orce and Solano del Zamborino (ORCEpluSZ)” (Ref: PID2021-125098NB-I00).

L.A.C. is funded by the Spanish Ministry of Science, Innovation and Universities, with an FPI Predoctoral Grant (Ref. PRE2019-089411), associated to project RTI2018-099850-B-I00 and the University of Salamanca. D.H.R. was supported by the Ministry of Science, Innovation and Universities, under the contract REF: PEJ2019-005420-A as part of the i + D + I Garantía Juvenil.

## Author contributions

**Lloyd A. Courtenay** – Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, review and editing, visualization. **José Yravedra** – Formal analysis, investigation, data curation, resources, writing – original draft, review and editing, supervision, project administration, funding acquisition. **Darío Herranz-Rodrigo** – Data curation. **Juan José Rodríguez-Alba** – Writing – reviewing and editing. **Alexia Serrano-Ramos** – Data curation. **Verónica Estaca-Gómez** – Writing – reviewing and editing. **Diego González-Aguilera** – Resources, writing – reviewing and editing, supervision, funding acquisition. **José Antonio Solano** – Data curation, supervision. **Juan Manuel Jiménez-Arenas** – Resources, supervision, Project administration, funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data has been provided as supplementary materials

## Aknowledgements

We would like to thank all researchers, students and volunteers involved in the excavations, preparations and study of Barranco Léon. The corresponding author would also like to thank Rosa Huguet for her helpful comments and suggestions. Finally, the corresponding author is very grateful for the comments and suggestions made by Julia Aramendi to improve the quality of this line of research. We would also like to thank Jean-Philip Brugal, Antoine Souron, and one anonymous, reviewer for their constructive comments on earlier versions of this manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.quascirev.2022.107912.

## References

Alcántara-García, V., Barba Egido, R., Barral del Pino, J.M., Crespo Ruiz, A.B., Eiriz Vidal, A.I., Falquina Aparicio, Á., Herrero Calleja, S., Ibarra Jiménez, A., Megías González, M., Pérez Gil, M., Pérez Tello, V., Rolland Calvo, J., Yravedra, J., Vidal, A., Domíngez-Rodrigo, M., 2006. Determinación de procesos de fractura sobre huesos: un sistema de análisis de los ángulos de los planos de fracturación como discriminador de agentes bióticos. Trab. Prehist. 61 (1), 25–38.

Anadón, P., Gabás, M., 2009. Paleoenvironmental evolution of the early Pleistocene lacustrine sequence at Barranco León archeological site (Orce, Baza basin, southern Spain) from stable isotopes and Sr and Mg chemistry of ostracod shells. J. Paleolimnol. 42, 261–279. https://doi.org/10.1007/s10933-008-9275-6.

Anadón, P., Juliá, R., Oms, O., 2003. Estratigrafía y estudio sedimentológico preliminar de diversos afloramientos en Barranco León y Fuente Nueva (Orce, Granada). In: Toro, I., Agustí, J., Martínez, B. (Eds.), El Pleistoceno inferior de Barranco León y Fuente Nueva 3, Orce (Granada). Memoria científica Campañas 1999-2002, Monografías de Arqueología, vol. 17. Junta de Andalucía. Consejería de Cultura, pp. 47–72.

Anadón, P., Oms, O., Violeta, R., Ramon, J., 2015. The geochemistry of biogenic carbonates as a paleoenvironmental tool for the Lower Pleistocene Barranco León sequence (BL-5D, Baza Basin, Spain). Quat. Int. 389, 70–83.

Andrés, M., Gidna, A.O., Yravedra, J., Domínguez-Rodrigo, M., 2012. A study of dimensional differences of tooth marks (pits and scores) on bones modified by small and large carnivores. Archaeological and Anthropological Sciences 4 (3), 209–219. https://doi.org/10.1007/s12520-012-0093-4.

Antón, M., 2013. Sabertooth. Indiana University Press.

Antón, M., Galobart, A., Turner, A., 2005. Co-existence of scimitar-toothed cats, lions and hominins in the European Pleistocene. Implications of the post-cranial anatomy of *Homotherium latidens* (Owen) for comparative palaeoecology. Quat. Sci. Rev. 24 (10–11), 1287–1301. https://doi.org/10.1016/j.quascirev.2004.09.008.

Antón, M., Salesa, M.J., Galobart, A., Tseng, Z.J., 2014. The Plio-Pleistocene scimitar-toothed felid genus *Homotherium* Fabrini, 1890 (Machairodontinae, Homotherini): diversity, palaeogeography and taxonomic implications. Quat. Sci. Rev. 96, 259–268. https://doi.org/10.1016/j.quascirev.2013.11.022.

Aramendi, J., Maté-González, M.A., Yravedra, J., Ortega, M.C., Arriaza, M.C., González-Aguilera, D., Baquedano, E., Domínguez-Rodrigo, M., 2017. Discerning carnivore agency through the three-dimensional study of tooth pits: revisiting crocodile feeding behaviour at FLK- Zinj and FLK NN3 (Olduvai Gorge, Tanzania). Palaeogeogr. Palaeoclimatol. Palaeoecol. 488, 93–102. https://doi.org/10.1016/j.palaeo.2017.05.021.

Arriaza, M.C., Aramendi, J., Maté-González, M.A., , Yravedra, J. D., Strantdford, D., 2019. Characterising leopard as taphonomic agent throught the use of Micro-photogrammetric reconstruction of tooth marks and pit to score ratio. Hist. Biol. https://doi.org/10.1080/08912963.2019.1598401.

Arriaza, M.C., Aramendi, J., Maté-González, M.A., Yravedra, J., Stratford, D., 2021. The hunted or the scavenged? Australopith accumulation by brown hyenas at Sterkfontein (South Africa). Quat. Sci. Rev. 273, 107252. https://doi.org/10.1016/j.quascirev.2021.107252.

Barsky, D., Celiberti, V., Cauche, D., Grégoire, S., Lebègue, F., Lumley, H., Toro Moyano, I., 2010. Raw material discernment and technological aspects of the Barranco León and Fuente Nueva 3 stone assemblages (Orce southern Spain). Quat. Int. 223–224, 201–219. https://doi.org/10.1016/j.quaint.2009.12.004.

Barsky, D., Vergès, J.M., Sala, R., Menendez, L., Toro-Moyano, I., 2015a. Limestone percussion tolos from the late early Pleistocene sites of Barranco León and Fuente Nueva 3 (Orce, Spain). Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 1682. https://doi.org/10.1098/rstb.2014.0352.

Barsky, D., Sala, R., Menéndez, L., Toro-Moyano, I., 2015b. Use and re-use: Re-knapped flakes from the mode 1 site of Fuente Nueva 3 (Orce, Andalucía, Spain). Quat. Int. 361, 21–33. https://doi.org/10.1016/j.quaint.2014.01.048.

Bearder, S.K., 1977. Feeding habits of spotted hyaenas in a woodland habitat. East Afr. Wildl. J. 15, 263–280.

Benjamin, D.J., Berger, J.O., 2019. Three recommendations for improving the use of p-values. Am. stat. 73, 186–191.

Binford, L.R., 1981. Bones: Ancient Men and Modern Myths. Academic Press, Inc., New York.

Blain, H.A., Bailon, S., Agustí, J., Martínez-Navarro, B., Isidro, T., 2011. Paleoenvironmental and paleoclimatic proxies to the Early Pleistocene hominids of Barranco León D and Fuente Nueva 3 (Granada, Spain) by means of their amphibian and reptile assemblages. Quat. Int. 243 (1), 44–53.

Blain, H.A., Lozano-Fernández, I, Agustí, J., Bailon, S., Menéndez Granda, L., Espígares Ortiz, Ros-Montoya, S., Jiménez Arenas, J.M., Toro-Moyano, I., Martínez-Navarro, B., Sala, R., 2016. Refining upon the climatic background of the Early Pleistocene hominid settlement in western Europe: Barranco León and Fuente Nueva-3 (Guadiz-Baza Basin, SE Spain). Quat. Sci. Rev. 144, 132–144.

Blumenschine, R.J., 1995. Percussion marks, tooth marks, and experimental determinations of the timing of hominid and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania. J. Hum. Evol. 29, 21e51.

Blumenschine, R.J., Marean, C.W., Capaldo, S.D., 1996. Blind tests of interanalyst correspondence and accuracy in the identification of cut marks, percussion marks, and carnivore tooth marks on bone surfaces. J. Archeol. Sci. 23, 493e507.

Bookstein, F.L., 1989. Principal warps: thin-Plate Splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. 11 (6), 567–585.

Bookstein, F.L., 1991. Morphometric Tools for Landmark Data. Cambridge University Press, Cambridge.

Bookstein, F.L., 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Med. Image Anal. 1, 225–243.

Bourguignon, L., Crochet, J.-Y., Capdevila, R., Ivorra, J., Antoine, P.-O., Agustí, J., Barsky, D., Blain, H.-A., Boubles, N., Bruxelles, L., Claude, J., Cochard, D., Filoux, A., Firmat, C., Lozano-Fernández, I., Magniez, P., Pelletier, M., Rios-Garaizar, J., Testu, A., Valensi, P., De Weyer, L., 2016. Bois-de-Riquet (Lézignan-la-Cèbe, Hérault): a late Early Pleistocene archeological occurrence in southern France. Quat. Int. 393, 24–40. https://doi.org/10.1016/j.quaint.2015.06.037.

Brain, C.K., 1981. Hunters or the Hunted? an Introduction to African Cave Taphonomy. University of Chicago Press, Chicago.

Bunn, H.T., 1982. Meat Eating and Human Evolution: Studies on the Diet and Subsistence Patterns of Plio-Pleistocene Hominids in East Africa. PhD Thesis. University of California, Berkeley.

Bunn, H.T., Pickering, T.R., 2010. Methodological recommendations for ungulate mortality analyses in paleoanthropology. Quat. Res. 74, 388–394.

Cheheb, R.C., Arzarello, M., Arnaud, J., Berto, C., Cáceres, I., Caracausi, S., Colopi, F., Daffara, S., Canini, G.M., Huguet, R., Karambatsou, T., Benedetto, S., Zambaldi, M., Berruti, G.L.F., 2019. Human behaviour and Homo-mammal interactions at the first European peopling: new evidence from the Pirro Nord site (Apricena, Southern Italy). Sci. Nat. 106, 16. https://doi.org/10.1007/s00114-019-1610-4.

Colquhoun, D., 2019. The False Positive Risk: a proposal concerning what to do about p-values. Am. Statistician 73, 192–201.

Cooper, S.M., 1990. The hunting behaviour of spotted hyaenas (*Crocuta crocuta*) in a region containing both sedentary and migratory populations of herbivores. Afr. J. Ecol. 28 (2), 131–141. https://doi.org/10.1111/j.1365-2028.1990.tb01145.x.

Courtenay, L.A., González-Aguilera, D., 2020. Geometric morphometric data augmentation using generative computational learning algorithms. Appl. Sci. 10 (24), 9133. https://doi.org/10.3390/app10249133.

Courtenay, L.A., Yravedra, J., Huguet, R., Aramendi, J., Maté-González, M.Á., González-Aguilera, D., Arriaza, M.C., 2019. Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks. Palaeogeogr. Palaeoclimatol. Palaeoecol. 522, 28–39. https://doi.org/10.1016/j.palaeo.2019.03.007.

Courtenay, L.A., Yravedra, J., Maté-González, M.Á., Vázquez-Rodríguez, J.M., Fernández-Fernández, M., González-Aguilera, D., 2020a. The effects of prey size on carnivore tooth mark morphologies on bone; the case study of *Canis lupus signatus*. Hist. Biol. 33 (11), 2760–2772. https://doi.org/10.1080/08912963.2020.1827239.

Courtenay, L.A., Herranz-Rodrigo, D., Huguet, R., Maté-González, M.Á., González-Aguilera, D., Yravedra, J., 2020b. Obtaining new resolutions in carnivore tooth pit morphological analyses: a methodological update for digital taphonomy. PLoS One 15 (10), e0240328. https://doi.org/10.1371/journal.pone.0240328.

Courtenay, L.A., Herranz-Rodrigo, D., González-Aguilera, D., Yravedra, J., 2021a. Developments in data science solutions for carnivore tooth pit classification. Sci. Rep. 11, 10209. https://doi.org/10.1038/s41598-021-89518-4.

Courtenay, L.A., Herranz-Rodrigo, D., Yravedra, J., Vázquez-Rodríguez, J., Huguet, R., Barja, I., Maté-González, M.A., Fernández, M., González-Aguilera, D., 2021b. Effects of captivity on carnivore implications in ecological studies of both the past and present. Animals 11, 2323. https://doi.org/10.3390/ani11082323.

Courtenay, L.A., González-Aguilera, D., Lagüela, S., del Pozo, S., Ruiz-Mendez, C., Barbero-García, I., Román-Curto, C., Cañueto, J., Santos-Durán, C., Cardeñoso-Álvarez, M.E., Roncero-Riesco, M., Hernandez-López, D., Guerrero-Sevilla, D., Rodríguez-Gonzalvez, P., 2021c. Hyperspectral imaging and robust statistics in non-melanoma skin cancer analysis. Biomed. Opt Express 12 (8), 5107–5127.

Domínguez-Rodrigo, M., 1997. Meat-eating by early Hominids at the FLK 22 Zinjanthropus Site, Olduvai Gorge, Tanzania: an experimental approach using cut mark data. J. Hum Evol. 33 (6), 669–690.

Domínguez-Rodrigo, M., 1999. Flesh availability and bone modifications in carcasses consumed by lions, palaeoecological relevance in hominind foraging patterns. Palaeogeogr. Palaeoclimatol. Palaeoecol. 149, 373–388.

Domínguez-Rodrigo, M., Barba, R., 2006. New estimates of tooth mark and percussion mark frequencies at the FLK Zinj site: the carnivore-hominid-carnivore hypothesis falsified. J. Hum. Evol. 50 (2), 170–194.

Domínguez-Rodrigo, M., Barba, R., Egeland, C., 2007. Deconstructing Olduvai. Springer, The Netherlands.

Domínguez-Rodrigo, M., Gidna, A.O., Yravedra, J., Musiba, C., 2012. A comparative neo-taphonomic study of felids, hyaenids and canids: an analogical framework based on long bone modification patterns. J. Taphon. 10 (3), 147–164.

Domínguez-Rodrigo, M., Yravedra, J., Organista, E., Gidna, A., Fourvel, J.-B., Baquedano, E., 2015. A new methodological approach to the taphonomic study of paleontological and archaeological faunal assemblages: a preliminary case study from Olduvai Gorge (Tanzania). J. Archaeol. Sci. 59, 35–53. https://doi.org/10.1016/j.jas.2015.04.007.

Domínguez-Rodrigo, M., Egeland, C.P., Cobo-Sánchez, L., Baquedano, E., Hulbert, R.C., 2022. Sabertooth carcass consumption behavior and the dynamics of Pleistocene large carnivoran guilds. Sci. Rep. 12 (1), 6045. https://doi.org/10.1038/s41598-022-09480-7.

Echassoux, A., 2004. Étude taphonomique, paléoécologique et archéozoologique des faunes de grands mammifères de la seconde moitié du Pléistocène inférieur de la grotte du Vallonet (Roquebrune-Cap-Martin, Alpes-Maritimes, France). L'anthropologie. 108, 11–53.

Espigares, M.P., Martínez-Navarro, B., Palmqvist, P., Ros-Montoya, S., Toro, I., Agustí, J., Sala, R., 2013. *Homo* vs. *Pachycrocuta*: earliest evidence of competition for an elephant carcass between scavengers at Fuente Nueva-3 (Orce, Spain). Quat. Int. 295, 113–125. https://doi.org/10.1016/j.quaint.2012.09.032.

Espigares, M.P., Palmqvist, P., Guerra-Merchán, A., Ros-Montoya, S., García-Aguilar, J.M., Rodríguez-Gómez, G., Serrano, F.J., Martínez-Navarro, B., 2019. The earliest cut marks of Europe: a discussion on hominin subsistence patterns in the Orce sites (Baza basin, SE Spain). Sci. Rep. 9, 15408. https://doi.org/10.1038/s41598-019-51957-5.

Estes, R.D., Goddard, J., 1967. Prey selection and hunting behaviour of the african wild dog. J. Wildl. Manag. 31, 52–69. https://doi.org/10.2307/3798360.

Fourvel, J.P., Magniez, P., Moigne, A.M., Testu, A., Joris, A., Lamglait, B., Vaccaro, C., Fosse, P., 2018. Wild dogs and their relatives: implication of experimental feedings in their taphonomical identification. Quaternaire 29 (1), 21–29.

https://doi.org/10.4000/quaternaire.8578.

Garrido, G., 2008. El registro de *Vulpes alopecoides* (Forsyth-Major, 1877), *Canis etruscus* (Forsyth-Major, 1877) y *Canis* cf. *Falconeri* (Forsyh-Major, 1877) (Canidae, carnívora, mammalia) en Fonelas P-1 (Cuenca de Guadix, Granada). In: Arribas, A. (Ed.), Vertebrados del Plioceno superior terminal en el suroeste de Europa: Fonelas P-1 y el Proyecto Fanals. Cuadernos del Museo Geominero, nº10. Instituto Geológico y Minero de España, Madrid.

Gibert, J., Gibert, L., Iglesias, A., Maestro, E., 1998. Two 'Oldowan' assemblages in the Plio-Pleistocene deposits of the Orce region, southeast Spain. Antiquity 72, 17–25. https://doi.org/10.1017/S0003598X00086233.

Gidna, A.J., Yravedra, J., Domínguez-Rodrigo, M., 2013. A cautionary note on the use of captive carnivores to model wild predator behaviour: a comparison of bone modification patterns on long bones by captive and wild lions. J. Archaeol. Sci. 40, 1903–1910.

Gittleman, J.L., 1985. Carnivore body size: ecological and taxonomic correlates. Oecologia (Berl.) 67 (4), 540–554. https://doi.org/10.1007/BF00790026.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge.

Gunz, P., Mitteroecker, P., 2013. Semilandmarks: a method for quantifying curves and surfaces. Hystrix it. J. Mammal. 24, 103–109. https://doi.org/10.4404/hystrix-24.1-6292.

Haynes, G., 1980. Evidence of carnivore gnawing on Pleistocene and recent mammalian bones. Paleobiology 6 (3), 341–351.

Haynes, G., 1983. A guide for differentiating mammalian carnivore taxa responsible for gnaw damage to herbivore limb bones. Paleobiology 9 (2), 164–172.

Henshel, J.R., 1986. The Socio-Ecology of Spotted hyaena *Crocuta crocuta* in the Kruger National Park. Ph.D. thesis, University of Pretoria.

Höhle, J., Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS J. Photogrammetry Remote Sens. 64, 398–406. https://doi.org/10.1016/j.isprsjprs.2009.02.003.

Huguet, R., Saladié, P., Cáceres, I., Díez, C., Rosell, J., Bennàsar, M., Blasco, R., Esteban-Nadal, M., Gabucio, M.J., Rodríguez-Hidalgo, A., Carbonell, E., 2013. Successful subsistence strategies of the first humans in south-western Europe. Quat. Int. 295, 168–182.

Kruuk, H., 1972. The Spotted Hyena: A Study of Predation and Social Behavior. University of Chicago Press, Chicago.

Kuhn, B.F., Berger, L.R., Skinner, J.D., 2009. Variation in tooth mark frequencies on long bones from the assemblages of all three extant bone-collecting hyaenids. J. Archaeol. Sci. 36 (2), 297–307. https://doi.org/10.1016/j.jas.2008.09.008.

Lakens, D., 2017. Equivalence tests: a pratical primer for T tests, correlations and meta analyses. Society of Psychological and Personality Sciences 8 (4), 355–362. https://doi.org/10.1177/1948550617697177.

Lopez-Fernandez, L., Rodriguez-Gonzalvez, P., Hernandez-Lopez, D., Ortega-Terol, D., González-Aguilera, D., 2017. Comparative analysis of triangulation libraries for modeling large point clouds from land and their infrastructures. Infrastructure 2 (1), 1–11. https://doi.org/10.3390/infrastructures2010001.

Lozano, S., Mateos, A., Rodríguez, J., 2016. Exploring paleo food-webs in the European Early and Middle Pleistocene: a network analysis. Quat. Int. 413, 44–54. https://doi.org/10.1016/j.quaint.2015.10.068.

Lucenti, S.B., Madurell-Malapeira, J., 2020. Unraveling the fossil record of foxes: an updated review on the Plio-Pleisotcene *Vulpes* spp. from Europe. Quat. Int. 236, 106296. https://doi.org/10.1016/j.quascirev.2020.106296.

Luzón, C., Yravedra, J., Courtenay, L.A., Saarinen, J., Blain, H.-A., DeMiguel, D., Viranta, S., Azanza, B., Rodríguez-Alba, J.J., Herranz-Rodrigo, D., Serrano-Ramos, A., Solano, J.A., Oms, O., Agustí, J., Fortelius, M., Jiménez-Arenas, J.M., 2021. Taphonomic and spatial analyses from the early Pleistocene site of Venta Micena 4 (Orce, Guadix-Baza Basin, southern Spain). Sci. Rep. 11 (1). https://doi.org/10.1038/s41598-021-93261-1.

Madurell-Malapeira, J., Martínez-Navarro, B., Ros-Montoya, S., Patrocinio Espigares, M., Toro, I., Palmqvist, P., 2011. The earliest European badger (*Meles meles*), from the late villafranchian site of Fuente Nueva 3 (Orce, Granada, SE iberian peninsula). Comptes Rendus Palevol 10 (8), 609–615. https://doi.org/10.1016/j.crpv.2011.06.001.

Malcolm, J.R., Van Lawick, B., 1975. Notes on wild dogs (*Lycaon pictus*) hunting zebras. Mammalia 39, 231–240. https://doi.org/10.1515/mamm.1975.39.2.231.

Marean, C.W., 1989. Sabertooth cats and their relevance for early hominid diet and evolution. J. Hum. Evol. 18, 559–582. https://doi.org/10.1016/0047-2484(89)90018-3.

Marean, C.W., Ehrhardt, C.L., 1995. Palaeanthropological and palaeoecological implications of the taphonomy of a sabertooth's den. J. Hum. Evol. 29, 515–547. https://doi.org/10.1006/jhev.1995.1074.

Martínez-Monzón, A., Sánchez-Bandera, C., Fagoaga, A., Oms, O., Agustí, J., Barsky, D., Solano-García, J., Jiménez-Arenas, J.M., Blain, H.A., 2021. Amphibian body size and species richness as a proxy for primary productivity and climate: the Orce wetlands (Early Pleistocene, Guadix-Baza Basin, SE Spain). Palaeogeogr. Palaeoclimatol. Palaeoecol. 110752. https://doi.org/10.1016/J.PALAEO.2021.110752.

Martínez-Navarro, B., Palmqvist, P., Madurell-Malapiera, J., Ros-Montoya, S., Espigares, M.P., Torregrosa, V., Pérez-Claros, J.A., 2010. La fauna de grandes mamíferos de Fuente Nueva 3 y Barranco León 5 – estado de la Cuestión. In: Martínez-Navarro: Ocupaciones Humanas en el Pleistoceno Inferior y Medio de la Cuenca de Guadix-Baza. Junta de Andalucía, Consejería de Cultura, pp. 197–236.

Maté-González, M.A., Aramendi, J., Yravedra, J., González-Aguilera, D., 2017. Statistical comparison between low-cost methods for 3D characterization of cutmarks on bones. Rem. Sens. 9 (9), 873. https://doi.org/10.3390/rs9090873.

Mattson, D.J., 1997. Use of ungulates by Yellowstone grizzly bears. Biol. Conserv. 81, 161–177. https://doi.org/10.1016/S0006-3207(96)00142-5.

Medin, T., Martínez-Navarro, B., Rivals, F., Madurell-Malapeira, J., Ros-Montoya, S., Espigares, M.P., Figueirido, B., Rook, L., Palmqvist, P., 2017. Late Villafranchian *Ursus etruscus* and other large carnivorans from the Orce sites (Guadix-Baza basin, Andalusia, Southern Spain): taxonomy, biochronology, paleobiology, and ecogeographical context. Quat. Int. 431, 20–41. https://doi.org/10.1016/j.quaint.2015.10.053.

Mills, M.G.L., 1984a. Prey selection and feeding habits of the large carnivores in the Southern Kalahari. Kuedue (suppl.) 21, 281–294. https://doi.org/10.4102/koedoe.v27i2.586.

Mills, M.G.L., 1984b. The comparative behavioural ecology of the brown hyaena *Hyaena brunnea* and the spotted hyaena *Crocuta crocuta* in the Southern Kalahari. Koedoe 27, 237–247. https://doi.org/10.4102/koedoe.v27i2.583.

Moclán, A., Domínguez-Rodrigo, M., Yravedra, J., 2019. Classifying agency in bone breakage: an experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using machine learning (ML) algorithms. Archaeol. Anthropol. Sci. 11, 4663–4680. https://doi.org/10.1007/s12520-019-00815-6.

Mora, R., Aramendi, J., Courtenay, L.A., González-Aguilera, D., Yravedra, J., Maté-González, M.Á., Prieto-Herráez, D., Vázquez-Rodríguez, J.M., Barja, I., 2022. *Ikhnos*: a novel software to register and analyze bone surface modiufications based on three-dimensional documentation. Animals 12 (20), 2861. https://doi.org/10.3390/ani12202861.

Oms, O., Pares, J.M., Martínez-Navarro, B., Agustí, J., Toro, I., Martínez-Fernandez, G., Turq, A., 2000. Early human occupation of Western Europe: paleomagnetic dates for two paleolithic sites in Spain. P. Natl. Acad. Sci. USA 97, 10666–10670. https://doi.org/10.1073/pnas.180319797.

Oms, O., Anadón, P., Agustí, J., Julià, R., 2011. Geology and chronology of the continental Pleistocene archeological and mammal sites of the Orce area (Baza Basin, Spain). Quat. Int. 243, 33–43. https://doi.org/10.1016/j.quaint.2011.03.048.

Pagès, J., 2004. Analyse factorielle de Données mixtes. Rev. Stat. Appl. 4, 93–111.

Palmqvist, P., Arribas, A., Gröcke, D.R., 2002. The early Pleistocene locality at Venta Micena (Orce, Guadix-Baza Basin, southeast Spain): remarks on the taphonomy, biogeochemistry and paleoecology of the large mammals assemblage. Pliocenica 2, 126–150.

Palmqvist, P., Arribas, A., Martínez-Navarro, B., 2007a. Ecomorphological study of large canids from the lower Pleistocene of southeastern Spain. Lethaia 32 (1), 75–88. https://doi.org/10.1111/j.1502-3931.1999.tb00583.x.

Palmqvist, P., Torregrosa, V., Pérez-Claros, J.A., Martínez-Navarro, B., Turner, A., 2007b. A re-evaluation of the diversity of *Megantereon* (Mammalia, Carnivora, Machairodontinae) and the problem of species identification in extinct carnivores. J. Vertebr. Paleontol. 27 (1), 160–175. https://doi.org/10.1671/0272-4634(2007)27[160:AROTDO]2.0.CO;2.

Palmqvist, P., Pérez-Claros, J.A., Janis, C.M., Gröcke, D.R., 2008. Tracing the ecophysiology of ungulates and predator–prey relationships in an early Pleistocene large mammal community. Palaeogeogr. Palaeoclimatol. Palaeoecol. 266 (1–2), 95–111. https://doi.org/10.1016/j.palaeo.2008.03.015.

Palmqvist, P., Martínez-Navarro, B., Pérez-Claros, J.A., Torregrosa, V., Figueirido, B., Jiménez-Arenas, J.M., Espigares, M.P., Ros-Montoya, S., De Renzi, M., 2011. The giant hyena *Pachycrocuta brevirostris*: modelling the bone-cracking behavior of an extinct carnivore. Quat. Int. 243 (1), 61–79. https://doi.org/10.1016/j.quaint.2010.12.035.

Parkinson, J.A., 2018. Revisiting the hunting-versus-scavenging debate at FLK-Zinj: a GIS spatial analysis of bone modifications produced by hominins and carnivores in the FLK 22 assemblage, Olduvai Gorge, Tanzania. Palaeogeogr. Palaeoclimatol. Palaeoecol. 511, 29–51.

Parkinson, J.A., Plummer, T.W., Bose, R., 2014. A GIS-based approach to documenting large canid damage to bones. Palaeogeogr. Palaeoclimatol. Palaeoecol. 409, 57–71.

Parkinson, J.A., Plummer, T.W., Harston-Rose, A., 2015. Characeterizing felid tooth marking and gross bone damage patterns using GIS image analysis: an experimental feeding study with large felids. J. Hum. Evol. 80, 114–134.

Parkinson, J.A., Plummer, T.W., Oliver, J.S., Bishop, L.C., 2022. Meat on the menu: GIS spatial distribution analysis of bone surface damage indicates that Oldowan hominins at Kanjera South, Kenya, had early access to carcasses. Quat. Sci. Rev. 277, 107314.

Périquet, S., Fritz, H., Revilla, E., 2015. The lion king and the hyaena queen: large carnivore interactions and coexistence. Biol. Rev. 90 (4), 1197–1214.

Pickering, T.R., Egeland, C.P., 2006. Experimental patterns of hammerstone percussion damage on bones: implications for inferences of carcass processing by humans. J. Archaeol. Sci. 33 (4), 459–469. https://doi.org/10.1016/j.jas.2005.09.001.

Pineda, A., Saladié, P., Huguet, R., Cáceres, I., Rosas, A., García-Tabernero, A., Estalrrich, A., Mosquera, M., Ollé, A., Vallverdú, J., 2015. Coexistence among large predators during the lower paleolithic at the site of La mina (Barranc de la Boella, tarragona, Spain). Quat. Int. 388, 177–187.

Pineda, A., Saladié, P., Huguet, R., Cáceres, I., Rosas, A., Estalrrich, A., García-Tabernero, A., Vallverdú, J., 2017. Changing competition dynamics among predators at the late Early Pleistocene site Barranc de la Boella (Tarragona, Spain). Paleogeog. Palaeoclimat. Palaeoecol. 477, 10–26.

Rahimi, A., Recht, B., 2007. Random features for large-scale kernel machines. In: Proceedings of the International Conference of Neural Information Processing Systems, 20, pp. 1–8. https://doi.org/10.5555/2981562.2981710.

Rhodes, R., Rhodes, G., 2004. Prey selection and use of natural and man-made barriers by African wild dogs while hunting. S. Afr. J. Wildl. Res. 34, 135–142.

Rodríguez, J., Rodríguez-Gómez, G., Martín-González, J.A., Goikoetxea, I., Mateos, A., 2012. Predator–prey relationships and the role of *Homo* in Early Pleistocene food webs in Southern Europe. Palaeogeogr. Palaeoclimatol. Palaeoecol. 365–366, 99–114. https://doi.org/10.1016/j.palaeo.2012.09.017.

Rodríguez-Gómez, G., Palmqvist, P., Rodríguez, J., Mateos, A., Martín-González, J.A., Espigares, M.P., Ros-Montoya, S., Martínez-Navarro, B., 2016. On the ecological context of the earliest human settlements in Europe: resource availability and competition intensity in the carnivore guild of Barranco León-D and Fuente Nueva-3 (Orce, Baza Basin, SE Spain). Quat. Sci. Rev. 143, 69–83. https://doi.org/10.1016/j.quascirev.2016.05.018.

Rodríguez-Martín, M., Rodríguez-González, P., Ruiz de Oña Crespo, E., González-Aguilera, D., 2019. Validation of portable mobile mapping system for inspection tasks in thermal and fluid-mechanical facilities. Rem. Sens. 11 (19), 2205. https://doi.org/10.3390/rs11192205.

Rohlf, F.J., 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. Syst. Biol. 47, 147–158. https://doi.org/10.1080/106351598261094.

Rohlf, F.K., 1999. Shape statistics: Procrustes superimpositions and tangent spaces. J. Classif. 16, 197–223. https://doi.org/10.1007/s003579900054.

Ros-Montoya, S., Bartolini-Lucenti, S., Espigares, M.P., Palmqvist, P., Martínez-Navarro, B., 2021. First review of lyncodontini material (Mustelidae, carnivora, mammalia) from the lower Pleistocene archaeo-palaeontological sites of Orce (southeastern Spain). Riv. Ital. Paleontol. Stratigr. 127, 33–47.

Saarinen, J., Oksanen, O., Žliobaitė, I., Fortelius, M., DeMiguel, D., Azanza, B., Bocherens, H., Luzón, C., Solano-Garcí, J., Yravedra, J., Courtenay, L.A., Blain, H.A., Sánchez-Bandera, C., Serrano-Ramos, A., Rodríguez-Alba, J.J., Viranta, S., Barsky, D., Tallavaara, M., Oms, O., Agustí, J., Ochando, J., Carrión, J.S., Jiménez-Arenas, J.M., 2021. Pliocene to Middle Pleistocene climate history in the Guadix-Baza Basin, and the environmental conditions of early *Homo* dispersal in Europe. Quat. Sci. Rev. 268, 107132. https://doi.org/10.1016/j.quascirev.2021.107132.

Saladié, P., Rodríguez-Hidalgo, A., Huguet, R., Cáceres, I., Díez, C., Vallverdú, J., Canals, A., Soto, M., Santander, B., Bermúdez de Castro, J.M., Arsuaga, J.L., Carbonell, E., 2014. The role of carnivorans and their relationship to hominin stellements in the TD6-2 level from Gran Dolina (Sierra de Atapuerca, Spain). Quat. Sci. Rev. 93, 47–66.

Sánchez-Bandera, C., Oms, O., Blain, H.-A., Lozano-Fernández, I., Bisbal-Chinesta, J.F., Agustí, J., Saarinen, J., Fortelius, M., Titton, S., Serrano-Ramos, A., Luzón, C., Solano-García, J., Barsky, D., Jiménez-Arenas, J.M., 2020. New stratigraphically constrained palaeoenvironmental reconstructions for the first human settlement in western Europe: the early Pleistocene herpetofaunal assemblages from Barranco León and Fuente Nueva 3 (Granada, SE Spain). Quat. Sci. Rev. 243, 106466. https://doi.org/10.1016/j.quascirev.2020.106466.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., Freitas, N., 2016. Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE 104 (1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218.

Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical Bayesian optimization of machine learning algorithms. Proc. Int. Conf. Neur. Inf. Process. Syst. 25, 2951–2959.

Sokal, R.R., Rohlf, F.J., 1981. Biometry: the Principles and Practice of Statistics in Biological Research. Freeman, New York.

Tancik, M., Srinivasan, P., Milenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R., 2020. Fourier features let networks learn high frequency functions in low dimensional domains. Proc. Int. Conf. Neur. Inf. Process. Syst. arXiv: 2006.10739v1.

Tappen, M., Lordkipanidze, D., Bukshianidze, M., Ferring, R., Vekua, A., 2007. Are you in or out (of Africa)? Site formation at Dmanisi and actualistic studies in Africa. In: Pickering, T.R., Schick, K., Toth, N. (Eds.), Breathing Life into Fossils: Taphonomic Studies in Honor of C.K. Brain. Bloomington. Stone Age Institution Press, pp. 119–135.

Tappen, M., Bukhsianidez, M., Ferring, R., Coil, R., Lordkipanidze, D., 2022. Life and death at Dmanisi, Georgia: taphonomic signals from the fossil mammals. J. Hum. Evol. 171, 103249. https://doi.org/10.1016/j.jhevol.2022.103249.

Tilson, R.L., Henschel, J.R., 1986. Spatial arrangement of spotted hyaena groups in a desert environment, Namibia. Afr. J. Ecol. 24 (3), 173–180. https://doi.org/10.1111/j.1365-2028.1986.tb00358.x.

Titton, S., Barsky, D., Bargallo, A., Vergès, J.M., Guardiola, M., Solano, J.G., Jimenez Arenas, J.M., Toro-Moyano, I., Sala-Ramos, R., 2018. Active percussion tools from the Oldowan site of Barranco León (Orce, Andalusia, Spain): the fundamental role of pounding activities in hominin lifeways. J. Archaeol. Sci. 96, 131–147. https://doi.org/10.1016/j.jas.2018.06.004.

Titton, S., Barsky, D., Bargalló, A., Serrano-Ramos, A., Vergès, J.M., Toro-Moyano, I., Sala-Ramos, R., García-Solano, J.G., Jimenez Arenas, J.M., 2020. Subspheroids in the lithic assemblage of Barranco León (Spain): recognizing the late oldowan in Europe. PLoS One 15 (1), e0228290. https://doi.org/10.1371/journal.pone.0228290.

Titton, S., Oms, O., Barsky, D., Bargalló, A., Serrano-Ramos, A., García-Solano, J., Sánchez-Bandera, C., Yravedra, J., Blain, H.-A., Toro-Moyano, I., Jiménez Arenas, J.M., Sala-Ramos, R., 2021. Oldowan stone knapping and percussive activities on a raw material reservoir deposit 1.4 million years ago at Barranco León (Orce, Spain). Archaeol. Anthropol. Sci. 13, 108. https://doi.org/10.1007/s12520-021-01353-w.

Toro-Moyano, I., Lumley, H. de, Fajardo, B., Barsky, D., Celiberti, V., Grégoire, S., Martínez-Navarro, B., Espigares, M.P., Ros-Montoya, S., 2009. L'industrie lithique des gisements du Pléistocène inférieur de Barranco León et Fuente Nueva 3, Granade, Espagne. L'Anthropologie 113, 111–124. https://doi.org/10.1016/j.anthro.2009.01.006.

Toro-Moyano, I., Martínez-Navarro, B., Agustí, J., 2010. Ocupaciones Humanas en el Pleistoceno inferior y medio de la cuenca de Guadix-Baza. Memoria Científica. Junta de Andalucía, Consejería de Cultura. EPG Arqueología Monografias.

Toro-Moyano, I., Barsky, D., Cauche, D., Celiberti, V., Grégoire, S., Lebegue, F., Moncel, M.H., Lumley, H., 2011. The archaic stone-tool industry from Barranco León and Fuente Nueva 3, (Orce, Spain): evidence of the earliest hominin presence in southern Europe. Quat. Int. 243, 80–91. https://doi.org/10.1016/j.quaint.2010.12.011.

Toro-Moyano, I., Martínez-Navarro, B., Agustí, J., Souday, C., Bermúdez de Castro, J.M., Martinón-Torres, M., Fajardo, B., Duval, M., Falguères, C., Oms, O., Parés, J.M., Anadón, P., Julià, R., García-Aguilar, J.M., Moigne, A.M., Espigares, M.P., Ros-Montoya, S., Palmqvist, P., 2013. The oldest human fossil in Europe, from Orce (Spain). J. Hum. Evol. 65, 1–9. https://doi.org/10.1016/j.jhevol.2013.01.012.

Turner, A., 1992. Large carnivores and earliest European hominids: changing determinants of resource availability during the Lower and Middle Pleistocene. J. Hum. Evol. 22 (2), 109–126. https://doi.org/10.1016/0047-2484(92)90033-6.

Turner, A., 1995. The Villafanchian large carnivore guild: geographic distribution and structural evolution. Il Quat. 8, 349–356.

Turner, A., Antón, M., 1996. The giant hyaena, *Pachycrocuta brevirostris* (mammalia, carnivora, Hyaenidae). Geobios 29 (4), 455–468. https://doi.org/10.1016/S0016-6995(96)80005-2.

Turner, A., Antón, M., 1997. The Big Cats and Their Fossil Relatives. Columbia University Press, New York.

Turq, A., Martínez Navarro, B., Palmqvist, P., Arribas, A., Agustí, J., Rodríguez Vidal, J., 1996. Le Plio-Pléistocène de la région d'Orce, province de Grenade, Espagne : bilan et perspectives de recherche. Paléo 8, 161–204.

Valtierra, N., Courtenay, L.A., López-Polín, L., 2020. Microscopic analyses of the effects of mechanical cleaning interventions on cut marks. Archaeological and Anthropological Sciences 12, 193. https://doi.org/10.1007/s12520-020-01153-8.

Vezina, A.F., 1985. Empirical Relationships between predator and prey size among terrestrial vertebrate predators. Oecologia (Berl.) 67 (4), 555–565. https://doi.org/10.1007/BF00790027.

Villa, P., Mahieu, E., 1991. Breakage patterns of human long bones. J. Hum. Evol. 21, 27–48. https://doi.org/10.1016/0047-2484(91)90034-S.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p < 0.05". Am. Statistician 73, 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Wiering, M.A., van der Ree, M.H., Embreschts, M.J., Stollenga, M.F., Meijster, A., Nolte, A., Schomaker, L. R. B Schomaker, 2013. The neural support vector machine. In: The 25th Benelux Artificial Intelligence Conference, 257–254.

Yravedra, J., 2005. Tafonomía Aplicada a Zooarqueología. Lerko Print, Madrid.

Yravedra, J., Lagos, L., Bárcena, F., 2011. A taphonomic study of wild wolf (*Canis lupus*) modification of horse bones in northwestern Spain. J. Taphonomy. 9, 37–65.

Yravedra, J., Andrés, M., Domínguez-Rodrigo, M., 2013. A taphonomic study of the African wild dog (*Lycaon pictus*). Archaeological and Anthropological Sciences 6 (2), 113–124. https://doi.org/10.1007/s12520-013-0164-1.

Yravedra, J., Maté-González, M.Á., Courtenay, L.A., González-Aguilera, D., Fernández, M.F., 2019. The use of canid tooth marks on bone for the identification of livestock predation. Sci. Rep. 9 (1). https://doi.org/10.1038/s41598-019-52807-0.

Yravedra, J., Solano, J.A., Courtenay, L.A., Saarinen, J., Linares-Matás, G., Luzón, C., Serrano-Ramos, A., Herranz-Rodrigo, D., Cámara, J.M., Ruiz, A., Titton, S., Rodríguez-Alba, J.J., Mielgo, C., Blain, H.A., Agustí, J., Sánchez-Bandera, C., Montilla, E., Toro-Moyano, I., Fortelius, M., Oms, O., Barsky, D., Jiménez-Arenas, J.M., 2021. Use of meat resources in the early Pleistocene assemblages from Fuente Nueva 3 (Orce, Granada, Spain). Archaeol. Anthropol. Sci. 13, 213. https://doi.org/10.1007/s12520-021-01461-7.

Yravedra, J., Solano, J.A., Herranz-Rodrigo, D., Linares-Matás, G.J., Saarinen, J., Rodríguez-Alba, J.J., Titton, S., Serrano-Ramos, A., Courtenay, L.A., Mielgo, C., Luzón, C., Cámara, J., Sánchez-Bandera, C., Montilla, E., Toro-Moyano, I., Barsky, D., Fortelius, M., Agusti, J., Blain, A.H., Oms, O., Jiménez-Arenas, J.M., 2022a. Unraveling hominin activities in the zooarchaeological assemblage of Barranco León (Orce, Granada, Spain). J. Palaeolithic Archaeol. 5, 6. https://doi.org/10.1007/s41982-022-00111-1.

Yravedra, J., Courtenay, L.A., Herranz-Rodrigo, D., Rodríguez-Alba, J.J., Linares-Matás, G., Estaca-Gómez, V., Luzón, C., Serrano-Ramos, A., Maté-González, M.Á., Solano, J.A., González-Aguilera, D., Jiménez-Arenas, J.M., 2022b. Taphonomic characterisation of extinct eurasion carnivores through geometric morphometrics. Sci. Bull. 67 (16), 1644–1648.

*Spanish Translation of Title and Abstract*

# Un nuevo enfoque para la caracterización morfológica de lesiones cutáneas de tipo cáncer no-melanoma mediante el análisis de Elípticos de Fourier e imágenes clinicas.

La detección precoz del cáncer de piel de tipo no-melanoma (CPNM) es crucial para conseguir los mejores resultados médicos. La morfología de las lesiones se considera uno de los principales parámetros para la identificación de algunos tipos de cáncer de piel, como el melanoma. En el caso del CPNM, la importancia de la forma como parámetro de detección visual no ha sido lo suficientemente investigada. En este estudio, se ha analizado un conjunto de 993 imágenes de diferentes tipos de lesiones de tipo CPNM, así como lesiones cutáneas benignas, obtenidas mediante una cámara estándar RGB. Para cada imagen, se extrajeron los límites de la lesión. Después, se calcularon los coeficientes morfológicos mediante un análisis de elípticos de Fourier (AEF), así como la asimetría para el límite de cada lesión. Estos datos fueron analizados mediante estadística multivariante, utilizando diferentes combinaciones de datos para una mayor reducción de dimensiones. Por último, se emplearon algoritmos de aprendizaje automático para evaluar la separación de los tipos de lesiones. Las tareas de clasificación se realizaron utilizando la asimetría, los coeficientes elípticos de Fourier, y la combinación de ambos. La separación entre lesiones malignas y benignas tuvo éxito en la mayoría de los casos. El enfoque que mejor funcionó fue el basado en la combinación de coeficientes elípticos de Fourier y la asimetría, que dio como resultado una precisión equilibrada de 0.786, y un Área Bajo la Curva (AUC) de 0.735. Los resultados obtenidos sugieren que la morfología, y, en particular, la combinación de coeficientes elípticos de Fourier y la asimetría, deberían integrarse como parámetro fundamental en futuras técnicas de detección.

*Supplementary Information and Links*

**Code available from:**

www.mdpi.com/article/10.3390/jcm11154392/s1

Journal of
*Clinical Medicine*

MDPI

*Article*

# A Novel Approach for the Shape Characterisation of Non-Melanoma Skin Lesions Using Elliptic Fourier Analyses and Clinical Images

**Lloyd A. Courtenay [1,*], Inés Barbero-García [1], Julia Aramendi [1,2], Diego González-Aguilera [1], Manuel Rodríguez-Martín [3], Pablo Rodríguez-Gonzalvez [4], Javier Cañueto [5,6,7] and Concepción Román-Curto [5,6]**

[1] Department of Cartographic and Terrain Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003 Ávila, Spain; ines.barbero@usal.es (I.B.-G.); juliaram@usal.es (J.A.); daguilera@usal.es (D.G.-A.)

[2] Deptartment of Geology, Facultad de Ciencia y Tecnología, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa, Spain

[3] Department of Mechanical Engineering, Universidad de Salamanca, 49029 Zamora, Spain; ingmanuel@usal.es

[4] Department of Mining Technology, Topography and Structures, University of León, 24401 Ponferrada, Spain; p.rodriguez@unileon.es

[5] Department of Dermatology, University Hospital of Spain, Paseo de San Vicente 58-182, 37007 Salamanca, Spain; javier.canueto@gmail.com (J.C.); croman@saludcastillayleon.es (C.R.-C.)

[6] Instituto de Investigación Biomédica de Salamanca (IBSAL), Paseo de San Vicente 58-182, 37007 Salamanca, Spain

[7] Instituto de Biología Molecular y Celular del Cáncer (IBMCC)/Centro de Investigación del Cáncer (Lab 7), Campus Miguel de Unamuno s/n, 37007 Salamanca, Spain

* Correspondence: ladc1995@gmail.com; Tel.: +34-633-647-825

**Citation:** Courtenay, L.A.; Barbero-García, I.; Aramendi, J.; González-Aguilera, D.; Rodríguez-Martín, M.; Rodríguez-Gonzalvez, P.; Cañueto, J.; Román-Curto, C. A Novel Approach for the Shape Characterisation of Non-Melanoma Skin Lesions Using Elliptic Fourier Analyses and Clinical Images. *J. Clin. Med.* **2022**, *11*, 4392. https://doi.org/10.3390/jcm11154392

Academic Editor: Moussa Antoine Chalah

Received: 3 July 2022
Accepted: 27 July 2022
Published: 28 July 2022

**Abstract:** The early detection of Non-Melanoma Skin Cancer (NMSC) is crucial to achieve the best treatment outcomes. Shape is considered one of the main parameters taken for the detection of some types of skin cancer such as melanoma. For NMSC, the importance of shape as a visual detection parameter is not well-studied. A dataset of 993 standard camera images containing different types of NMSC and benign skin lesions was analysed. For each image, the lesion boundaries were extracted. After an alignment and scaling, Elliptic Fourier Analysis (EFA) coefficients were calculated for the boundary of each lesion. The asymmetry of lesions was also calculated. Then, multivariate statistics were employed for dimensionality reduction and finally computational learning classification was employed to evaluate the separability of the classes. The separation between malignant and benign samples was successful in most cases. The best-performing approach was the combination of EFA coefficients and asymmetry. The combination of EFA and asymmetry resulted in a balanced accuracy of 0.786 and an Area Under Curve of 0.735. The combination of EFA and asymmetry for lesion classification resulted in notable success rates when distinguishing between benign and malignant lesions. In light of these results, skin lesions' shape should be integrated as a fundamental part of future detection techniques in clinical screening.

**Keywords:** Non-Melanoma Skin Cancer; Elliptic Fourier Analysis; Shape analysis; Skin lesion asymmetry; clinical images; computer vision

## 1. Introduction

Non-melanoma skin cancer (NMSC) is one of the most common malignancies, with an especially high incidence rate among elderly and white-skinned populations. NMSC includes different pathologies, with Basal Cell Carcinoma (BCC) and Squamous Cell Carcinoma (SCC) being the most common. The early detection and diagnosis of NMSC reduces the risk of bad prognoses, as well as the costs these pathologies entail on health systems due to their high incidence [1,2].

The first step in skin cancer diagnosis, including melanoma and NMSC, is a visual examination. For the detection of melanoma, the ABCDE mnemonic is a widely used tool [3,4], which considers the following variables: Asymmetry, Border irregularity, Colour uniformity, Diameter, and Evolution. In ABCDE, shape can be considered a combination of asymmetry and border irregularity. Other identification methodologies also include parameters that can be considered an important component of overall morphology, such as border irregularities, as presented by MacKie [5].

For NMSC, the dermoscopic features of its lesions are well-studied [6]. Nevertheless, studies taking shape into consideration are very limited, while this parameter is mostly evaluated in combination with others, such as lesion colour and texture [7].

While asymmetry is considered an important parameter for the visual detection of skin cancer, there is a lack of empirical data that relates the shape of the lesion to the probability of it being malignant. From this perspective, an objective characterisation and definition of lesions' shape may not only be useful for visual examination but could also aid the development of more precise and non-invasive methodologies. This variable can additionally be integrated into more developed methodologies using images and Artificial Intelligence, a field of research that has been of growing importance over the last few years [8,9].

The study of morphology is of growing interest in several fields of science [10], fueled primarily by the integration and improvement of advanced computer vision techniques towards the manipulation of different types of data. Many methods exist for the study of morphology, varying mostly by the means in which the data are defined and extracted. One common approach to the study of morphology is that of Geometric Morphometrics (GMM) [11–13]. GMM is a growing protagonist in fields related to biology and evolution [10], with other interesting applications in forensic sciences [14,15] and the study of microscopic elements on bones [16–18]. Nevertheless, GMM analyses are often hindered by the definition of landmark data; landmarks are precise homologous loci, of biological or geometric significance, that must always be identifiable across the sample [12,19].

In response to this, analysts began developing a different yet closely linked approach, making use of Fourier descriptors as a function of shape [20–22]. The principle of Fourier Analyses (FA) is to describe shape as a series of periodic functions along the curvature of an outline [21]. From this perspective, FA overcomes the limitations presented by GMM approaches, providing a means of analysing forms without a strict definition of homologous landmarks [11,13]. This type of methodology has been employed in a wide array of applications, ranging from the study of leaf shapes in biology [23], anthropological applications [24], or even the analysis of object design over time [25,26].

In this study, we present a novel approach to analysing skin lesions' shape, employing FA to investigate the shape of different skin lesion outlines. Thus, the aim of this research is to highlight the possible differences among NMSCs and benign skin lesions, proposing shape as a useful parameter for skin lesion classification. From this perspective, the data presented may provide an empirical approximation to the characterisation of skin lesions' shape.

## 2. Materials and Methods

### *2.1. Image Dataset*

The images used for the analysis were obtained from the Dermofit Dataset [7] (Figure 1), provided by the University of Edinburgh. This dataset has proven to be useful for the training of neural networks for skin lesion classification [27,28] and the segmentation of images via generative adversarial networks [29]. The scale of each image is unknown, and they were taken using a standard camera, thereby covering the visible area of the electromagnetic spectrum. No dermoscopes were employed for the collection of data. The Dermofit dataset additionally contains a mask delimiting each lesion area.

The original dataset consists of 10 classes, covering different cutaneous lesions. For the present study, the number of classes included was reduced, giving preference to well-defined NMSC lesions, and joining benign pathologies into one class, as a distinction between them was not considered clinically relevant. Under this premise, this study analyses a total of 4 different skin lesion types: Basal Cell Carcinoma (BCC, $n = 239$; Figure 1a), Intraepithelial Carcinoma (IEC, $n = 78$; Figure 1b), Squamous Cell Carcinoma (SCC, $n = 88$; Figure 1c), and a collection of benign lesions (BEN, $n = 588$), joining Seborrhoeic Keratosis (Figure 1d) and Melanocytic Nevus (Figure 1e).

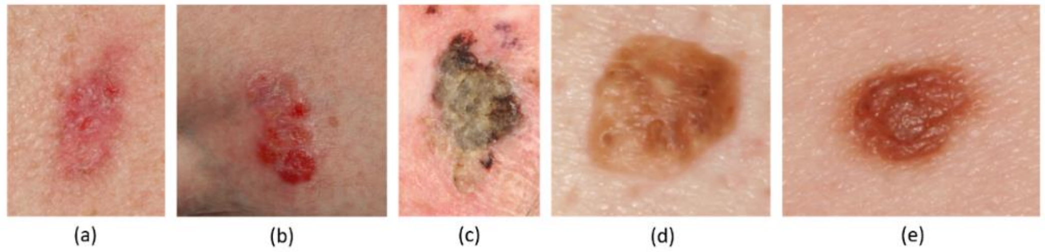


**Figure 1.** Example images for the different skin lesions, including BCC (**a**), IEC (**b**), SCC (**c**), and BEN: Seborrhoeic Keratosis (**d**) and Melanocytic Nevus (**e**).

## *2.2. Definition of Lesion Boundaries*

To define the borders of the lesions, a combined approach was followed. First, the original image segmentation provided by the Dermofit dataset was considered. This segmentation was manually established by the medical experts who curated the dataset. Nevertheless, in some cases, these classifications were found to have important differences regarding visual segmentation (the visual appearance of lesions; Figure 2a), resulting in a simplification of the boundaries (Figure 2b). The expert definition of boundaries is considered the optimal segmentation from a medical point of view; nevertheless, pixel level analyses to fit these boundaries to the point of highest spectral change can yield a higher level of detail and precision, thus providing a more empirical definition of the visual shape of the lesion. From this perspective, the automated refinement of segmentations using computer vision-based techniques allow for a more reproducible segmentation of the image, while the criteria given by the dermatologist remains crucial.

To obtain pixel-level segmentation, the present approach modified the manual technique by including an automatic computer vision technique. For this, each image was segmented using a k-means clustering algorithm [30] with 4 classes, defining areas mostly inside the pre-established boundary as a lesion, and thus refining the manual segmentation, based on their characteristics in the visual spectrum. Then, this was followed by a morphological closing algorithm [31], removing isolated areas and thus cleaning the segmented image to provide a single outline. This technique facilitated a better definition of lesion borders, especially in images where manual segmentation was observed to not fit well around the visual edges of the lesion (Figure 2b,c).

The obtained lesion boundary for each image was defined by 300 points, which were considered enough for a detailed representation of shape (Figure 2d).

Segmentation processes were performed using the Python programming language (v.3.7.6) and the OpenCV library.

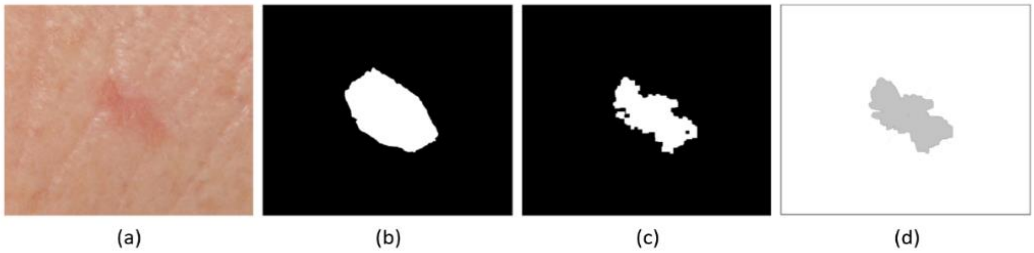


**Figure 2.** Example of original image (**a**), lesion mask provided as part of Dermofit dataset (**b**), recalculated lesion mask using k-means clustering (**c**), and obtained lesion boundary (**d**).

### *2.3. Elliptic Fourier Analyses*

Once outlines had been extracted, geometries were aligned and centered using variance–covariance matrices and eigenvalues. This step ensures that further calculations are invariant of the outline location, rotation, and origin. Outlines were also scaled using geometry centroid sizes as a scale factor (measured in pixels) to ensure pixel size and camera distance were not conditioning factors for the description of morphology. Centroid size was calculated as the distance from the edge to the centre (centroid) of each lesion along multiple points along the outline. Size features could not be further considered in the analyses because the Dermofit dataset does not provide a scale bar for each photo, which also obstructs any type of analysis of the lesion's *form* (shape + size; [32]). After normalisation procedures, outlines were analysed using an FA approach.

Fourier series are used to describe shapes by decomposing a periodic function into a sum of simpler trigonometric functions, such as sine and cosine values. These periodic functions can consider: (1) the distance of any point on the outline to the centroid [33], (2) the variation of the tangent angle for any point [33], or (3) a series of linearly transformed circular coordinates [34,35]. These approaches are known as Fourier Radius Variation, Fourier Tangent Angle, and Elliptic Fourier analyses, respectively. While each approach has its advantages and disadvantages, Elliptic Fourier Analyses (EFA) are more robust to irregularities along the outline [34,35] without the need for points to be equally spaced, thus enabling EFA to be easily fitted to any type of geometry. For this reason, EFA was selected as the most optimal approach for the present study.

Once calculated, each of these periodic functions can then be decomposed using Fourier series, resulting in a harmonic sum of trigonometric functions weighted with harmonic coefficients. Using EFA, Fourier coefficients are divided into 4 main groups, labelled $a$ through $d$. Coefficients $a$ and $b$ can be simply defined as the trigonometric moments around $x$ coordinate values, while coefficients $c$ and $d$ define the $y$ coordinate projection from circular to linear space [21,34,35]. Depending on the number of harmonics ($n$) used to describe the Fourier series, a set of coefficients—$a_n$, $b_n$, $c_n$ and $d_n$—can then be subjected to multivariate statistical analyses to empirically define each outline.

### *2.4. Multivariate Statistics*

For the present study, the first 19 harmonics of the elliptical Fourier series were used as descriptors of skin lesions' outlines. The optimal number of harmonics was calculated by estimating the cumulative power for each harmonic, with 19 harmonics representing up to 98.3% of the cumulative power. As is common practice in EFA, coefficients $a_1$, $b_1$, and $c_1$ were then used to normalise data [21], eliminating any remaining influence that size or rotation may have on subsequent analyses. This resulted in a final dataset of 73 morphological descriptors per individual.

Following this, dimensionality reduction was performed across coefficients through Principal Components Analyses (PCA). Principal Component (PC) Scores were then carefully assessed to evaluate the percentage of variance represented, selecting only those PC scores representing up to 95% of variance. Following PCA, analyses were carried out to assess the homogeneity of sample distributions using multiple Shapiro tests [36]. If samples were found to fit a Gaussian distribution, then subsequent analyses adopted a parametric approach, while non-Gaussian distributions were studied using robust statistical methods [8,37].

To assess statistical differences and similarities among groups, Multivariate Analyses of Variance (MANOVA) were performed. For normally distributed PC scores, the Hotelling–Lawley test statistic was used [38]. When normality was rejected, robust alternatives such as the Wilk's Lambda test statistic were used [39].

Additional analyses considered the use of Mahalanobis distances. For this purpose, within-group covariance distributions were first calculated, and then compared with distances to members of other groups. Statistical assessments of distribution differences were performed using either univariate Analysis of Variance tests (ANOVA) or Kruskal–Wallis tests, for Gaussian and Non-Gaussian distributions, respectively [40].

Changes in outline shape were visualised with the aid of transformation grids and warpings, computed using Thin Plate Splines (TPS) [41]. TPS grids minimise the bending between shapes to express changes in the relative position of points along the outline as the deformation of a grid. Therefore, TPS were used to fit central shape configurations for each of the groups separately and to visually calculate deformations when compared with other samples. Final calculations of outline deformations were then performed with the help of an isoline contour function. Additionally, oscilloscopes were used to evaluate changes in x and y coordinates across outlines. A trapezoidal integration was then computed to approximate an estimation of the area of each function ($\alpha$), thus evaluating the smoothness of oscilloscope curves. To provide a frame of reference, a perfect theoretical ellipsoid was computed to have an $\alpha = 0.0$.

Considering recent criticism on the "blind" use of $p$-values in applied statistics, the present study evaluated the hypothetical results while excluding the $p < 0.05$ rule for defining "statistical significance". In accordance with the most recent recommendations set forth by the American Statistician [42,43], $p$-values were evaluated by accompanying calculations of the probability of observations being a Type I statistical error, or the False Positive Risk (FPR) [44]. FPR values were calculated using the Sellke–Berger approach to define the likelihood ratio of the null hypothesis against the alternative hypothesis [45,46]. In general, prior probabilities of 0.5 were used for $p$-value calibrations, as suggested by Colquhoun [44,47]. Nevertheless, where possible, calibration confidence intervals were constructed using prior probabilities of 0.8 and 0.2 as well [8]. Throughout the study, FPR value calculations were only excluded for $p$-values over a 0.368 threshold, considering these values to be too high to accept the alternative hypothesis on any grounds [8]. Finally, a more robust $p$-value threshold of 0.003 was adopted as a threshold for more conclusive results, considering how this value is 3 standard deviations ($3\sigma$) from the mean, and associated with a 4.5% chance of being a Type I statistical error when using 0.5 prior probabilities [8].

All statistical applications were performed in the R (v.4.0.4) programming language. The R code to calculate EFA coefficients is available in the Supplementary Materials. Visualisations of EFA results were performed, in part, using the Momocs R library [48].

### *2.5. Asymmetry Calculations*

To empirically quantify and analyse lesion asymmetry, an index was designed and implemented. For each lesion, the centroid was calculated and then used to transpose outlines so that the x or y axis aligned with 0. Once centered, the absolute values of the axis in question were calculated, removing the line of "symmetry" between each value ($x$ or $y$) and the corresponding point on the opposite side of the outline ($x'$ or $y'$). The Euclidean distance, $d\ (x_i, x'_i)$, was then calculated between each point, and used to derive a quantitative measurement of outline displacement. An asymmetry index ($a(x)$ or $a(y)$) was then assigned to axis $x$ and axis $y$, respectively, through the root mean square Euclidean distance across each outline (Equation (1));

$$a(x)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}d\left(x_i,\left|x'_i\right|\right)^2},\quad a(y)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}d\left(y_i,\left|y'_i\right|\right)^2}, \tag{1}$$

The final asymmetry index for each skin lesion was calculated as the maximum index among the $x$ and $y$ axes.

Once asymmetry indices had been obtained for each sample, samples were tested for normality using Shapiro–Wilk tests, and then described using either traditional or robust statistical approaches [8,37,49]. For traditional descriptive statistics, sample mean and standard deviation were used to calculate central tendency and dispersion, respectively. For robust statistics, these were replaced by the median and the Square Root of the Biweight Midvariance (√BWMV). Similarly, 95% confidence intervals were also constructed using a [0.05, 0.95] interquartile range. Next, distributions of samples were analysed for statistical differences via ANOVA or Kruskal–Wallis tests. In addition to this, all aforementioned Fourier analyses were then repeated incorporating the asymmetry index into PCA, including the calculation of multivariate differences through MANOVA and Mahalanobis distances.

### 2.6. Machine Learning

To test the degree of separation amongst samples, classification tasks were performed using machine learning techniques. Therefore, a $k$-fold cross-validated ($k = 10$) Support Vector Machine (SVM) with a Radial Basis Function (RBF) was used [50]. SVMs are customizable and flexible models that use a *kernel-trick* to adjust for the existence or inexistence of parametric components, such as normality. Thus, this *kernel-trick* allows SVMs to construct non-linear decision boundaries. The SVM is additionally characterised using a soft maximised-margin as a decision boundary, thus avoiding overfitting of the data used for training.

SVMs were trained on 70% of data, separating the remaining 30% for testing and model evaluation. SVMs were mostly trained on raw PC scores, filtering only those PC scores representing up to 95% sample variance. For this purpose, the first experiment trained SVMs on PC scores obtained from EFA coefficients (Figure 3), while the second experiment trained SVMs on PC scores calculated when asymmetry indices were also included. Nevertheless, two additional experiments were also performed (Figure 3): one calculating the degree of univariate separability on asymmetry indices alone, and a final experiment appending the PC scores obtained from EFA coefficients with the asymmetry indices, enabling an assessment of the effect asymmetry has on classification results prior to a combined dimensionality reduction (Figure 3).

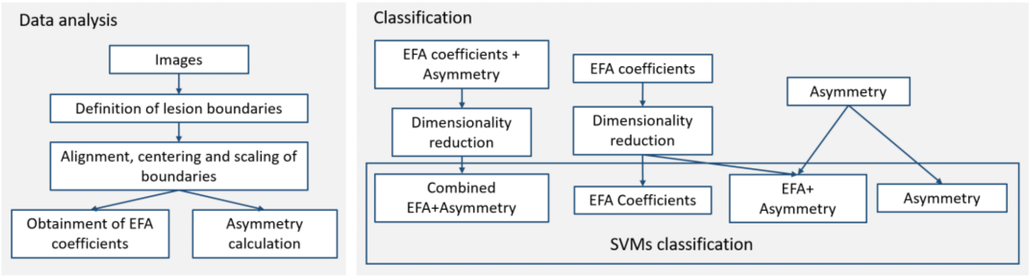


**Figure 3.** Methodological workflow.

For the selection of each SVM's optimal cost ($c$) and gamma ($\gamma$) hyperparameters, Bayesian Optimization Algorithms (BOAs) were employed [51–53]. BOA was initialised using a random optimization algorithm, thus defining the prior distribution for hyperparameter selection [53]. This was then followed by an Expected Improvement (EI) BOA algorithm for 50 iterations. While Gaussian Process Upper Confidence Bound

(GPUCB) and Probability of Improvement (PI) selection functions were also experimented with, they did not provide notable differences from their EI counterpart [53,54].

SVMs were evaluated on test sets, taking into consideration the general balance and imbalance of different sample sizes within the dataset while choosing appropriate evaluation criterion. While the selection of lesions from the Dermofit dataset does not present an extreme imbalance between benign and malignant tumours (≈29:20), when comparing between individual samples, this imbalance increases greatly (≈97:13 in the worst of cases). From this perspective, the present study chose to use evaluation metrics less susceptible to changes in sample balance [55], namely, Accuracy, Precision, Recall, the F1 Score, and the Area Under the precision–recall Curve (AUC). Each of these metrics, except for AUC, were calculated using confusion matrices, measuring the ratio of correctly classified individuals (True Positive & True Negative), as well as miss-classified individuals (False Positive & False Negative). AUC curves were calculated on the probability of label association values.

Machine learning applications were performed in the R programming language (v.4.0.4), primarily using the "caret" library.

## 3. Experimental Results

### 3.1. Elliptic Fourier Analyses

The analyses of the skin lesion morphologies revealed border irregularity to be a fundamental descriptive component of mainly malignant tumours. In general, PCA dimensionality reduction produces a high number of inhomogeneous PC scores (Shapiro $w = 0.86$, $p = 1.1 \times 10^{-28}$, FPR = $2.0 \times 10^{-24}$%), with the first 15 PC scores representing ≈90% of variance and 21 PC scores reaching ≈95% cumulative variance. The PCA plots (Figure 4) reveal a strong concentration of benign lesions (red colour in Figure 4) in the centre of each dimension (median [$x$, $y$] shape space coordinates = [0.009, 0.0009]), represented by more circular lesions, while all three malignant samples present much higher variance across the shape space (√BWMV Benign = 0.094 and Malignant = 0.115).

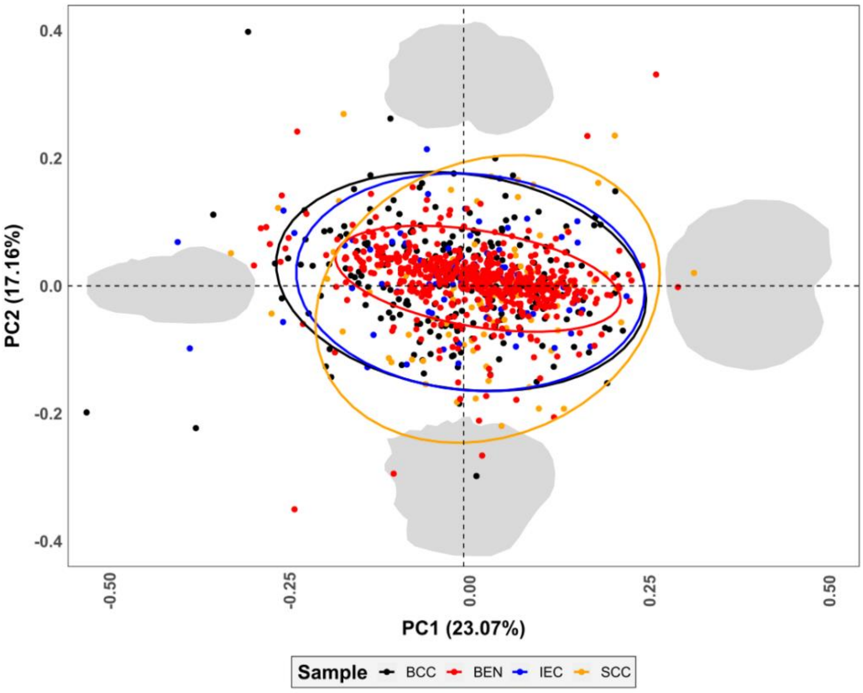


**Figure 4.** Principal Component Analysis (PCA) scatter plots with 95% confidence intervals presenting variance in skin lesions' shape, as represented by Elliptic Fourier Analyses.

Morphological variance calculated through Thin Plate Spline grid warpings are presented at the extremity of each PC score in grey. Shape space coordinate (0,0) is represented by circular lesions with no border irregularities. BCC = Basal Cell Carcinoma, BEN = Benign, IEC = Intraepithelial Carcinoma, SCC = Squamous Cell Carcinoma.

Upon analysing the projection and the morphological variations along the curvilinear abscissa (Figure 5), the oscilloscopes confirmed Benign samples to be the most elliptical lesions in nature ($\alpha$ = 0.78), with hardly any deviations from a theoretical ellipsoid (Figure 5). The SCC ($\alpha$ = 3.46) and IEC ($\alpha$ = 4.22) samples, on the other hand, appear to deviate the most along the outline, with frequent irregularities along the lesion borders. Interestingly, the BCC samples present a relatively smooth curve, where the malignant samples are of the greatest spherical nature ($\alpha$ = 1.06).

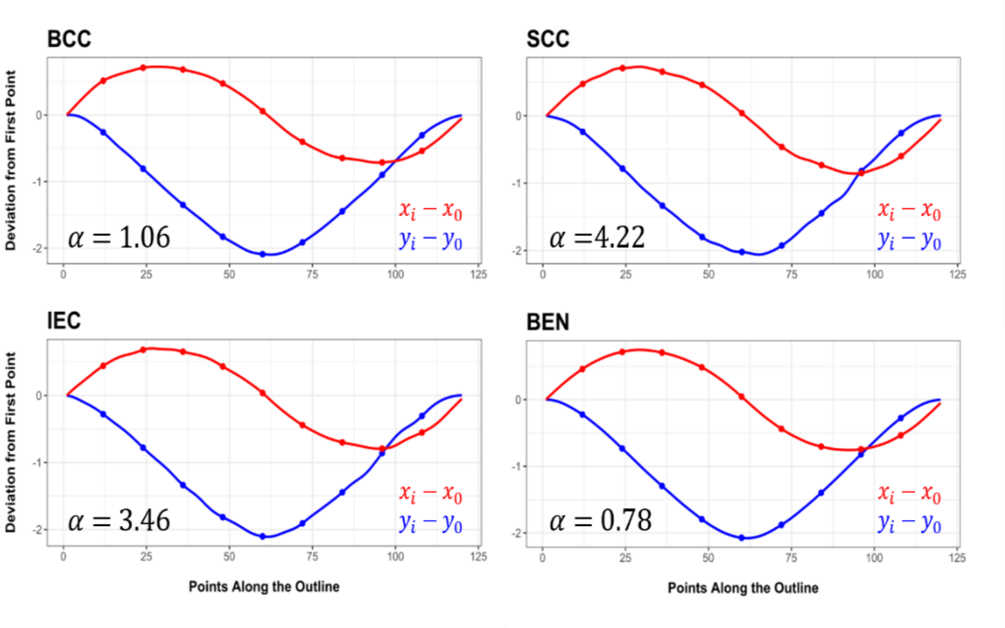


**Figure 5.** Oscilloscope curves reflecting variations along the outline of each of the samples according to elliptical Fourier descriptors. $\alpha$ values represent the results obtained from computing the area of each oscilloscope function. Perfect elliptical outlines would be presented by smooth sinusoidal curves with no irregular deviations ($\alpha$ = 0.0).

When analysing the differences between each of the malign lesions in comparison with the benign samples, the Thin Plate Splines (TPS) and isoline plots confirm these deformations (Figure 6), with samples such as SCC and IEC presenting distinct lateral constrictions. The BCC samples, on the other hand, are characterised by a more irregular-oval shape. Overall, the isoline heatmaps reveal all the malignant samples to present highly localised deformations, which would indicate shape variation to be a product of edge irregularities, as opposed to an overall change across the entire elliptical nature of the lesion. From this perspective, it could be assumed that lesion asymmetry is a powerful conditioning factor in diagnoses of malignant and benign lesions.

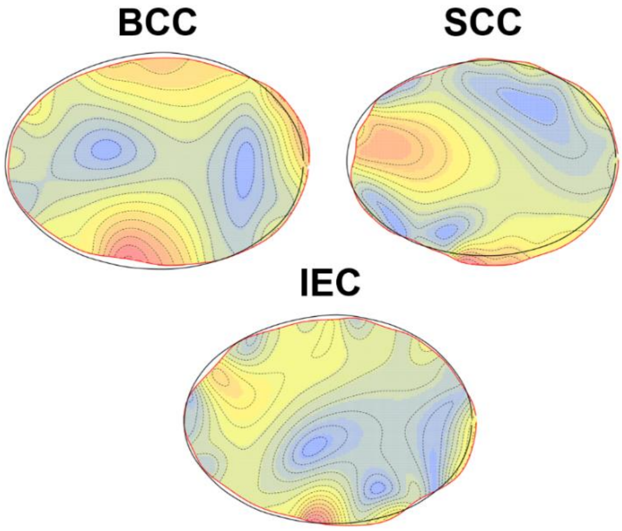


**Figure 6.** Deformation grid visualisations via isoline plots, projecting each of the central configurations for malignant samples onto the central shape of benign skin lesions. Red areas reflect areas of greater deformation from benign samples.

The multivariate quantification of the sample differences based on EFA shows that benign lesions frequently separate from all three types of malignant samples (MANOVA $p = 0.002$, FPR = 3.3 +/− [0.8, 11.9]%). When considering the Fourier coefficients alone, the separation between the Benign and IEC samples becomes a little less clear (Table 1), with a 5.7 +/− [1.5, 19.4]% chance of being a Type I statistical error when using MANOVA testing. Similarly, while the MANOVA results hint towards a possible separation among some of the malignant samples, the FPR calculations are too high to consider these observations conclusive, indicating that the malignant tumours are morphologically similar among themselves.

**Table 1.** Multivariate Analysis of Variance (MANOVA) and Mahalanobis distance testing to assess the degree of statistical differences between sample outlines. BCC = Basal Cell Carcinoma, BEN = Benign, IEC = Intraepithelial Carcinoma, SCC = Squamous Cell Carcinoma.

| | | **MANOVA** | | | **Mahalanobis Distances** | | |
|---|---|---|---|---|---|---|---|
| | | **BCC** | **BEN** | **IEC** | **BCC** | **BEN** | **IEC** |
| BEN | *p*-Value | 0.001 | | | $9.7 \times 10^{-47}$ | | |
| | FPR | 1.8% | | | $2.8 \times 10^{-42}$% | | |
| IEC | *p*-Value | 0.756 | 0.004 | | 0.228 | $1.6 \times 10^{-27}$ | |
| | FPR | - | 5.7% | | 37.9% | $2.7 \times 10^{-23}$% | |
| SCC | *p*-Value | 0.030 | 0.001 | 0.023 | 0.738 | $2.9 \times 10^{-22}$ | 0.292 |
| | FPR | 22.2% | 1.8% | 19.1% | - | $2.9 \times 10^{-10}$% | 49.4% |

When considering the Mahalanobis distances (Table 1), the calculations reveal much larger differences between the sample distributions, especially when separating between the Benign and Malignant lesions as a whole ($p = 2.5 \times 10^{-74}$, FPR = $1.2 \times 10^{-69}$ +/− [$2.9 \times 10^{-70}$, $4.6 \times 10^{-69}$]%). In this case, none of the malignant lesions appear to be similar, while benign lesion multivariate distributions appear to be notably separate from each of the carcinoma samples.

### 3.2. *Lesion Asymmetry*

Upon calculating the asymmetry indices for each of the samples, great differences emerged between the benign lesions and each of the carcinoma samples (Table 2, Figure

7). In most of the cases, the malignant lesions present much higher variability (Interquantile Range = 0.37, √BWMV = 0.093) as opposed to benign lesions (Interquantile Range = 0.19, √BWMV = 0.046). The differences between these samples are also of great importance ($\chi^2$ = 103.3, $p = 2.2 \times 10^{-16}$, FPR = $2.2 \times 10^{-12}$ +/− [$5.4 \times 10^{-13}$, $8.6 \times 10^{-12}$]%). When considering each malignant sample separately, boxplots indicate that BCC is the sample with the greatest degree of asymmetry (Figure 7); nevertheless, robust metrics (Table 2) reveal the IEC and BCC to have the same central index, with IEC presenting the largest robust interquartile range (BCC = 0.375, SCC = 0.400, IEC = 0.435).

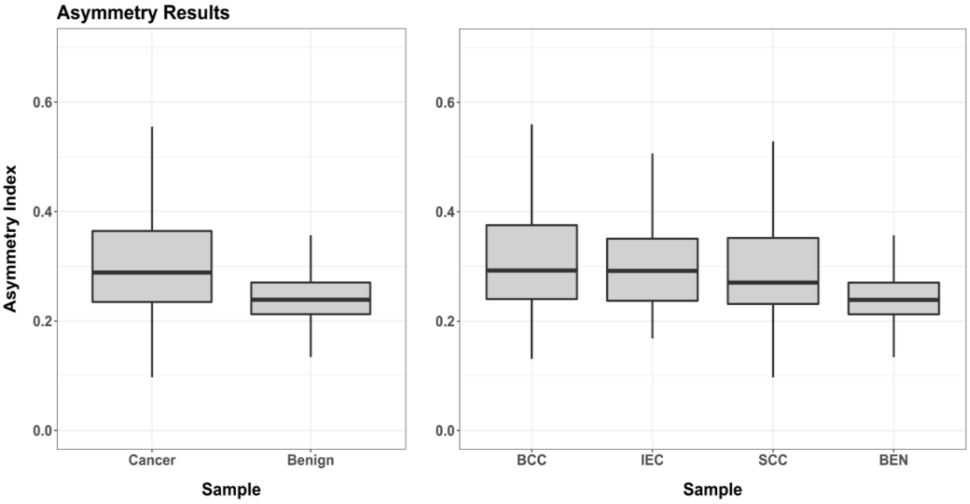


**Figure 7.** Boxplots presenting the maximum asymmetry index calculations for each of the samples.

**Table 2.** Descriptive statistics for the asymmetry indices of each of the samples. For space restrictions, FPR values were excluded from the present table considering all *p*-values were far below the 3σ threshold. L (0.05) = Lower bound 95% confidence interval; U (0.95) Upper bound 95% confidence interval; √BWMV = Square Root of the Biweight Midvariance.

| | **Shapiro** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ***w*** | ***p*** | **Min.** | **L (0.05)** | **Median** | **√BWMV** | **U (0.95)** | **Max.** |
| BCC * | 0.782 | $2.2 \times 10^{-16}$ | 0.131 | 0.179 | 0.292 | 0.099 | 0.554 | 1.064 |
| IEC * | 0.647 | $2.6 \times 10^{-12}$ | 0.168 | 0.186 | 0.292 | 0.081 | 0.621 | 1.285 |
| SCC * | 0.566 | $1.2 \times 10^{-14}$ | 0.097 | 0.172 | 0.270 | 0.084 | 0.572 | 1.390 |
| BEN | 0.616 | $2.2 \times 10^{-16}$ | 0.134 | 0.183 | 0.239 | 0.046 | 0.373 | 1.213 |
| Cancer * | 0.679 | $2.2 \times 10^{-16}$ | 0.097 | 0.183 | 0.289 | 0.093 | 0.554 | 1.390 |
| Benign | 0.616 | $2.2 \times 10^{-16}$ | 0.134 | 0.183 | 0.239 | 0.046 | 0.373 | 1.213 |

* Malignant (cancerous) samples.

Integrating asymmetry indices into the multivariate statistical analyses produces similarly complex non-gaussian PCA distributions ($w$ = 0.86, $p = 1.1 \times 10^{-28}$, FPR = $2.0 \times 10^{-24}$%), with ≈90% of the cumulative sample variance appearing in the first 15 PC scores and ≈95% in the first 21. As can be seen in the sample biplots (Figure 8), the asymmetry index represents the variable of greatest importance in the description of the sample morphology, correlating strongly with both PC1 (24.8% variance, $p = 2.1 \times 10^{-238}$, FPR = $3.1 \times 10^{-233}$%) and PC2 (16.6%, $p = 6.5 \times 10^{-24}$, FPR = $9.5 \times 10^{-20}$%).

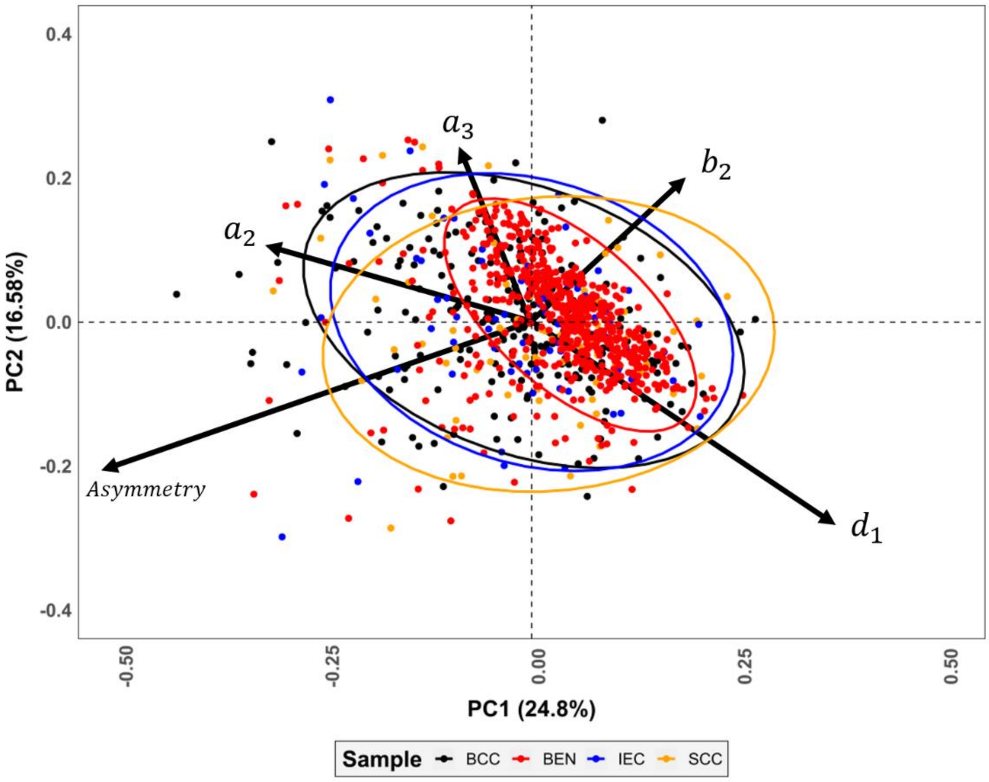


**Figure 8.** PCA biplot combining asymmetry indices with elliptical Fourier coefficients. For visual simplicity, only the first 5 most important variables were included in the biplot. Variables a, b, and d represent elliptic Fourier coefficients.

Through an in-depth analysis of the PCA scatter plots, it was observed that the asymmetry index produces a notable irregular dispersion among the malignant samples (skewness = −3.0, kurtosis = 12.0), pushing the benign lesions into a much more concentrated distribution (skewness = −3.9, kurtosis = 33.8), which is better described by the original elliptic Fourier coefficients.

As opposed to the calculations performed on Fourier coefficients alone, the inclusion of asymmetry presents a notable improvement in both the MANOVA and Mahalanobis results (Table 3), with all malignant lesions appearing clearly separable from benign lesions (MANOVA $p = 0.001$, FPR = 1.8 +/− [0.5, 7.0]%; Mahalanobis $p = 3.6 \times 10^{-75}$, FPR = $1.7 \times 10^{-70}$ +/− [$4.2 \times 10^{-71}$, $6.7 \times 10^{-70}$]%).

**Table 3.** Multivariate Analysis of Variance (MANOVA) and Mahalanobis distance testing to assess the degree of statistical differences between sample morphologies combining shape information and asymmetry.

| | | MANOVA | | | Mahalanobis Distances | | |
|---|---|---|---|---|---|---|---|
| | | BCC | BEN | IEC | BCC | BEN | IEC |
| BEN | *p*-Value | 0.001 | | | $3.6\times 10^{-45}$ | | |
| | FPR | 1.8% | | | $1.0 \times 10^{-40}$ % | | |
| IEC | *p*-Value | 0.814 | 0.001 | | 0.242 | $3.4 \times 10^{-27}$ | |
| | FPR | - | 1.8% | | 48.3% | $5.6 \times 10^{-23}$ % | |
| SCC | *p*-Value | 0.058 | 0.001 | 0.021 | 0.452 | $4.0 \times 10^{-23}$ | 0.051 |
| | FPR | 31.0% | 1.8% | 18.1% | - | $5.6 \times 10^{-19}$ % | 29.2% |

Thus, all the multivariate statistical results conclude asymmetry to be a considerable component for the identification of malignant lesions, with benign lesions being mostly characterised by their elliptical shape and greater overall symmetry.

### *3.3. Machine Learning*

The SVMs were found to successfully learn from the morphological data on most accounts (Table 4), with the worst performing models being univariate SVMs trained solely on asymmetry indices. The evaluation metrics also concur with the multivariate statistical results, revealing the combination of EFA data with the asymmetry index to be the most efficient means of differentiating between malignant and benign tumours (Table 5). Interestingly, SVMs appear to identify benign lesions with much greater accuracy than malignant lesions. This is likely because not all malignant lesions present an irregular border, while a much greater percentage of benign lesions are found to be concentrated around the elliptical-symmetric portion of the shape space. Nevertheless, the true positive to true negative rates remain high, resulting in a fairly balanced AUC metric.

**Table 4.** Overall evaluation metrics on test sets using Support Vector Machines for the classification of Benign and Malignant lesions. AUC = Area Under the precision–recall Curve. The combined EFA & Asymmetry category represents PCA dimensionality reduction techniques performed on both EFA coefficients and Asymmetry indices, prior to SVM training.

| Training Variables | Accuracy | Precision | Recall | F-Statistic | AUC |
|---|---|---|---|---|---|
| Asymmetry | 0.690 | 0.646 | 0.447 | 0.528 | 0.696 |
| EFA Coefficients | 0.772 | 0.887 | 0.717 | 0.794 | 0.693 |
| EFA & Asymmetry | 0.765 | 0.883 | 0.711 | 0.788 | 0.685 |
| Combined EFA & Asymmetry | 0.786 | 0.915 | 0.717 | 0.804 | 0.735 |

**Table 5.** Confusion Matrix calculated on test sets using the Combined EFA & Asymmetry dataset.

| | | True | |
|---|---|---|---|
| | | **Benign** | **Malignant** |
| **Predicted** | **Benign** | 71.67% | 10.53% |
| | **Malignant** | 28.33% | 89.47% |

## 4. Discussion

A well-known feature for the characterisation of malignant and benign skin lesions is their shape. While most diagnostic criteria try to assess these variables as a function of border regularity and overall symmetry, few studies have tried to empirically quantify these morphological traits. The present study analyses the morphological differences amongst NMSC and benign lesions using Fourier descriptors and asymmetry calculations. To the authors' knowledge, this is one of the first approximations to objectively defining these dermatological criteria using computer vision and multivariate statistical techniques.

In recent years, morphological tools such as GMM and EFA have proven highly efficient in the evaluation of medical data. From this perspective, interesting studies have employed landmark-based techniques for the diagnosis and evaluation of patients with several diseases and syndromes. These include, but are not exclusive to, Beta Thalassaemia [56], Glut1 Deficiency Syndrome [57], Fetal Alcohol Syndrome [58], Obstructive Sleep Apnea Syndrome [59], and (though not strictly using GMM) the study of Down Syndrome patients [60]. Fourier descriptors, on the other hand, have been used to a lesser extent, with applications in ovarian tumour analysis [61], as well as the study of optic nerve head morphology and glaucoma [62]. The present study contributes to these efforts, expanding dermatological analyses to include these types of tools as well.

While GMM has proven to be more popular in medical analyses over EFA, this is likely due to the large corpus of pre-existing research using GMM in the analysis of cranial morphology in physical anthropology [10]. From this perspective, the definition of landmarks for these types of analyses are already well-defined, while post-cranial soft-tissue research in medicine is notably lacking. Considering the difficulties that may exist in defining truly homologous landmarks on soft tissue, EFA presents the distinct advantage of being able to describe morphological data in elements where landmarks may not exist. Nevertheless, a fundamental component in any of these studies is the method though which these data are obtained.

A correct and objective definition of skin lesions' boundaries is a complex task, with most advanced techniques involving methodologies such as spectroscopy or hyperspectral imaging [8,29,63,64]. The task is especially challenging given the nature of the images used for this study, covering only the visible area of the spectrum. Similarly, while benign samples are composed mostly of pigmented lesions, NMSC classes often contain un-pigmented lesions, whose delimitation is especially complex. For this reason, the present study worked with a dataset of visually delimited lesions, whose boundaries could later be refined using K-means algorithms. From this perspective, future research should address the use of morphological analyses on boundaries that have been extracted using automated methods. These could include techniques such as those provided when combining multispectral or hyperspectral imagery, alongside advanced computational learning techniques.

Employing image segmentation techniques on ultrasound images, a methodologically similar study by Martínez-Más and colleagues [61] was able to successfully characterize ovarian tumours, reaching up to ≈87% accuracy and an ≈0.87 Area under the Receiver Operator Characteristic Curve. While applied to a different medical case study, these authors present an additional account of how Fourier shape descriptors and machine learning algorithms can be considered useful tools in medical diagnostics.

The present study revealed notable statistical differences between benign and malignant skin samples, wherein most of the statistical tests appeared conclusive (FPR < 6%). In addition, machine learning algorithms were able to reach up to 78.6% accuracy (AUC = 0.735). If integrated into a practical tool, combining both the use of automated outline extraction using computer vision techniques and the additional analyses of these outlines via Fourier shape descriptors, this methodological workflow could prove to be a powerful tool, especially at the screening stage of skin cancer diagnoses.

Clearly, the present results reveal shape and asymmetry to be more of an indicator of malignancy than the type of malignancy. The results obtained within this study hint that all malignant cutaneous tumours are mainly characterised by morphological irregularities in comparison with asymmetry, while not much else can be obtained through the current methodology. Nevertheless, it is important to note that most diagnostic criteria in skin lesion research are based on a combination of variables [3–5] and no single variable alone. At present, the EFA approaches described herein have been limited to a description of pure shape, while the lack of a scale bar in the Dermofit dataset hinders the possibility of studying form (shape & size; [32]). Likewise, an important component of skin lesion diagnostics is found in colour [3–5], a variable that may be integrated into future analyses through more advanced computer vision techniques.

In conclusion, this study describes a new methodological approach to the characterisation of non-melanoma- as well as benign-type skin lesions. Through a combined use of computer vision techniques, elliptical Fourier analyses, and computational learning, a ≈79% separation has been achieved between malignant and benign lesions, supported by notable statistical results ($p < 0.003$). Similarly, asymmetry has been found to be a fundamental variable in the description of cutaneous carcinomas. Nevertheless, future investigation should be dedicated to the analysis of more efficient and accurate segmentation procedures, while searching for means to integrate morphological and electromagnetic information into a more robust and well-rounded diagnostic tool.

**Supplementary Materials:** The following supporting information can be downloaded at: www.mdpi.com/article/10.3390/jcm11154392/s1, Supplementary File: R code for the calculation of Elliptic Fourier Coefficients.

**Author Contributions:** Conceptualization, L.A.C. and I.B.-G.; methodology, L.A.C. and I.B.-G.; software, L.A.C. and I.B.-G.; validation, L.A.C., I.B.-G. and J.A.; formal analysis, L.A.C. and I.B.-G.; investigation, L.A.C. and I.B.-G.; resources, D.G.-A.; data curation, L.A.C. and I.B.-G.; writing—original draft preparation, L.A.C. and I.B.-G.; writing—review and editing, L.A.C., I.B.-G., J.A., D.G.-A., M.R.-M. and P.R.-G.; visualization, L.A.C. and I.B.-G.; supervision, D.G.-A., J.C. and C.R.-C.; project administration, D.G.-A.; funding acquisition, D.G.-A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research has been funded by the European Regional Development Fund and the Junta de Castilla y León, under the project name DETECCTHIA (Ref. SA097P20). L.A.C. is funded by the Spanish Ministry of Science, Innovation, and Universities, with an FPI Predoctoral Grant (Ref. PRE2019-089411) associated to project RTI2018-099850-B-I00.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All data used in the present study originate from the Dermofit dataset and are available for purchase through the University of Edinburgh; https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html (accessed on 27 April 2021).

**Acknowledgments:** This research has made use of the high-performance computing resources of the Castilla y León Supercomputation Center (SCAYLE; www.scayle.es, accessed on 27 April 2021), financed by the European Regional Development Fund (ERDF).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Weinberg, A.S.; Ogle, C.A.; Shim, E.K. Metastatic cutaneous squamous cell carcinoma: An update. *Dermatol. Surg.* **2007**, *33*, 885–899.
2. Hoorens, I.; Vossaert, K.; Ongenae, K.; Brochez, L. Is early detection of basal cell carcinoma worthwhile? Systematic review based on the WHO criteria for screening. *Br. J. Dermatol.* **2016**, *174*, 1258–1265.
3. Friedman, R.J.; Rigel, D.S.; Kopf, A.W. Early detection of malignant melanoma: The role of physician examination and self-examination of the skin. *CA Cancer J. Clin.* **1985**, *35*, 130–151.
4. Tsao, H.; Olazagasti, J.M.; Cordoro, K.M.; Brewer, J.D.; Taylor, S.C.; Bordeaux, J.S.; Chren, M.M.; Sober, A.J.; Tegeler, C.; Bhusan, R.; Begolka, W.S. Early detection of melanoma: Reviewing the ABCDEs. *J. Am. Acad. Dermatol*. **2015**, *72*, 717–723.
5. MacKie, R.M. An Illustrated Guide to the Recognition of Early Malignant Melanoma, University of Glasgow, Glasgow, Scotland, 1986.
6. Fargnoli, M.C.; Kostaki, D.; Micantonio, T. Dermoscopy in the diagnosis and management of non-melanoma skin cancers. *Artic. Eur. J. Dermatol.* **2012**, *22*, 456–463. https://doi.org/10.1684/ejd.2012.1727.
7. Ballerini, L.; Fisher, R.B.; Aldridge, B.; Rees, J. A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions. *Lect. Notes Comput. Vis. Biomech.* **2013**, *6*, 63–86.
8. Courtenay, L.; Gonzalez-Aguilera, D.; Lagüela, S.; del Pozo, S.; Ruiz Méndez, C.; Barbero-García, I.; Román-Curto, C.; Cañueto, J.; Santos-Durán, C.; Cardeñoso-Álvarez, M.E.; et al. Hyperspectral Imaging and Robust Statistics in Non-Melanoma Skin Cancer Analysis. *Biomed. Opt. Express* **2021**, *12*, 5107–5127.
9. Dildar, M.; Akram, S.; Irfan, M.; Khan, H.U.; Ramzan, M.; Mahmood, A.R.; Alsaiari, S.A.; Saeed, A.H.M.; Alraddadi, M.O.; Mahnashi, M.H. Skin Cancer Detection: A Review Using Deep Learning Techniques. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5479.
10. Weber, G.W.; Bookstein, F.L. *Virtual Anthropology: A Guide to a New Interdisciplinary Field*; Springer: Vienna, Austria, 2011.
11. Bookstein, F.L. *Morphometric Tools for Landmark Data*; Cambridge University Press: Cambridge, UK, 1992.
12. Bookstein, F.L. Landmark Methods for Forms without Landmarks: Morphometrics of Group Differences in Outline Shape. *Med. Image Anal.* **1997**, *1*, 225–243.
13. Dryden, I.L.; Mardia, K.V. *Statistical Shape Analysis: Wiley Series in Probability and Statistics*; John Wiley Sons Ltd.: New York, NY, USA, 1998.

14. Kieser, J.; Bernal, V.; Gonzalez, P.; Birch, W.; Turmaine, M.; Ichim, I. Analysis of experimental cranial skin wounding from screwdriver trauma. *Int. J. Leg. Med.* **2008**, *122*, 179–187.
15. Komo, L.; Grassberger, M. Experimental sharp force injuries to ribs: Multimodal morphological and geometric morphometric analyses using micro-CT, macro photography and SEM. *Forensic Sci. Int.* **2018**, *288*, 189–200.
16. Aramendi, J.; Maté-González, M.A.; Yravedra, J.; Ortega, M.C.; Arriaza, M.C.; González-Aguilera, D.; Baquedano, E.; Domínguez-Rodrigo, M. Discerning carnivore agency through the three-dimensional study of tooth pits: Revisiting crocodile feeding behaviour at FLK- Zinj and FLK NN3 (Olduvai Gorge, Tanzania). *Palaeogeogr. Palaeoclim. Palaeoecol.* **2017**, *488*, 93–102.
17. Courtenay, L.A.; Huguet, R.; González-Aguilera, D.; Yravedra, J. A Hybrid Geometric Morphometric Deep Learning Approach for Cut and Trampling Mark Classification. *Appl. Sci.* **2020**, *10*, 150.
18. Courtenay, L.A.; Huguet, R.; Yravedra, J. Scratches and grazes: A detailed microscopic analysis of trampling phenomena. *J. Microsc.* **2020**, *277*, 107–117.
19. Gunz, P.; Mitteroecker, P.; Bookstein, F.L. Semilandmarks in Three Dimensions. In *Modern Morphometrics in Physical Anthropology*; Slice, D.E., Ed.; Springer: Boston, MA, USA, 2005; pp. 73–98.
20. Rohlf, F.J.; Archie, J.W. A Comparison of Fourier Methods for the Description of Wing Shape in Mosquitoes (Diptera: Culicidae). *Syst. Biol.* **1984**, *33*, 302–317.
21. Ferson, S.; Rohlf, F.J.; Koehn, R.K. Measuring Shape Variation of Two-dimensional Outlines. *Syst. Biol.* **1985**, *34*, 59–68.
22. Rohlf, F.J. Relationships among eigenshape analysis, Fourier analysis, and analysis of coordinates. *Math. Geol.* **1986**, *18*, 845–854.
23. Chitwood, D.H.; Sinha, N.R. Evolutionary and Environmental Forces Sculpting Leaf Development. *Curr. Biol.* **2016**, *26*, 297–306.
24. Caple, J.; Byrd, J.; Stephan, C.N. Elliptical Fourier analysis: Fundamentals, applications, and value for forensic anthropology. *Int. J. Leg. Med.* **2017**, *131*, 1675–1690.
25. Iovită, R. Comparing Stone Tool Resharpening Trajectories with the Aid of Elliptical Fourier Analysis. In *New Perspectives on Old Stones*; Springer: New York, NY, USA, 2010; pp. 235–253.
26. Chitwood, D.H. Imitation, Genetic Lineages, and Time Influenced the Morphological Evolution of the Violin. *PLoS ONE* **2014**, *9*, e109229.
27. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118.
28. Mukherjee, S.; Adhikari, A.; Roy, M. Malignant Melanoma Classification Using Cross-Platform Dataset with Deep Learning CNN Architecture. *Adv. Intell. Syst. Comput.* **2019**, *922*, 31–41.
29. Izadi, S.; Mirikharaji, Z.; Kawahara, J.; Hamarneh, G. Generative adversarial networks to segment skin lesions. *Proc. Int. Symp. Biomed. Imaging* **2018**, *15*, 881–884.
30. Lloyd, S.P. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.
31. Serra, J. *Image Analysis and Mathematical Morphology*; Academic Press: New York, NY, USA, 1982.
32. Jungers, W.L.; Falsetti, A.B.; Wall, C.E. Shape, relative size, and size-adjustments in morphometrics. *Am. J. Phys. Anthropol.* **1995**, *38*, 137–161.
33. Zahn, C.T.; Roskies, R.Z. Fourier Descriptors for Plane Closed Curves. *IEEE Trans. Comput.* **1972**, *21*, 269–281.
34. Giardina, C.R.; Kuhl, F.P. Accuracy of curve approximation by harmonically related vectors with elliptical loci. *Comput. Graph. Image Process.* **1977**, *6*, 277–285.
35. Kuhl, F.P.; Giardina, C.R. Elliptic Fourier features of a closed contour. *Comput. Graph. Image Process.* **1982**, *18*, 236–258.
36. Mohd Razali, N.; Bee Wah, Y. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
37. Höhle, J.; Höhle, M. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS J. Photogramm Remote Sens.* **2009**, *64*, 398–406.
38. Hotelling, H.A.; Generalized, T. Test and Measure of Multivariate Dispersion. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* Neyman, J., Eds; University of California Press: Berkeley, CA, USA, 1951; pp. 23–42.
39. Rao, C.R. An asymptotic expansion of the distribution of Wilk's criterion. *Bull. L'institut. Int. Stat.* **1951**, *33*, 177–180.
40. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; John Wiley & Sons: New York, NY, USA, 1973.
41. Bookstein, F.L. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 567–585.
42. Wasserstein, R.L.; Lazar, N.A. The ASA Statement on p-Values: Context, Process, and Purpose, *Am. Stat.* **2016**, *70*, 129–133.
43. Wasserstein, R.L.; Schirm, A.L.; Lazar, N.A. Moving to a World Beyond "p < 0.05." *Am. Stat.* **2019**, *73*, 1–19.
44. Colquhoun, D. The False Positive Risk: A Proposal Concerning What to Do About p-Values. *Am. Stat.* **2019**, *73*, 192–201.
45. Benjamin, D.J.; Berger, J.O. Three Recommendations for Improving the Use of p-Values. *Am. Stat.* **2019**, *73*, 186–191.
46. Sellke, T.; Bayarri, M.J.; Berger, J.O. Calibration of p Values for Testing Precise Null Hypotheses. *Am. Stat.* **2012**, *55*, 62–71.
47. Colquhoun, D. The reproducibility of research and the misinterpretation of p-values. *R. Soc. Open Sci.* **2017**, *4*, 171085.
48. Bonhomme, V.; Picq, S.; Gaucherel, C.; Claude, J. Momocs: Outline analysis using R. *J. Stat. Softw.* **2014**, *56*, 1–24.

49. Courtenay, L.A.; Herranz-Rodrigo, D.; Huguet, R.; Maté-González, M.Á.; González-Aguilera, D.; Yravedra, J. Obtaining new resolutions in carnivore tooth pit morphological analyses: A methodological update for digital taphonomy. *PLoS ONE* **2020**, *15*, e0240328.
50. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
51. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv* **2012**, arXiv:1203.2944.
52. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2016**, *104*, 148–175.
53. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization Yoshua Bengio. *J. Mach. Learn. Res*. **2012**, *13*, 281–305.
54. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. *Int. Conf. Neural. Inf. Process. Syst*. **2011**, *24*, 2546–2554.
55. He, H.; Ma, Y. *Imbalanced Learning*; IEEE Press: Piscataway, NJ, USA, 2013.
56. Roussos, P.; Mitsea, A.; Halazonetis, D.; Sifakakis, I. Craniofacial shape in patients with beta thalassaemia: A geometric morphometric analysis. *Sci. Rep.* **2021**, *11*, 1–13.
57. Pucciarelli, V.; Bertoli, S.; Codari, M.; de Amicis, R.; De Giorgis, V.; Battezzati, A.; Veggiotti, P.; Sforza, C. The face of Glut1-DS patients: A 3D Craniofacial Morphometric Analysis. *Clin. Anat*. **2017**, *30*, 644–652.
58. Mutsvangwa, T.E.M.; Meintjes, E.M.; Viljoen, D.L.; Douglas, T.S. Morphometric analysis and classification of the facial phenotype associated with fetal alcohol syndrome in 5- and 12-year-old children. *Am. J. Med. Genet. Part. A* **2010**, *152*, 32–41.
59. Turam Ozdemir, S.; Ercan, I.; Ezgi Cam, F.; Ocakoglu, G.; Demirdogen, E.; Ursavas, A. Three-Dimensional Analysis of Craniofacial Shape in Obstructive Sleep Apnea Syndrome Using Geometric Morphometrics Análisis. *Int. J. Morphol*. **2019**, *37*, 338–343.
60. Starbuck, J.M.; Cole, T.M.; Reeves, R.H.; Richtsmeier, J.T. The Influence of trisomy 21 on facial form and variability. *Am. J. Med. Genet. Part. A* **2017**, *173*, 2861–2872.
61. Martínez-Más, J.; Bueno-Crespo, A.; Khazendar, S.; Remezal-Solano, M.; Martínez-Cendán, J.P.; Jassim, S.; Du, H.; Al Assam, H.; Bourne, T.; Timmerman, D. Evaluation of machine learning methods with Fourier Transform features for classifying ovarian tumors based on ultrasound images. *PLoS ONE* **2019**, *14*, 1–14.
62. Sanfillipo, P.G.; Grimm, J.L.; Flanagan, J.G.; Lathrop, K.L.; Sigal, I.A. Application of Elliptic Fourier Analysis to describe the Lamina Cribrosa Shape with age and intraocular pressure. *Exp. Eye Res.* **2014**, *128*, 1–7.
63. Leon, R.; Martinez-Vega, B.; Fabelo, H.; Ortega, S.; Melian, V.; Castaño, I.; Carretero, G.; Elmeida, P.; Garcia, A.; Quevedo, E.; et al. Non-Invasive Skin Cancer Diagnosis Using Hyperspectral Imaging for In-Situ Clinical Support. *J. Clin. Med.* **2020**, *9*, 1662.
64. Zhang, Y.; Moy, A.J.; Feng, X.; Nguyen, H.T.M.; Sebastian, K.R.; Reichenberg, J.S.; Markey, M.K.; Tunnell, J.W. Diffuse reflectance spectroscopy as a potential method for nonmelanoma skin cancer margin assessment. *Transl. Biophotonics* **2020**, *2*, e202000001.

*Publications: Section 5*

# A New Mathematical Model for Morphological Analyses

*Spanish Translation of Title and Abstract*

# Un nuevo enfoque de morfometría geométrica basado en la Teoría de Grafos para el análisis de los radios de primates: un nuevo modelo matemático para el tratamiento de datos tipo landmark.

La morfometría geométrica es una herramienta que describe la morfología a partir de una serie de coordenadas, una vez eliminados los efectos de la posición, la rotación y la escala. Aquí podemos encontrar los conceptos de forma "pura" (*shape*) y forma "total" (*form*), excluyendo esta última el procedimiento de escalado de los datos. La reducción de dimensiones en morfometría geométrica es una herramienta común para la representación de los datos en un espacio de características simplificado, manteniendo la mayor parte posible de la variación morfológica original. En este trabajo, presentamos un nuevo modelo matemático que puede utilizarse para mejorar la calidad de las técnicas de reducción de dimensiones (ej., el análisis de componentes principales) en forma de una nueva librería de R, *GraphGMM*. La librería utiliza elementos del aprendizaje geométrico y la teoría de grafos para agregar e integrar la información morfológica en un nuevo conjunto de coordenadas transformadas, posterior a la superposición Procrustes. Utilizando como caso de estudio los radios de grandes simios actuales, mostramos cómo los puntos de referencia (*landmarks*) integrados capturan eficazmente la información morfológica antes de la reducción de dimensiones, lo que implica una construcción más eficiente del espacio de características final. Con la ayuda de este modelo, y en combinación con *landmarks* y *semilandmarks*, hemos sido capaces de detectar patrones morfológicos en la caña del radio que suelen quedar oscurecidos en los enfoques tradicionales basados en técnicas de reducción de dimensiones menos potentes. Dichos patrones también pueden llegar a obviarse como resultado de la carencia de descriptores morfológicos en ciertas porciones del hueso. Además, la aplicación de las técnicas aquí presentadas proporciona afinidades morfológicas más estables entre los grupos, independientemente de la entrada de datos, lo que podría ser muy beneficioso para estudios paleoantropológicos en los que la recogida de datos tiende a estar limitada por la naturaleza fragmentaria del registro fósil.

Submitted to **Journal of Anatomy**

*Supplementary Information and Links*

**Current version of Supplementary Information is available from:**
https://figshare.com/s/15ea6077bd6fbd7883bf

**R library and code available from:**
https://github.com/LACourtenay/GraphGMM

# A Graph Based Geometric Morphometric approach to the analysis of primate radii: A new mathematical model for the ordination of landmark data.

Lloyd A. Courtenay[1,2*], Julia Aramendi [3*], Diego González-Aguilera[1]

[1] Department of Cartographic and Land Engineering, Higher Polytechnic School of Ávila, University of Salamanca, Hornos Caleros 50, 05003, Ávila, Spain.
[2] Department of Prehistory, Ancient History and Archaeology, Complutense University of Madrid, Prof. Aranguren s/n, 28040, Madrid, Spain.
[3] Department of Geology, Facultad de Ciencia y Tecnología, Universidad del País Vasco-Euskal Herriko Unibertsitatea (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa, Spain.

* These authors contributed equally

**Corresponding author**
Email: ladc1995@gmail.com
Mobile: +34 633 647 825
ORCID ID: 0000-0002-4810-2001

## Abstract

Geometric Morphometrics is a tool that describes morphology as a series of coordinates after the effects of variation in translation, rotation, and scale are removed. This can be further divided into the notion of shape and form, where the latter excludes the scaling procedure from analyses. Dimensionality reduction in geometric morphometrics is necessary for the representation of this data in a more simplistic feature space, while maintaining as much of the original variation as possible. Here we present a new mathematical model that can be used to enhance the quality of dimensionality reduction techniques such as Principal Component Analyses. Integrated into a new R library, the GraphGMM framework uses elements of geometric learning and graph theory to aggregate and embed morphological information into a new set of transformed coordinates, posterior to Procrustes superimposition. Using a case study of great ape radii, we show how embedded landmarks efficiently capture morphological information prior to dimensionality reduction, leading to a more efficient construction of the final feature space. With the help of this model, in combination with landmarks and semilandmarks, we have been able to detect morphological patterns in the radius midshaft that are usually obscured in traditional approaches based on less powerful dimensionality reduction techniques, or as a result of the unavailability of descriptors in certain bone portions. Furthermore, the application of the techniques presented here also provide more stable morphometric affinities among groups, regardless of the data input, which could be very beneficial for paleoanthropological studies in which data collection tends to be limited by the fragmentary nature of the fossil record.

**Key Words:** Graph Theory, Neighbourhood Aggregation and Embedding, Dimensionality Reduction, Human Evolution, Morphology, Biomechanics.

## 1. Introduction

Morphology is a key source of information for the study of biological organisms. This type of data combines insights into structure and geometry in order to understand and classify evolutionary traits. The integration of morphological analyses has thus played a particular role in the field of human evolution. Nevertheless, this branch of science has seen a long evolution itself, with multiple innovations and debates into the most optimal means of capturing morphological patterns.

Most early research into the description of evolutionary morphological traits can be found to work from linear measurements, known as traditional morphometrics (Rohlf and Marcus, 1993). From this perspective, shape was described as a series of Euclidean distances, and an evaluation of their ratios (Dodson, 1978; Humphries et al., 1981; Jungers et al., 1995, *inter alia*). This was closely followed by the advent of both landmark (O'Higgins and Johnson, 1988; Bookstein, 1991; Rohlf and Marcus, 1993), and outline (Rohlf and Archie, 1984; Ferson et al., 1985; Rohlf, 1986), based approaches, specialised more in the recovery of geometric representations and original morphologies from which measurements are obtained. In development of both traditional morphometric, and landmark based studies, Euclidean Distance Matrix Analyses (EDMA), are also based on the construction of pairwise matrices of distances, describing metric relationships between landmarks (Lele and Richtsmeier, 1991). Finally, other approaches exist and are under current development, such as diffeomorphism based analyses, which are landmark free (e.g. Avants et al., 2006; Millet et al., 2014; Beaudet et al., 2017, Braga et al., 2019; Urciuoli et al., 2021; Zanolli et al., 2022).

Currently landmark based approaches are by far the most widely used with a variety of applications. This particular discipline within Virtual Anthropology is known as Geometric Morphometrics (GMM). GMMs describe shape and form through the digitisation of a set of anatomically, mathematically, or geometrically distributed homologous loci (Bookstein, 1991, 1997; Dryden and Mardia, 1998). Landmarks are represented as coordinates in 2D or 3D space, which are then projected onto a common coordinate system. This process involves the superimposition of landmarks through a series of procedures, including scaling, rotation and translation, based on the algorithm first published by Gower (1975), known as Generalised Procrustes Analyses (GPA) (Rohlf and Slice, 1990). GPA is thus useful for the direct comparison and visualisation of landmark configurations, providing a means of quantifying minute displacements of coordinates in space (Bookstein, 1989).

An additional advantage of superimposed coordinates is the ability to calculate multivariate statistical data regarding their overall position within the configuration. A popular technique for analysing this type of data consists in the eigendecomposition of superimposed landmark coordinates (Rohlf, 1996, 2000; Klingenberg and Monteiro, 2005). In this context, Principal Component Analyses (PCA) are used to describe the major trends of landmark variation in as "few statistically orthogonal dimensions as possible" (Rohlf, 1996). From this perspective, linear and orthogonal relationships are constructed to describe slight displacements in coordinate values along the $x$, $y$, and when included, $z$ axes.

The present study parts from the hypothesis that this approach may result in the loss of information on the integrity of the landmark in the configuration as a whole when visualising this data, as we are blind to the existing relationships among multiple landmarks and their neighbours. While this technique is a necessary step for Dimensionality Reduction, and facilitates many subsequent statistical analyses, as will be shown in the present study, these approaches are prone to losing some structural information.

The present study intends to rectify some of these issues by adapting graph theory for a more structurally aware visual representation of landmark data. Here we show that a Graph-based GMM approach is able to represent a higher degree of morphological variability in fewer dimensions of feature space. We additionally discuss the advantages of using non-linear dimensionality reduction techniques for the study and representation of this type of information. Finally, and as a means of testing this mathematical model, we use Graph-based GMM to analyse the morphological variability of modern day great ape radii, with implications for the analysis of biomechanical traits in evolution.

## 2. Methods

### 2.1. The GraphGMM R library

All of the methods described in the present paper have been implemented in the R programming language, compatible with R v.3.0 and R v.4.0, and are available from the corresponding author's GitHub page (https://github.com/LACourtenay/GraphGMM). A detailed description of the library, alongside instructions and a guide to its installation and usage, have also been provided as supplementary materials (Sup. Files 1 & 2).

### 2.1.1. Mathematical Model: Graph-based Geometric Morphometrics

The mathematical model described in this study proposes a means of embedding landmark configurations into a new geometrically and structurally-meaningful feature space, prior to dimensionality reduction. The goal of information embedding is to map data into a new $\mathbb{R}^k$ feature space, with $k$ being the number of dimensions, such that entities that are structurally similar are embedded closer together (Hoff et al., 2012; Grover and Leskovec, 2016; Hamilton et al., 2018; Leskovec, 2019). For this purpose, we use a message passing mechanism for neighbourhood aggregation (Fig. 1a). In the context of GMM, this can be adapted as a means of mapping landmark configurations into a new feature space, such that similarity in the embedded feature space approximates similarity in biological and geometrical structure (Fig. 1b). This proximity in the new feature space can be considered a means of revealing landmark homophily (Fortunato, 2010; Yang and Leskovec, 2014), as well as structural equivalence (Henderson et al., 2012), in the global context of the entire configuration (Fig. 1c).

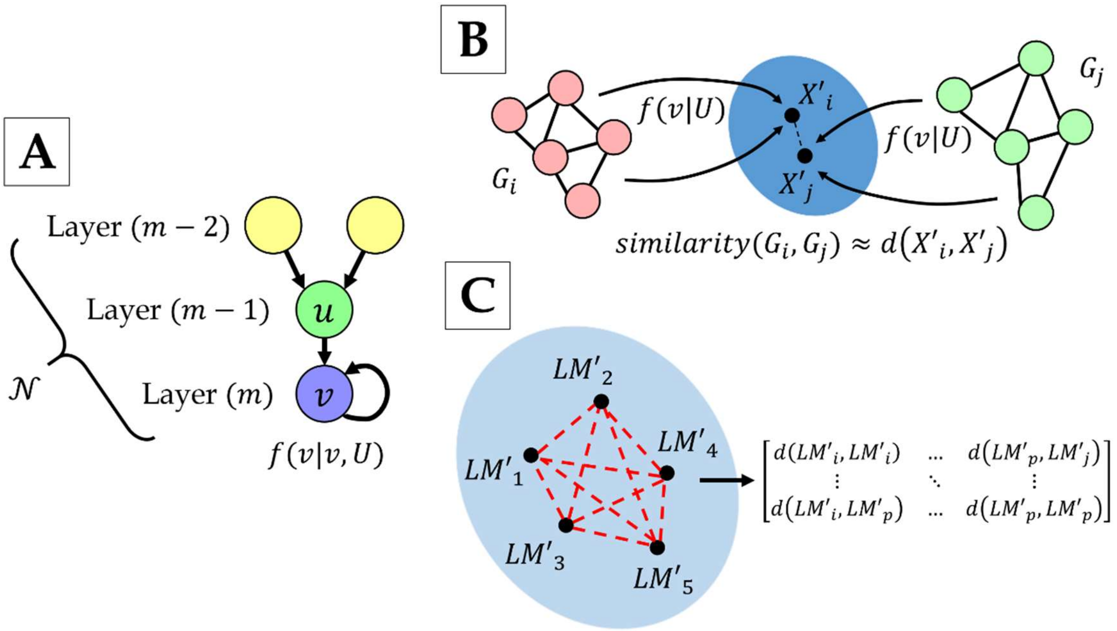

A
Layer $(m-2)$
Layer $(m-1)$
$u$
Layer $(m)$
$v$
$\mathcal{N}$
$f(v|v,U)$
B
$G_i$
$f(v|U)$
$X'_i$
$X'_j$
$f(v|U)$
$G_j$
$similarity(G_i,G_j) \approx d(X'_i,X'_j)$
C
$LM'_2$
$LM'_1$
$LM'_4$
$LM'_3$
$LM'_5$
$\begin{bmatrix} d(LM'_i,LM'_i) & \dots & d(LM'_p,LM'_j) \\ \vdots & \ddots & \vdots \\ d(LM'_i,LM'_p) & \dots & d(LM'_p,LM'_p) \end{bmatrix}$

**Figure 1** – Graphic visualisation of the core concepts behind Graph-based Geometric Morphometrics. (A) The message passing mechanism used to represent each node *v* as a function of itself and its neighbours (*u*). (B) The core concept that the distance between landmark configurations (*d(X'ᵢ,X'ⱼ)*) in the embedded feature space approximates similarity in geometric structure. (C) A visual representation of similarity matrix constructions, representing each landmark configuration as a matrix of similarity measurements, such as *cosine* similarities, between landmark vertices in the embedded feature space.

Let each set of landmarks (*LM*) be represented as a series of vertices $LM_v \in V$, with neighbours $LM_u$, connected by a series of undirected edges in a computational graph *G*. *LM*s should also be connected through a self-looping edge. Each configuration of *G* can thus be represented as an adjacency matrix $A \in \mathbb{R}^{p \times p}$, with feature matrix $X \in \mathbb{R}^{p \times k}$ of Procrustes superimposed landmark coordinates, where *p* is the number of landmarks, and *k* is the number of dimensions. Landmarks can then be embedded using a message-passing mechanism (eq. 1, Fig. 1a), similar to those proposed by Kipf and Welling (2017);

$$v_i^{(m)} = \left(v_i^{(m-1)}, F_{j \in N(i)}\left(v_i^{(m-1)}, v_j^{(m-1)}, e_{j,i}\right)\right) \qquad (1)$$

where $v_i^{(m-1)} \in \mathbb{R}^X$ are the node features of node *i* in layer (*m*-1), with each node *j* and $i \in G$ being connected by an edge $e_{j,i} \in \mathbb{R}^K$.

The definition of $e_{j,i}$ can be computed in a number of ways, either by the use of biological data (*sensu* Adams, 1999), or using vertex spatial distributions (see Section 3 and Supplementary File 1). The spatial attributes of a given landmark in layer *m* can thus be redefined by aggregating the spatial attributes of neighbouring landmarks in layer *m-1*, with the landmark's own spatial attributes, and *N* being the neighbourhood of landmarks (Fig. 1a). *F* (eq. 1) is a differentiable, permutation invariant function that convolves freely over *G* (Kipf and Welling, 2017; Wang et al., 2019). In this application, let *F* be (eq. 2);

$$v_j^{(m)} = \sum_{j \in N(i) \cup \{i\}} \frac{1}{\sqrt{\deg(i)} \cdot \sqrt{\deg(j)}} \cdot v_j^{(m-1)} \tag{2}$$

which can be simplified using adjacency matrix *A*, identity matrix $I_N$ of *A*, feature matrix *X*, and the landmark degree centrality *D* ∀ *v* ∈ *G,* to embed landmarks *X'* into a new feature space (eq. 3-5);

$$\widetilde{A} = A + I_N \tag{3}$$

$$\widetilde{D} = I_N \cdot D \tag{4}$$

$$X' = \left(\sqrt{D} \cdot \widetilde{A} \cdot \sqrt{D}\right) \cdot X \tag{5}$$

Through this, the neighbourhood of a node (*v*) is computed using the square root of the identity matrix to normalise the effect of highly central landmarks in the context of *G* (Kipf and Welling, 2017).

The convolutional operation across *G* is calculated by (eq. 6-8);

$$h_v^0 = x_v \tag{6}$$

$$h_v^{(m)} = \left( \sum_{u \in N(v)} \frac{h_u^{(m-1)}}{|N(v)|} + h_v^{(m-1)} \right), \forall\ m \in \{1, \ldots, M\} \tag{7}$$

$$z_v = h_v^M \tag{8}$$

where *z* is the output of each prior embedding in layer *h* (*sensu* Bruna et al., 2014), with $LM_i'$ ∈ *G* now being represented by feature matrices *X*' in a new $\mathbb{R}^k$ feature space. From this perspective, the attributes of neighbouring landmarks can be passed through the computational graph in a series of convolutional steps. With a single pass, $LM_i'$ is now a function and representation of its own attributes, as well as the attributes of neighbouring vertices (Fig. 1a).

Once projected into the new feature space, landmark configurations can be represented in two different ways; (1) using the raw embedded landmark coordinates, or (2) in terms of a similarity matrix (Fig. 1c). The first representation can be processed using eigendecomposition to visualise patterns in a similar way to traditional GMM approaches, with the new PCA feature space presenting the added advantage of preserving more geometrical and structurally meaningful data from the feature matrix. The second representation, on the other hand, assess and quantifies the degree of landmark homophily and structural equivalence, by computing the proximity of landmarks in the embedded feature space (Grover and Lescovec, 2016; Hamilton et al., 2018). The most common means of calculating this is through the cosine distance (eq. 9), between landmarks X and Y;

$$d(X,Y)=\frac{\sum_{i=1}^{n}X_i\cdot Y_i}{\sqrt{\sum_{i=1}^{n}X_i^2}\cdot\sqrt{\sum_{i=1}^{n}Y_i^2}} \tag{9}$$

This results in a distance value between [0, 1], with 1 describing structurally similar landmarks.

The resulting representations can then be flattened into vectors and processed using different dimensionality reduction techniques. Nevertheless, each approach results in different sized matrices (eq. 10, 11), which may prove computationally expensive to process. From this perspective, the embedded landmark configurations will result in vectors equal in size to those used in traditional GMMs (eq. 10);

$$x\in\mathbf{R}^{p\times k} \tag{10}$$

while similarity matrices result in vectors of exponentially increasing sizes (eq. 11);

$$x\in\mathbf{R}^{\frac{p^2-p}{2}} \tag{11}$$

As opposed to Representation and Geometric Learning applications, where an additional complexity is required in formulae (eq. 1, 2, 6-8) to ensure generalizability (Bronstein and Kokkinos, 2010; Bruna et al., 2014; Henaff et al., 2015; Bronstein et al., 2017; Kipf and Welling, 2017; Murphy et al., 2019; Wang et al., 2019; Xu et al., 2019), the use of Procrustes-based feature matrices as input simplifies the message passing mechanism in a number of ways. Firstly, the laws behind landmark homology ensures a standardised topological structure across graphs, while graph size also remains constant (Bookstein, 1990, 1991; Rohlf, 1996; Dryden and Mardia, 1998). Secondly, Procrustes superimposition procedures remove the need for $F$ to be invariant to vertex rotation and relative position, as Generalised Procrustes Analysis (GPA) already ensures this. Finally, due to points 1 and 2, dataset sizes are no longer as large an issue as no learnable parameters are needed to optimise graph embeddings.

It is also important to point out that the depth of convolutions across graphs is not analogous with the concept of "depth" in computer vision (Henaff et al., 2015). While graph- and image-based convolutions are similar in how their receptive field captures neighbouring nodes or pixels, as opposed to computer vision, the depth of a graph convolution is dependent on the total number of nodes in the graph (i.e., the diameter of $G$). From this perspective, a single convolution represents the neighbours of a vertex, as well as the vertex itself. A second convolution will then represent the neighbours of the neighbouring vertices, as well as the vertex itself (Fig. 2). As we increase the number of convolutions, the embedded feature space will capture more information from vertices further across the graph, which may not be relevant to the vertex itself. In light of this, graph-based convolutions are typically restricted to a few convolutions (typically between two and five), while convolving beyond the natural diameter of the graph is theoretically superfluous.

Additionally, the embedding process can also be considered an approximation to a non-linear exponential transformation (see proofs for Theorem 1, Sup. File 3), yet remains in Euclidean space. In that sense, the first convolution produces a linear transformation of coordinates into a new embedded coordinate space, while subsequent convolutions produce a gradually non-linear transformation.

While this transformation can be considered increasingly "non-linear" with each convolution, coordinates still lie in a Euclidean feature space. From this perspective, performing PCA on the embedded shape variables will still produce a feature space of structural variables that present a linear relationship with the original Procrustes coordinates. Therefore, linear models can be used to predict landmark coordinates, facilitating the visualisation of morphological patterns across the feature space using methods such as Thin Plate Splines (TPS; Bookstein, 1989).

It must be noted that the use of embedded landmarks or *cosine*-similarity matrices present both advantages and disadvantages. The *cosine*-similarity approach is theoretically similar to Euclidean Distance Matrix Analysis (EDMA), in that pairwise relationships are described without the need for superimposition (Lele, 1991; Richtsmeier et al., 1992). Nevertheless, as pointed out by O'Higgins and Dryden (1993), as well as Rohlf (1996), this results in a large increase of variables (eq. 11), and a loss of reflection information. The Graph-based GMM approach overcomes the latter point by deriving matrices from Procrustes data, while *cosine* similarity is used as a means of measuring homophily, not distance. In addition, the use of the *cosine* as a metric for measuring similarity is theoretically and mathematically more adept to the characteristics of shape space (Slice, 2001). Nevertheless, the size of similarity matrices can still prove to be computationally expensive for models with high numbers of landmarks.

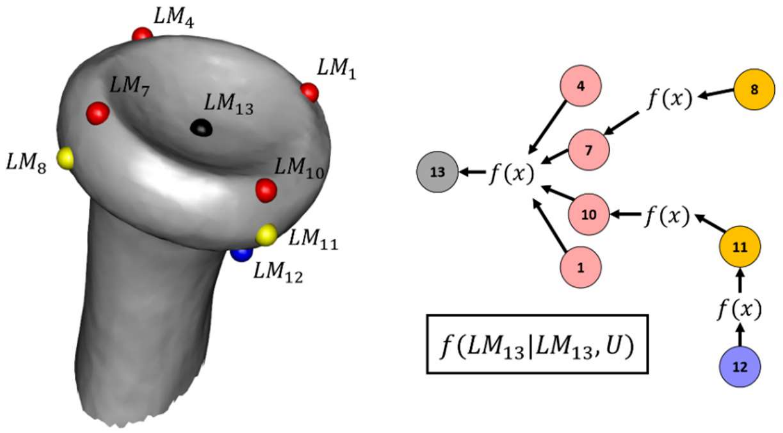


**Figure 2** – Graphic visualisation of the message passing mechanism after three convolutions across a set of landmarks placed on the proximal epiphysis of a chimpanzee radius. Landmark 13 ($LM_{13}$) can thus be represented as a function of itself, and its neighbours ($u$).

### 2.1.2. Mathematical Model: Kernel-PCA Dimensionality Reduction

The process of embedding landmarks is useful for the compression and representation of geometric information. Feature matrices can then be subject to dimensionality reduction, such as eigendecomposition. The present study discusses three main dimensionality reduction techniques that can be applied to; (1) the original landmark data, (2) the embedded landmark data, as well as the (3) embedded *cosine*-similarity matrices. These include; (1) traditional linear Principal Component Analyses (PCA), via single value decomposition; and (2) an adaptation of PCA (kernel-PCA) for the description of non-linear relationships using kernel functions (Schölkopf et al., 1998).

Traditional GMM applications use PCA as means of representing multivariate relationships efficiently (Rohlf, 2000; Klingenberg and Monteiro, 2005). Nevertheless, PCA is a dimensionality reduction technique that works by finding linear and orthogonal relationships, thus proving to be statistically parametric in nature. As is common-knowledge, statistical normality and linearity is not always present in real-world applications (Diaconsis and Freedman, 1984), implying that PCA may not be the most efficient means of representing certain types of data.

Let each individual landmark configuration be represented as a vector $\vec{x}$ in dataset $X$, with features $\{x_{i,1}, \ldots, x_{i,n}\}$. The objective of PCA originally would be to calculate the covariance matrix of a centred and scaled version of $X$, followed by the eigendecomposition of this matrix. This transformation is linear in nature. Nevertheless, non-linearity can be imposed prior to decomposition through the use of kernel functions, $(x_i, X_{:, n})$ (Bishop, 2006).

Many kernels exist for the non-linear transformation of data, almost all of which are computed considering the spatial proximity of each point with the rest of $X$. From this perspective, we define $\Delta$ as the distance between each point with the rest of the observations in each dimension; $x_i - X_{:,n}$ . Once $\Delta$ has been defined, the kernel transformation of each individual can be defined through each of the kernels described by;

$$\mathrm{K}\left(x_i, X_{:,n}\right) = e^{\frac{-\|\Delta\|^2}{2\alpha^2}} \tag{12}$$

$$\mathrm{K}\left(x_i, X_{:,n}\right) = e^{\frac{-\|\Delta\|}{\alpha}} \tag{13}$$

$$\mathrm{K}\left(x_i, X_{:,n}\right) = \left\|X_{:,n} x_i\right\|^d \tag{14}$$

$$\mathrm{K}\left(x_i, X_{:,n}\right) = e^{-\gamma\|\Delta\|^2} \tag{15}$$

$$\mathrm{K}\left(x_i, X_{:,n}\right) = 1 + X_{:,n} x_i + X_{:,n} x_i \min\left(X_{:,n}, x_i\right) - \frac{X_{:,n} + x_i}{2} \min\left(X_{:,n}, x_i\right)^2 + \frac{1}{3} \min\left(X_{:,n}, x_i\right)^3 \tag{16}$$

$$\mathrm{K}\left(x_i, X_{:,n}\right) = \frac{2\pi}{1 + \|\Delta\|^2} \tag{17}$$

where equation 12 is the Gaussian kernel (Rahimi and Recht, 2007), equation 13 is the Laplacian kernel (Rahimi and Recht, 2007), equation 14 is the Polynomial kernel (Bishop, 2006; Cortes and Vapnik, 1995), equation 15 is the Radial Basis Function kernel

(Bishop, 2006; Cortes and Vapnik, 1995), equation 16 is an adaptation of the Spline kernel (Gunn, 1998), and equation 17 is an adaptation of the Cauchy kernel (Rahimi and Recht, 2007).

Considering how there are multiple means of describing non-linear relationships, the majority of these functions are dependent on a hyperparameter (α, *d*, γ), defining the degree of the transformation. Values approaching 0 thus imply an increasingly linear transformation (Fig. 3). Equations 16 and 17 are yet to be adapted into tuneable non-linear functions in the present line of research. Nevertheless, a trainable version of the Cauchy kernel (eq. 17) does exist in neural-network applications (Rahimi and Recht, 2007).

Once coordinates are projected into the new feature space, PCA is performed for the extraction of final eigenvectors and eigenvalues.

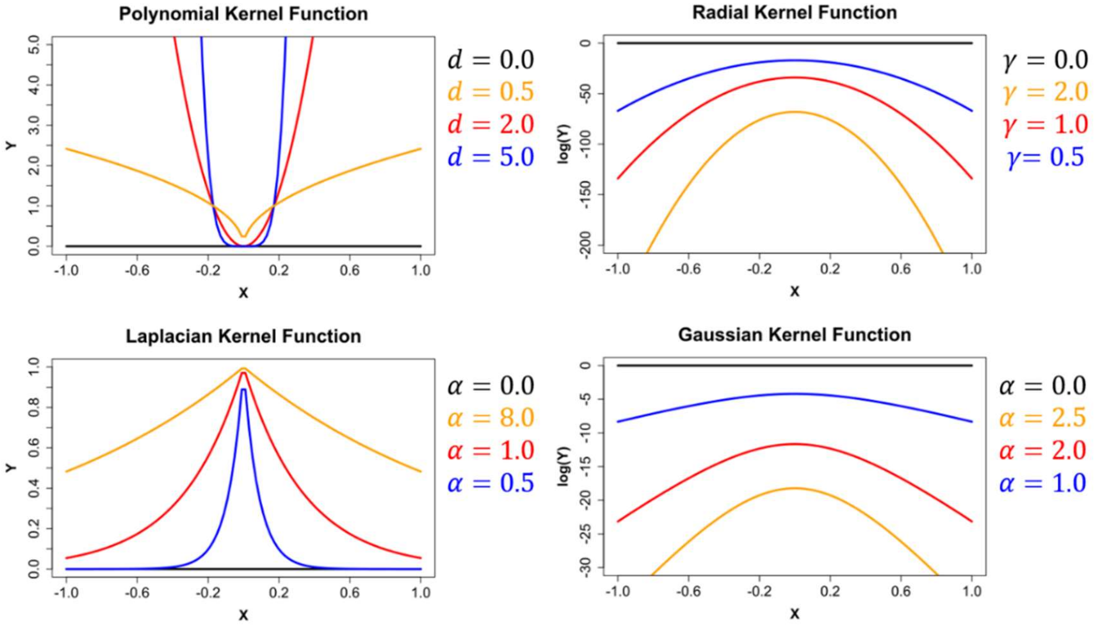


**Figure 3** – Graphic demonstrations of the types of non-linear transformations that can be performed using Kernel functions prior to dimensionality reduction.

### 2.1.3. Additional Reflections on Graph Theory

The advantages of representing GMM data as a set of graphs goes beyond embedding and dimensionality reduction techniques for ordination. Graph theory presents numerous possibilities for the analysis of landmark configurations as a whole, once transformed into a connected graph. From this perspective, metrics such as Eigenvector, Betweenness and Degree centrality calculations can indicate landmark importance in the context of *N*, and more importantly *G*. These metrics have been found to be particularly important in defining the degree of transformation observed during embedding processes (see proofs for Theorem 1, Sup. File 3). Likewise, landmark community detection using different algorithms (e.g., the Louvian algorithm: Blondel et al, 2008), can be used for a more automated and empirically defined modularity analyses (Klingenberg, 2009). Finally, the representation of data as a structured graph can be considered a valuable contribution to morphological visualisation techniques, alongside other approaches (some similar), proposed by Bookstein (1978, 1989), O'Higgins and Dryden (1993), among others.

### 2.2. Sample

The present study demonstrates the applicability of Graph-based GMM using three open-source case studies, available from the geomorph (v.4.0.2; Adams et al., 2021) and shapes (v.1.2.6; Dryden, 2021) R libraries. These include the graph-based analysis of; (1) 167 large primate skulls described using 2D 8-landmark models, originally presented by O'Higgins and Dryden (1993); (2) 40 *Plethodon jordani* and *Plethodon teyahlee* heads described using 2D 12-landmark models, originally described by Adams (2004, 2010); (3) 18 Macaque skulls described using 3D 7-landmark models, originally described by Dryden and Mardia (1998). In each of these case studies, Supplementary File 1 presents how graph embeddings can be used in replicable case-studies, demonstrating a number of different techniques for Graph-based GMM.

- In case study 1, we show how Graph-based GMM is able to describe greater morphological variability across multiple PC scores, while the use of *cosine*-similarity matrices reveals a new source of information for the description of morphological variability.
- In case study 2, we show the different possible techniques available for the conversion of landmark data into graphs, as well as the different patterns these techniques may reveal.
- In case study 3, we show how these techniques are applicable to both 2D and 3D data, as well as the statistical differences that can be revealed when using kernel-PCA approaches.

Additionally, for the purpose of the present study, 3D models of 84 modern great ape radii, including chimpanzees (*n* = 20), gorillas (*n* = 17), orangutans (*n* = 17) and anatomically modern humans (*n* = 30), were analysed using the mathematical models provided in the GraphGMM library. The materials used in this study come from different open-access online archives (e.g., MorphoSource, KUPRI, NMDID), where either the raw CT data, or the rendered 3D volumes from the CT files, are available (see Table S1 in Sup. File 4 for further details).

Image data were processed based on the available information to generate 3D meshes of the external surface of the radii, including volume cropping, segmenting, rendering, translation and mirroring when necessary. For this purpose, the morphomap R package (Profico et al., 2021), and Avizo software (Solid Works, Visualisation Sciences Group, USA), were used.

Radii were landmarked in two steps using the EVAN Toolbox (http://evan-society.org/). First, 41 fixed anatomical landmarks (Fig. 4) were identified following previous works (Pérez-Criado and Rosas, 2017). Most of these points are located on the epiphyses (Table S2 in Sup. File 4). A net made of 160 points (10 fixed landmarks and

150 sliding semilandmarks) was then projected and slid along the diaphysis, covering from below the radial tuberosity to the proximal point of the sigmoid notch (Fig.4; see Sup. File 4 for further details). The sliding process proposed by Bookstein (1991, 1997) based on the TPS formalism was used to minimise the bending energy between a calculated mean specimen used as reference and the rest of the sample through repetitive sliding of all semilandmarks simultaneously. The application of this technique accounts for local shape deformation using a mathematical model that interpolates the space found between two sets of homologous landmarks as smoothly as possible. The sliding process is constrained by the adjacent fixed landmarks and semilandmarks that act as starting point for the sliding algorithm (Gunz and Mitteroecker, 2013; Mitteroecker and Gunz, 2009). The sliding process was repeated five times per specimen in order to get the best minimisation of the bending energy.

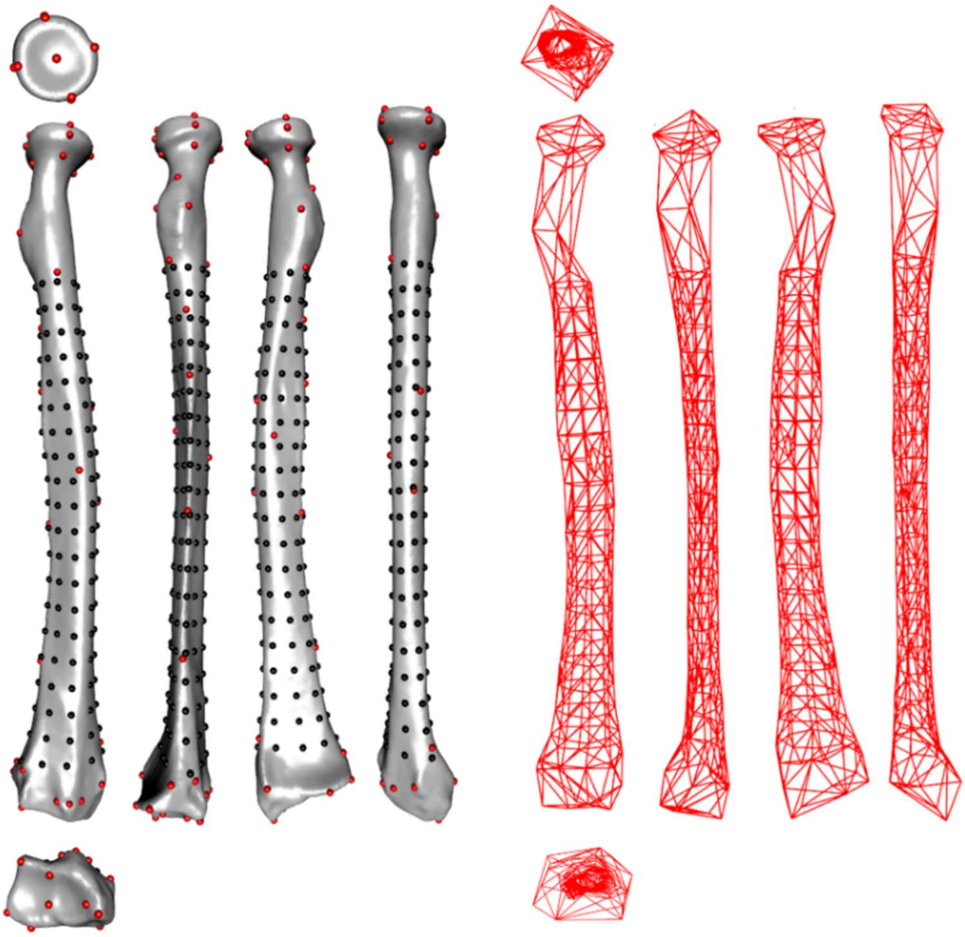


**Figure 4** – (Left) Example of both fixed (red) and sliding semilandmarks (black) placed on an example of an Anatomically Modern Human radii, alongside the constructed graph using pivot-ball triangulation (Right).

Once digitised, landmarks were superimposed using full GPA both in shape (with scaling) and form (without scaling), producing a set of orthogonally projected Procrustes coordinates. An experiment was also performed excluding semilandmarks, however the majority of this study includes them. For graph construction, we used a pivot ball triangulation algorithm (radius ρ = 25.5: Bernardini et al. 1999) on the median shape and form configuration, using triangulation so as to compute landmark relationships spatially. The characteristics of each of the computed graphs are explained in Table 1. Graphs were then embedded using two convolutions. The resulting landmark embeddings were used for dimensionality reduction, according to the optimal means of representing this type of data. Experiments were performed calculating both linear and kernel PCA on the original Procrustes coordinates (here referred to as Original PCA), landmark embeddings, and similarity matrices.

**Table 1.** Descriptive statistics and characteristics of landmark graphs in shape and form, while including or excluding semilandmarks.

| | **Landmarks & Semilandmarks** | | **Landmarks** | |
|---|---|---|---|---|
| | Shape | Form | Shape | Form |
| Graph diameter | 23 | 22 | 4 | 4 |
| Clustering coefficient | 0.394 | 0.393 | 0.474 | 0.483 |
| Graph Density | 0.032 | 0.031 | 0.121 | 0.115 |
| Mean LM degree | 6.45 | 6.35 | 4.83 | 4.59 |

## 3. Results

Initial observations of PCA distributions reveal sample scattering among the different approaches to present more pronounced differences in shape space (Fig. 5), as opposed to form space (Fig. 6). This is likely due to the weight allometry has on the first principal components in any GMM analysis (Jungers et al., 1995), and the influence Centroid Size has on the apparition of the Pinocchio effect (Klingenberg, 2021, *inter alia*; Fig. S13 and proofs for Theorems 1 & 2, from Sup. File 3). Sample distributions are relatively stable across reduction techniques in shape space, including degrees of overlapping among samples, and the established variational relationships among groups. From this perspective, chimpanzees and gorillas slightly overlap on all accounts, while AMH and orangutans constitute their own independent clusters. AMH and orangutans only present a slight overlap in the highly explanatory Polynomial Graph-based PCA (Fig. 5). Despite certain differences in general robusticity and straightness of radii, the majority of morphological changes expressed by PC1 tend to be relatively stable, with all types of PCA representing similar traits.

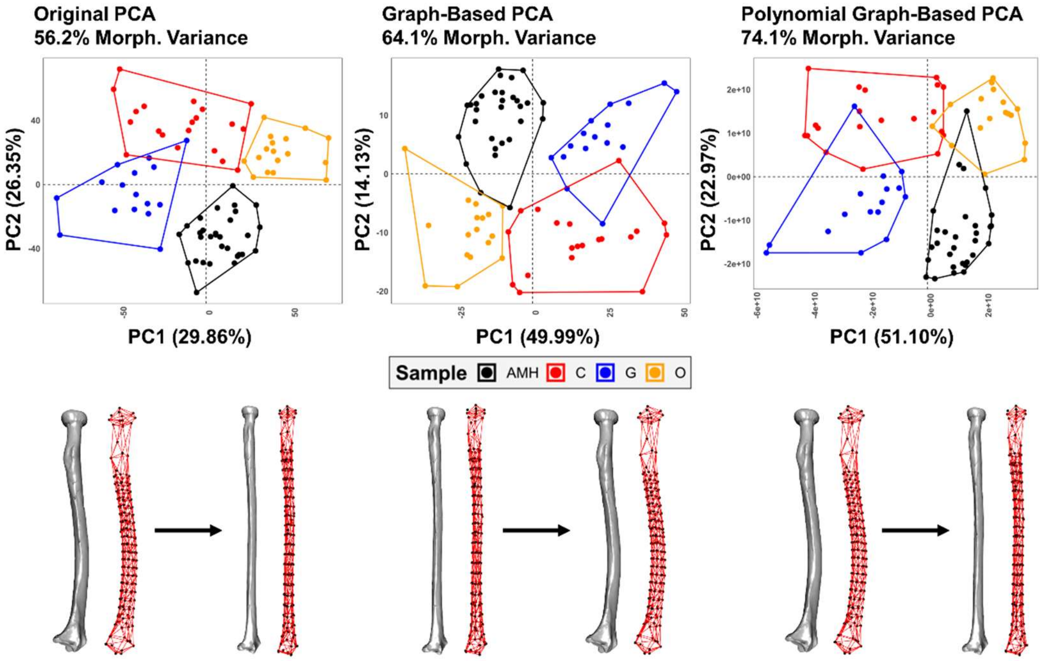


**Figure 5** – Examples of the three different Dimensionality Reduction results in shape space using traditional Principal Component Analyses in Geometric Morphometrics, as well as linear and non-linear Graph-based Principal Component Analyses to analyse the radius morphology in great apes using anatomical landmarks and sliding semilandmarks. Changes in shape space are visualised across the extremities of PC1. AMH = Anatomically Modern Human, C = Chimpanzee, G = Gorilla, O = Orangutan.

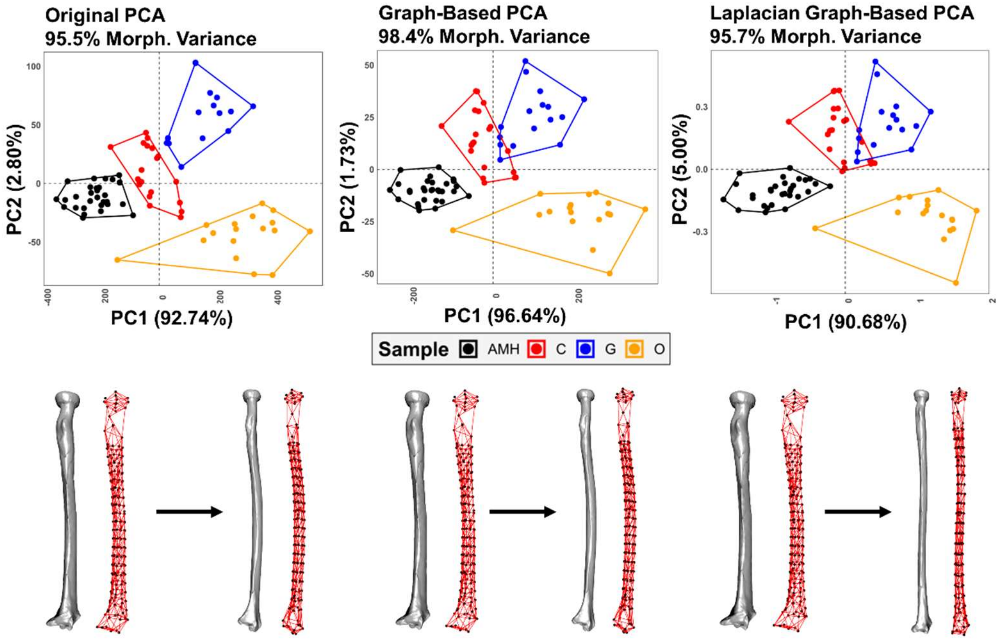


**Figure 6** – Examples of the three different Dimensionality Reduction results in form space using traditional Principal Component Analyses in Geometric Morphometrics, and linear and non-linear Graph-based Principal Component Analyses, to analyse the radius morphology in great apes using anatomical landmarks and sliding semilandmarks. Changes in form space are visualised across the extremities of PC1. AMH = Anatomically Modern Human, C = Chimpanzee, G = Gorilla, O = Orangutan.

The largest differences among analyses, however, can be observed in dimensionality reduction performance. The original PCA-based method explains slightly more than 50% of the total variance along the first two PCs, whereas the Graph-based technique on the embedded data (64.1%), and the non-linear Graph-based PCA on the *cosine*-similarity matrices (74.1%), using a polynomial kernel ($d$ = 10), are able to increase the amount of variance contained in the first two components. The original PCA reaches the same percentage of variance explained in the linear and non-linear Graph-based versions of the PCA in the first three and five PCs, respectively. Graph-based methods also imply a considerable reduction of morphologically significant variables (Fig. 7), particularly in the non-linear version of the analysis, where significant explanatory power is provided by just one PC.

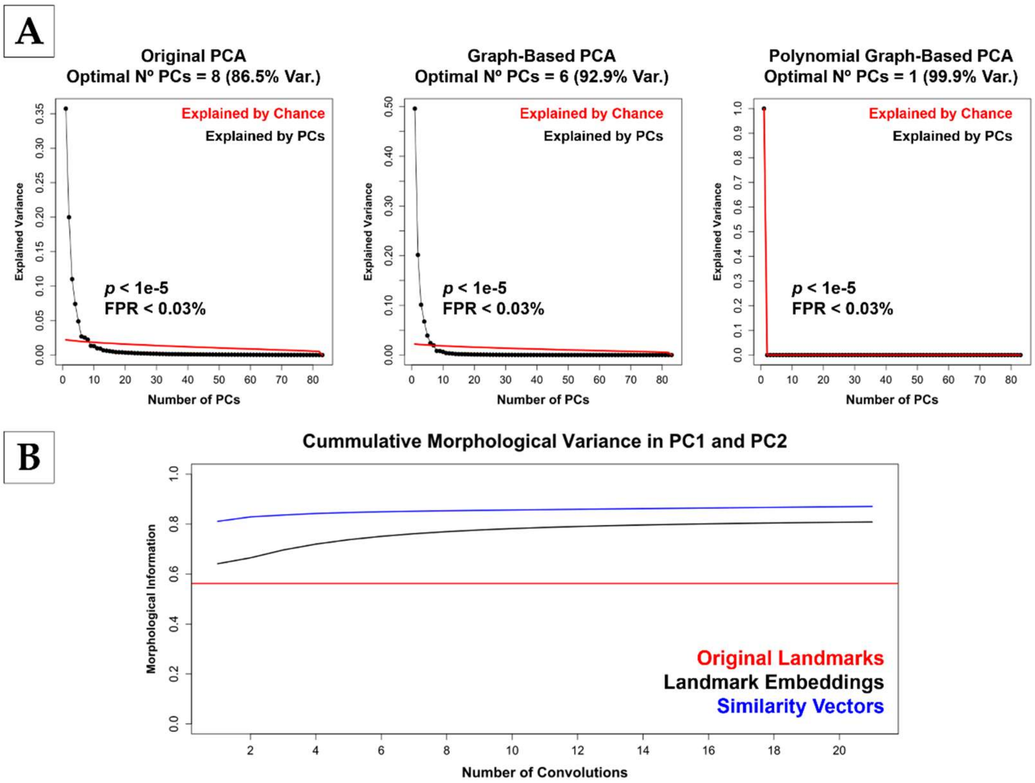


**Figure 7** – Plots describing the percentage of morphological variance represented across Principal Component (PC) Scores using each of the different techniques. (A) Comparisons of percentage of morphological variance explained by PC scores and the percentage of morphological variance explained by chance after 10,000 permutations. FPR = False Positive Risk. (B) Percentage of cumulative morphological variance in PC1 and PC2 when performing PCA on the original landmarks, the embedded landmarks, and the similarity vectors, after different numbers of graph convolutions.

Form PCAs do not vary that extremely in terms of explained variance in PC1 *versus* PC2 space, although a slight increase in explanatory power can be observed in the two Graph-based methods. This is especially relevant when comparing the original PCA with the graph-based PCA, where PC1 presents an increase by 3.9%. Thus, contrary to the results in shape space, linear approaches on embedded data in form space provide a better reduction of morphologically significant variables than the non-linear Graph-based PCA using a Laplacian kernel ($\alpha = 5000$). In general, graphical sample distribution is very similar regardless of the dimensionality reduction technique, with orangutans and AMH being clearly separated in form space, while chimpanzees and gorillas tend to slightly overlap when larger percentages of variance are considered. Differences in the morphological variance expressed by PC1 are more pronounced in the non-linear Graph-based PCA, where the slenderness and straightness of the radius is more marked towards the positive extremity of PC1, as opposed to a more robust radius towards the negative extremities, where AMH are located (Fig. 6). PC1 in the original and linear Graph-based methods do not only account for a larger percentage of form variance than PC1 in the non-linear Graph-based method, but also contain more similar morphological features, that diverge from the extremely slender and straight representation of the radius in the Laplacian approach, by providing a slightly more curved, less slender midshaft portion, and a more angled-positioned distal epiphysis towards the positive PC1 axis end.

When considering PCA biplots (Fig. 8), so as to understand the specific variables conditioning sample distributions, traditional PCA reveals shape space to be conditioned strongly by displacements in the y coordinates of landmarks 1, 4, 10, 21 and 22. This is relevant considering landmarks 1 to 10 are located on the head of the radius, while landmarks 21 and 22 delimit the proximal and distal extremes of the pronator teres attachment (Fig. 8). In the non-linear Graph-based PCA, sample distributions are determined by the structural relationship between pairs of landmarks. The most significant pairs contain landmarks located in the midshaft of the radius (see Table S2 in Sup. File 4), of which structural relationships are more significant in the human radius (Fig. 8). Landmark 77 plays a pivotal role, and its structural relationship with the rest of the landmarks seems to be particularly relevant, seeing how the established links are equally important in defining each specimen's position along the first two PCs. On the contrary, in original PCA space, eigenvectors y21 and y22 seem to play a major role in morphological variance along PC2, whereas differences in the head of the radius are represented by eigenvectors y1, y4 and y10, carrying almost the same weight along PC1.

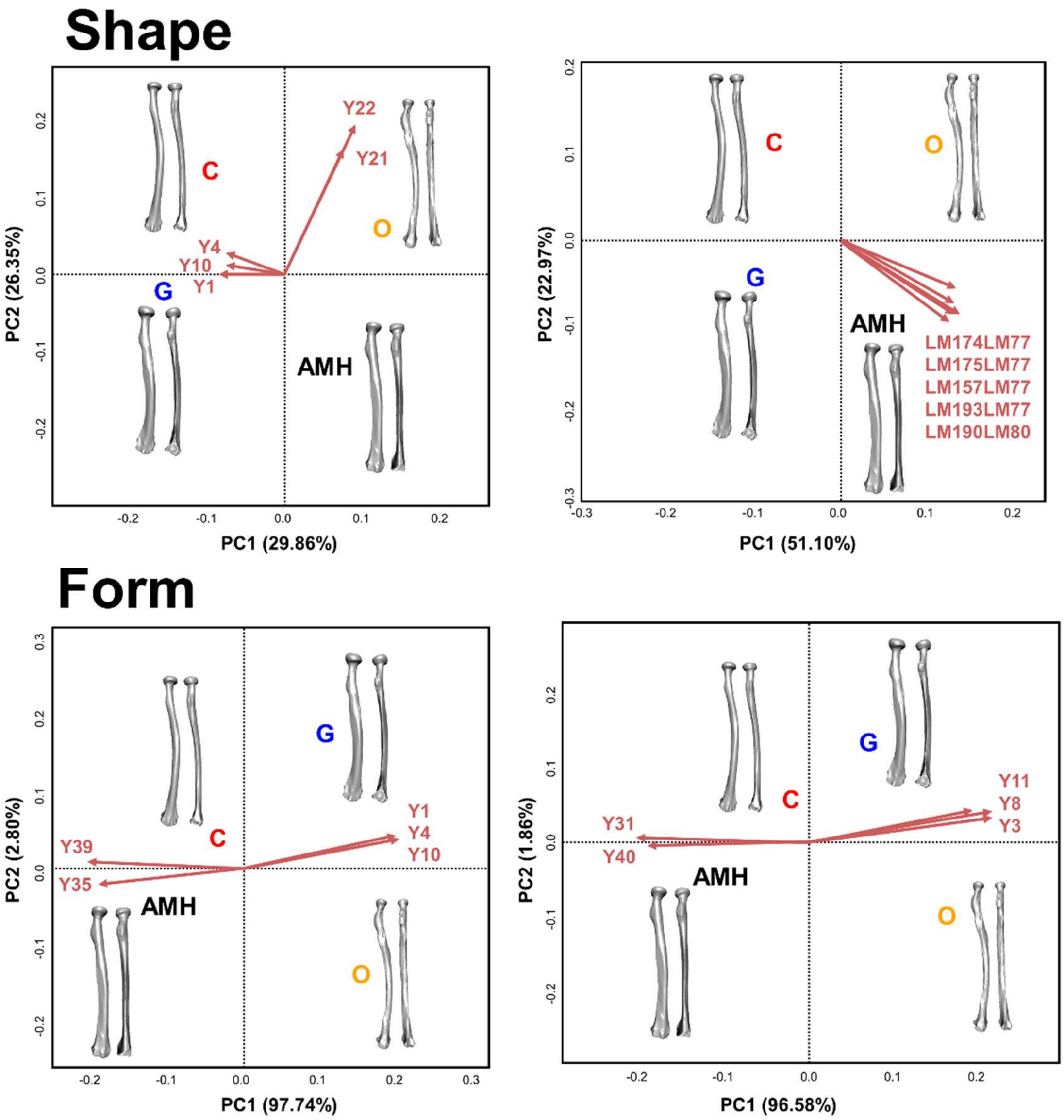


**Figure 8** –PCA biplots comparing (Left panels) dimensionality reduction performed on the original landmark data, and (Right panels) the best performing Graph-Based PCAs. 3D models represent the calculated median configurations for each group. AMH = Anatomically Modern Human, C = Chimpanzee, G = Gorilla, O = Orangutan.

Such differences are also palpable when great ape groups are compared in pairs (Fig. 9, Fig. S2 in Sup. File 4), though in most cases differences tend to be subtle, with the non-linear Graph-based approach providing more marked differences among the sample as a result of a more effective compression of information. Most differences are assembled on the proximal and distal portions of the bone, especially in the head and neck of the radius, the radial tuberosity, the sigmoid notch, the scaphoid, and the styloid process. Many of the differences are not only related to changes in relative proportions (e.g., distance between the head and the radial tuberosity), but also in the orientation and location of certain osseous landmarks (e.g., the angle of the sigmoid notch relative to the bone’s longitudinal axis, the position of the radial tuberosity relative to the interosseous crest). Although the diaphysis presents less abrupt changes, important differences in curvature and width are observed among groups. Morphological differences along the shaft are variably marked based on the PCA approach (as expected from the results in Fig. 8), though in general terms midshaft curvature degree along the posterior view presents the most notable changes on the diaphysis. In certain cases, however, important differences are also observed in the distal diaphyseal portion (e.g., when comparing orangutan *versus* gorilla).

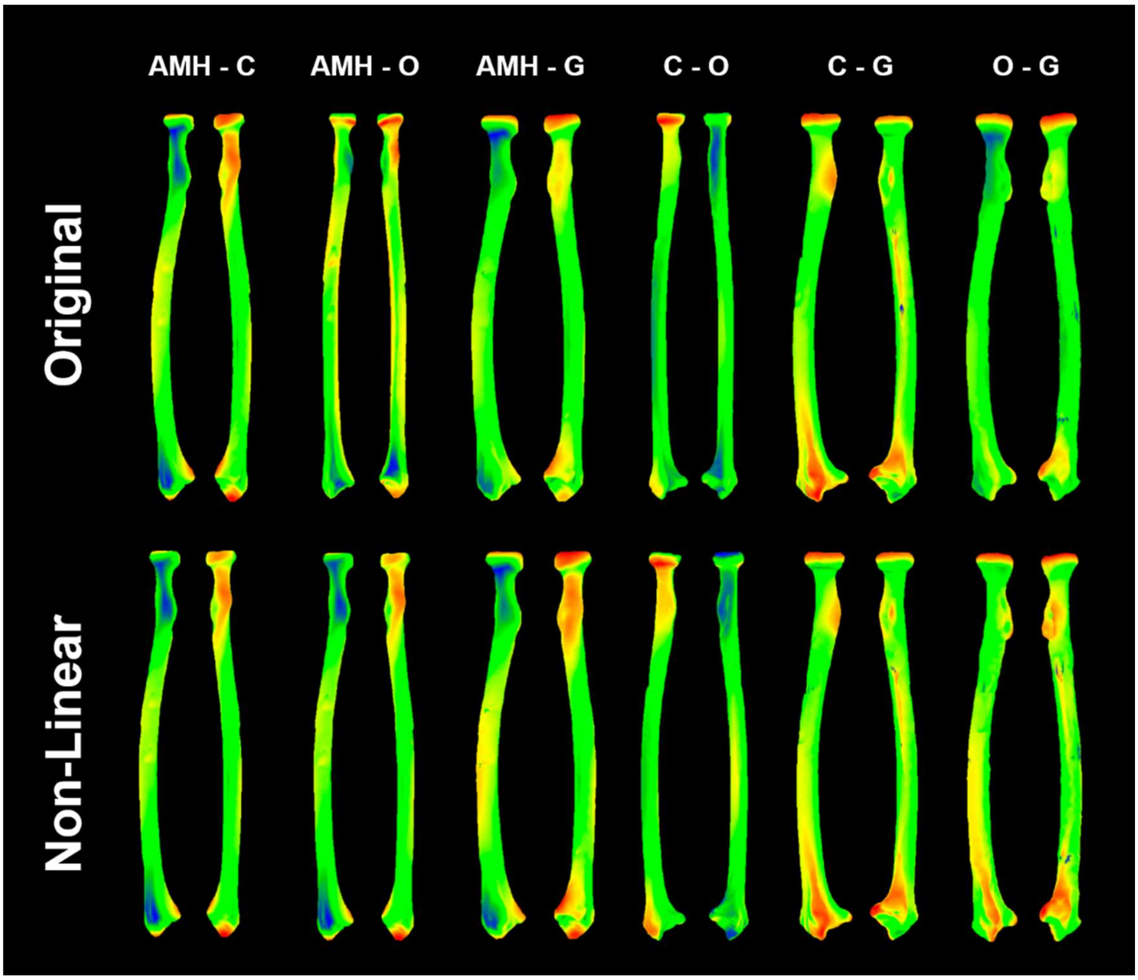


**Figure 9** – Calculations of the differences between mesh warpings when predicting median shape changes using traditional Geometric Morphometrics (Upper) and *cosine*-similarity matrices (Lower). Heat maps indicate areas where mesh warpings differ from the original Thin Plate Spline prediction (red = positive deformations, blue = negative deformations). AMH = Anatomically Modern Human, C = Chimpanzee, G = Gorilla, O = Orangutan.

As expected from the homogeneity observed across form PCA biplots in Figure 6, the exploratory power of morphological features underlying variance in PC1 *versus* PC2 form space are also similar, with displacements in singular coordinate values located in the distal (y39, y35; y31, y40) and proximal (y1, y4, y10; y3, y8, y11) epiphyses, conditioning sample scattering along PC1. This structure is clearly associated with the importance of size in form space, in this occasion mainly related to differences in length, with orangutans presenting the longest radii in opposition to AMH, who have the shortest radii among modern great apes in relative and absolute terms.

When the morphology of the radii in different great ape groups is analysed by means of fixed anatomical landmarks (most of them situated on the proximal and distal epiphyses), more strongly marked changes in shape space can be observed depending on the dimensionality reduction technique (Fig. 10). The traditional GMM approach results in a biplot that does not resemble those seen in Figure 5, but instead presents different clustering patterns, with AMH clearly separated from nonhuman great apes, and orangutans and gorillas overlapping with chimpanzees, which fall right in between the former groups. On the other hand, Graph-based PCAs show a sample distribution that resembles those seen in Figure 5 across all PCAs, with four distinct clusters slightly overlapping in pairs formed by AMH and orangutans, on one side, and chimpanzees and gorillas, on the other side. Hence, Graph-based PCAs, apart from achieving a more comprehensive two-dimensional graphical representation of morphological variance, tend to be more stable regardless of the data input. Traditional GMM PCAs, on the other hand, are more volatile and dependent upon the availability on morphological information. When comparing the warped surfaces using both methods, differences are most noticeable in the way of capturing changes in diaphyses (Fig. 5).

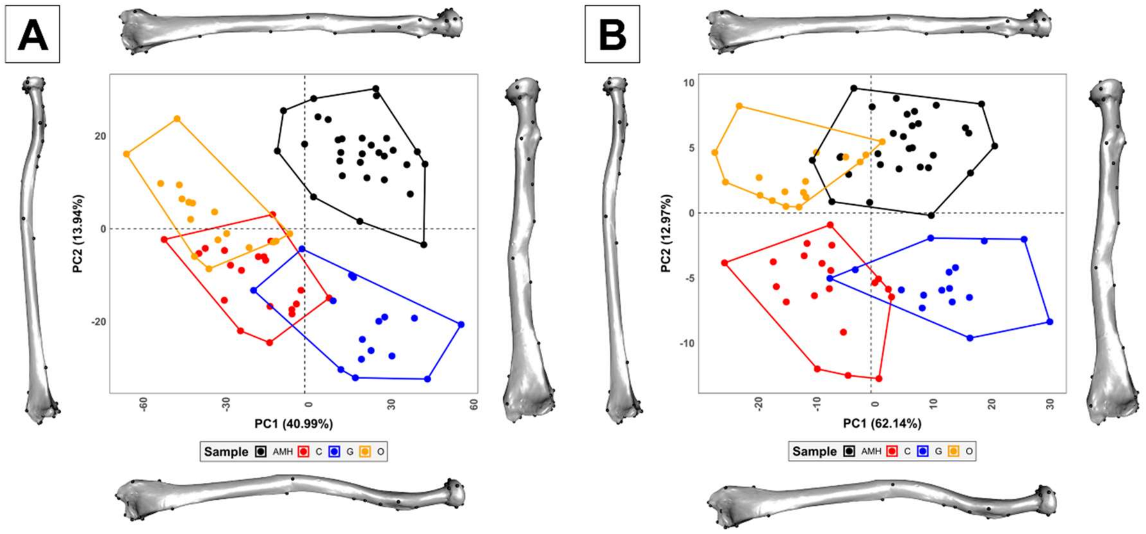


**Figure 10** – Dimensionality reduction on fixed anatomical landmarks using; (A) traditional PCA in Geometric Morphometrics and (B) linear Graph-based PCA. Changes in shape space are visualised across the extremities of PC1 and PC2. AMH = Anatomically Modern Human, C = Chimpanzee, G = Gorilla, O = Orangutan.

Differences in PCA performance based on the technique are not that evident in form space (Fig. S3 in Sup. File 4), though great ape groups appear to be graphically closer in the highly explanatory PC1 *versus* PC2 space resulted from the Graph-based technique (99.32%). Separation between groups in form space is less marked when semilandmarks are not considered, regardless of the approach. Once again, this is proof of the weight size has on the type of information represented in each type of PCA.

Finally, experiments adjusting the number of convolutions represented in Figure S4 (Supplementary File 4) show how adjusting this parameter can fine tune the visualisation of regional or global structural features as convolutions increase. In the case of the present study, just 2 convolutions is sufficient in capturing a large percentage of morphological variance, however other studies may wish to weight their feature spaces according to the research questions at hand.

## 4. Discussion

### 4.1. Graph-Based methods in Geometric Morphometrics

Dimensionality Reduction (DR) techniques are an extremely useful tool in data science for a more efficient representation of information. They are additionally frequently used as a means and basis of pattern recognition (Bishop, 2006). PCA is an essential means of performing DR in GMM, as it extracts vectors from a large set of variables (eq. 10), that allow for the compression of information into fewer dimensions. This additionally maintains the Procrustes distances between the different specimens, accounting for as much of the original variation as possible. The objective of DR in GMM should therefore be to compress information into as little and as meaningful dimensions as possible, presenting a more manageable dataset describing morphological, structural, biomechanical and evolutionary patterns, as well as providing a direct means of testing hypotheses regarding this data.

Eigendecomposition reduces matrices into constituent parts which successively explain decreasing proportions of the total variance. This orthogonal projection of data onto a lower dimensional principal subspace also has the additional function of normalising data, however this does not necessarily guarantee a Gaussian feature space (Diaconsis and Freedman, 1984). In most cases, this technique is used to extract the most relevant information from a dataset in the form of the first few PC scores of PCA, thus removing any residual noise that may impede more complex analyses. Nevertheless, the elimination of the "least important" PC scores often raises concerns, as no unique accepted rule exists for this selection (Jolliffe, 2002). In many types of analyses, the first few PC scores may not explain all of the necessary variance to capture true morphological patterns. The only exception to this is if a strong distinctive feature is existent, such as the variable size. However, in these cases, form feature spaces are mostly biased, with the first PC being explained almost exclusively by size, with little explanatory power in terms of shape.

The present study parted from the hypothesis that traditional approaches to DR in GMM result in the slight loss of information in higher dimensions of principal subspace. In general, the majority of the feature spaces described in this study seem to prove this hypothesis, as almost all graph-based approaches describe higher levels of variance in just two PC scores as opposed to other methods. If we consider the type of information represented in PCA biplots derived from coordinate data, most trends appear to be analysed in terms of slight displacements of single coordinate values in only one dimension (e.g., the most important variable of the first panel of Fig. 8 is the movement of LM22 along the y axis). Our original hypothesis considered this to be a poor representation of the landmark configuration as a whole, as this single eigenvalue loses sight of how a singular displacement may affect the point's neighbours. Nevertheless, it could also be argued that this statement is an oversimplification of the problem at hand.

The displacement of even a single landmark, no matter how small, will have a consequent effect on the entire configuration during superimposition procedures. This, in part, is a product of a phenomenon known as the *Pinocchio effect* (Chapman, 1990; Walker, 2000; Hallgrimson et al., 2015; Klingenberg, 2021; Supplementary File 3; Theorem 1 & 2). From this perspective, it can be argued that variations in morphology cannot be ascribed to an individual landmark, but more to the relationship between the landmarks (Klingenberg, 2021). This is due to how landmark displacements consequently adjust the position of the configuration's centroid as well, resulting in a change not only in the point of reference for superimposition procedures, but also in centroid size (Supplementary File 3, Theorem 2 & 3). Nevertheless, while this is true during Procrustes analyses, the moment this information is subject to PCA, the constructed feature space loses sight of these processes.

While the objective of any multivariate analysis is the evaluation of multiple variables, and no single PC score alone, the embedding procedures described here are more likely to encode inter-landmark and global relationships prior to PCA (Supplementary File 3, Theorem 3). The embedding procedure is thus able to capture geometric properties of coordinate data, prior to any type of GMM analyses (Supplementary File 3, Theorem 3).

Finally, it is common knowledge that few real-world cases comply with the assumptions made by parametric statistics. The present study has attempted to present just some of the methods available for analysts to extract and visualise non-linear relationships from morphological data. In most cases, non-linear PCAs have additionally proved to present the most efficient means of DR, representing the highest degree of morphological variance. The only downside is the possible "subjective" nature of selecting and fine tuning a kernel function for these purposes.

### 4.2. The link between form and function in the radius of extant great apes

From a locomotor point of view, modern great apes could first be split into two groups: (1) humans and (2) nonhuman great apes. Humans are characterised by possessing an  upper limb free from locomotor impositions, while in the latter group the arm, forearm and hand are involved in day-to-day locomotion, although to different extents and in different ways. In fact, there are two main patterns among nonhuman great apes that also help distinguish between extant African and Asian great apes: the larger the body mass, the less arboreal are great apes, as in gorillas; while the more arboreal, the more elongated is the upper limb, especially the forearm, as in orangutans.

Most anatomical specialisations shared among great apes are concentrated on the trunk and upper limbs, in response to the requirements imposed by positional and locomotor behaviours, as well as the acquisition of food that might appear differently dispersed in the landscape (Aiello, 1981; Fleagle, 1998; Larson 1998; Young, 2003; Hunt, 2016; Schmitt *et al*., 2016). Such shared features have been originally associated to suspensory behaviours between discontinuous supports that, nevertheless, encompass a wide range of locomotor patterns and thus require different adaptations (Hunt, 1991; Fleagle, 1998; Ward, 2007). Terrestrial locomotion is also important among great apes. Not only humans move on the ground, but also nonhuman great apes have developed particular ways of dealing with the stresses imposed by terrestrial locomotion, and through adopting different quadrupedal strategies (Tuttle, 1967, 1969; Susman, 1974; Doran, 1992; Schmitt *et al*., 2016).

Orangutans present extreme specialisations (e.g., very long forelimbs, hook-like hands and feet, short and extremely mobile lower limb), adapted for suspensory locomotion in the upland forest areas, including hand-foot hanging, quadramanous climbing, bridging or transferring (Cant, 1987; Thorpe and Crompton, 2005, 2006; Thorpe *et al*., 2009; Manduell *et al*., 2011), whereas no significant anatomical features have been associated with quadrupedal fist-walking.

In accordance with their large bodies, gorillas are the most terrestrial nonhuman great apes, though arboreal locomotion has been also detected in variable degrees among species (e.g., the western lowland gorilla *versus* the mountain gorilla). This is especially relevant amongst the smaller specimens, namely females and subadults (Remis, 1995; Doran, 1997; Doran and McNeilage, 1998; Fleagle, 1998; Remis, 1998). When on the ground, gorillas are knuckle-walkers (Schmitt *et al*., 2016). Gorilla's locomotion repertoire is reflected in its anatomy, characterised by the presence of a short and broad thorax, accompanied by long forelimbs relative to the short hindlimbs.

Chimpanzees are somehow intermediate between gorillas and orangutans and have adopted a "compromise" anatomy that is neither fully adapted to the demands imposed by arboreal nor terrestrial locomotion (Hunt, 2016). While suspensory behaviours such as vertical climbing or arm-hanging are vital for fruit harvesting (Susman *et al*. 1980; Doran, 1993; Thorpe and Crompton, 2006), terrestrial and on-branches knuckle-walking is one of the main locomotion modes recorded among chimpanzees (Tuttle, 1967, 1969; Inouye, 1989; Doran, 1992; Schmitt *et al*., 2016). In order to practise efficient suspensory behaviours, chimpanzees display more balanced fore-to-hindlimb lengths; narrower scapulae, hands and feet; and more markedly curved and slenderer digits, in comparison to gorillas, which present anatomical compromises that might have decreased the efficiency during knuckle-walking (Pontzer and Wrangham, 2004).

On the other hand, bipedalism in humans provides a particular advantage by freeing the upper limb from locomotor purposes, and thus enabling the appearance of adaptations for additional functions (Napier, 1956; Marzke, 1986). The hyper-specialised locomotor repertoire and anatomy in modern humans alongside its more in-depth study, has helped establish more clear links between form and function in the human musculoskeletal system, whereas the more varied locomotor repertoire in nonhuman great apes often makes it difficult to link certain biomechanical behaviours to specific anatomical morphologies (Rose, 1991; Tallman, 2012; Thomson, 2021).

The generalised division among modern great apes based on the preferential use of the landscape (terrestrial *versus* arboreal), and the main modes of locomotion (e.g., bipedalism, suspension, quadrupedal walking), is well-reflected in the results obtained in the present study.

Morphological similarities in the radius between chimpanzees and gorillas seem to be the most strongly marked among the four groups, probably in accordance with the weight-bearing function of the radius during knuckle-walking in both groups. Differences in the range of movement at the elbow and wrist joints are noticeable among modern humans and apes, being movements in the latter group much more restricted due to the constraints imposed by locomotion (Carlsoo and Johansson, 1962; Tuttle, 1967, 1969; Jenkins, 1973; Jenkins and Fleagle, 1975; Corruccini, 1978; Sarmiento, 1985, 1988; Rose, 1988, 1993). Differences in this sense are also remarkable between the more terrestrial and the more arboreal nonhuman great apes, with orangutans displaying the largest range of motion (Begun, 2003; Drapeau, 2008). Meanwhile, humans are characterised by having a much more flexible range of motions that provide high dexterity in relation to the habitual manufacture and use of tools (Napier, 1962; Marzke, 1971, 1986; Wolfe *et al*., 2006).

The main anatomical differences in the radius that have been linked to weight-bearing functions and stability maintenance during quadrupedal locomotion are located in the proximal (e.g., bevelled head of the radius) and distal epiphyses (e.g., angled, rectangular and deeply concave radiocarpal articular surface; enlarged, deep and concave scaphoid articulation relative to a small lunate articulation). These features are meant to provide a close-packing mechanism at the wrist, and to avoid displacement at the elbow during full pronation. On the other hand, both humans and orangutans display a wider range of movements that do not result, however, in the same morphologies (e.g., more angled distal epiphysis in orangutans, more marked styloid process in humans), which, in turn, might cause a clearer distinction between these two groups. This is reflected by the results obtained in the present study, through all types of PCA and their derived TPS warpings.

Differences are more remarkable when size is included in the analyses, since orangutans are characterised by having exceptionally long and slender radii in comparison to the straight and short human radius, which provides greater efficiency for carrying loads. Anatomical differences along the diaphysis of the radius are not only perceptible in overall size-related features or curvature degrees, but also in the relative position and orientation of specific osseous landmarks linked to the attachment of soft tissues (e.g., the medially positioned radial tuberosity in orangutans and chimpanzees increases the lever advantage of *biceps brachii* in supination, whereas the posteriorly placed interosseous crest with respect to the radial tuberosity in humans serves radio-ulnar stabilisation and load transfer). The presence of morphologically and biomechanically significant features along the midshaft, highlights the relevance of incorporating detailed quantitative descriptions of the diaphyseal portion of the radius in morphometric studies. This is also noteworthy for the detection of form-and-function relationships in the primate postcranial skeleton. Likewise, the application of alternative DR techniques, as the ones presented here, can be seen important in exploring intra and interspecific variability in skeletal elements, which, in turn, might be an interesting way of revealing morphological patterns that are significant in group characterisation.

Despite the links that have been established between form and function in modern great apes, the translation of specific biomechanical capabilities to the skeleton and vice versa, as well as the evolution of certain anatomical traits, are not fully understood yet. On top of that, the increasing and heterogeneous hominoid fossil record has emphasised the multiple appearance and mosaic nature of certain osseous features associated to suspensory behaviours (e.g., Pilbeam *et al*., 1990; Moyà-Solà and Köhler, 1996; Finarelli and Clyde, 2004; Moyà-Solà *et al*., 2004; Almécija *et al*., 2007, 2009; Begun, 2010). This in turn has resulted in the proposal of different hypotheses which aim at explaining the appearance of specific morphological characteristics throughout hominoid evolution, and more specifically, throughout human evolution (Gregory, 1927, 1928; Morton, 1926; Washburn, 1967; Tuttle, 1967, 1969; Stern, 1975; Fleagle *et al*., 1981; Tuttle, 1981; Gebo, 1992, 1996; Richmond and Strait, 2000; Thorpe *et al*., 2007; Crompton *et al*., 2008; Arias-Martorell *et al*., 2014).

In conclusion, the combination of different techniques that enable researchers to look at already repeatedly studied skeletal structures from more varied perspectives, which in the present case also appear to provide more stable morphometric affinities regardless of the data input, could be very beneficial for future studies. In this sense, not only the anatomical characterisation of living primates could benefit from a wider range of available ordination, visualisation and analytical approaches, but also paleontological studies, which tend to be limited by the fragmentary nature of the fossil record. From this perspective, the GraphGMM framework provides a toolkit that is less sensitive to data input, and thus is able to maintain interspecific morphometric relationships across highly explanatory and easily interpretable graphics.

## Funding

L.A.C. is funded by the Spanish Ministry of Science, Innovation and Universities with an FPI Predoctoral Grant (Ref. PRE2019-089411), associated with the project RTI2018-099850-B-I00 and the University of Salamanca. J.A. is funded by the Euskal Herriko Unibertsitatea through its programme for postdoctoral specialization (Ref. ESPDOC21/05)

## Acknowledgements

The corresponding author would like to express his gratitude to Guillermo Blanco for his insights and advice into earlier versions of this study.

## Author Contributions

**L.A.C.** Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft, Review & Editing.
**J.A.** Data Curation, Formal Analysis, Investigation, Methodology, Resources, Validation, Writing – Original Draft, Review & Editing.
**D.G.A.** Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing.

## Data Availability Statement

The data underlying this article has been referenced where necessary. Access to additional data may be provided upon reasonable request to the authors.

## Conflict of Interest

The authors wish to declare that they have no conflicts of interest, and take full responsibility for the contents of this study.

## References

Adams, D.C. (1999) Methods for shape analysis of landmark data from articulated structures. *Evolutionary Ecological Research*. 1:959-970

Adams, D.C. (2004) Character displacement via aggressive interference in applachian salamanders. *Ecology*. 85:2664-2670

Adams, D.C. (2010) Parallel evolution of character displacement by competitive selection in terrestrial salamanders. *Evolutionary Biology.* 10:1-10

Adams, D.C., Collyer, M., Kaliontzopoulou, A., Baken, E. (2021) Geomorph: software for geometric morphometric analyses. R package version 4.0.2. Available online: https://cran.r-project.org/package=geomorph (Accessed: 22/04/2022).

Aiello, L.C. (1981) The allometry of primate body proportions, in: Day, E.H. (Eds.), *Vertebrate Locomotion. Symposia for the Zoological Society of London* 48. London: Academic Press.

Almécija, S., Alba, D.M., Moyà-Solà, S., Köhler, M. (2007) Orang-like manual adaptations in the fossil hominoid *Hispanopithecus laietanus*: first steps towards great ape suspensory behaviours. *Proceedings of the Royal Society B.* 274:2375-2384.

Almécija, S., Alba, D.M., Moyà-Solà, S. (2009) *Pierolapithecus* and the functional morphology of Miocene ape hand phalanges: paleobiological and evolutionary implications. *Journal of Human Evolution*. 57:284-297.

Arias-Martorell, J., Tallman, M., Potau, J.M., Bello-Hellegouarch, G., Pérez-Pérez, A. (2014) Shape Analysis of the Proximal Humerus in Orthograde and Semi-Orthograde Primates: Correlates of Suspensory Behavior. *American Journal of Primatology*. 77:7159162

Avants, B.B., Schoenemann, P.T., Gee, J.C. (2006) Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex, *Medical Image Analysis*. 10:397-412

Begun, D.R. (2003) Knuckle-walking and the original of human bipedalism. In: Meldrum, D.J., Hilton, C.E. (Eds.) *From Biped to Strider. The Emergences of Modern Human Walking, Running, and Resource Transport*. Boston: Springer. p. 9-33.

Begun, D.R. (2010) Miocene hominids and the origins of the African apes and humans. *Annual Review of Anthropology*. 39:67-84.

Bernardini, F., Mittleman, J., Rushmeier, H., Silva, C., Taubin, G. (1999) The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics.* 5(4):349-359.

Beaudet, A., Dumoncel, J., Beer, F., Durrleman, S., Gilissen, E., Oettlé, A., et al. (2017) The endocranial shape of *Australopithecus africanus*: surface analysis of the endocasts of Sts 5 and Sts 60. *Journal of Anatomy.* 232(2):296-303

Bishop, C. (2006) *Pattern Recognition and Machine Learning.* Singapore: Springer

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment.* 10:1-12. https://arxiv.org/abs/0803.0476

Bookstein, F.L. (1978) *The measurement of Biological Shape and Shape Change*, Lecture notes on Biomathematics: 24. New York: Springer-Verlag.

Bookstein, F.L. (1989) Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 11:567-585

Bookstein, F.L. (1990) Introduction to methods for landmark data. In: Bookstein, F.L., Rohlf, F.J. (Eds.) *Proceedings of the Michigan Morphometrics Workshop*. Ann Arbor: University of Michigan Museum of Zoology. p. 215-225

Bookstein, F.L. (1991) *Morphometric tools for landmark data: geometry and biology.* Cambridge: Cambridge University Press.

Bookstein, F.L. (1997) Landmark methods for forms without landmarks: morphometrics of group differences in outline shape, *Medical Image Analysis.* 1:225-243

Braga, J., Zimmer, V., Dumoncel, J., Samir, C., Beer, F., Zanolli, C., et al. (2019) Efficacy of diffeomorphic surface matching and 3D geometric morphometrics for taxonomic discrimination of Early Pleistocene hominin mandibular molars, *Journal of Human Evolution.* 130:21-35

Bruna, J., Zaremba, W., Szlam, A., LeCun, Y. (2014) Spectral networks and deep locally connected networks on graphs. *International Conference on Learning Representations.* 1-10, https://arxiv.org/abs/1312.6203

Bronstein, M.M., Kokkinos, I. (2010) Scale-invariant heat kernel signatures for non-rigid shape recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1704-1711, https://doi.org/10.1109/CVPR.2010.5539838

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vanderghenynst, P. (2017) Geometric Deep Learning: Going Beyond Euclidean Data. *IEEE Signal Processing Magazine*. 34(4):18-42. https://doi.org/10.1109/MSP.2017.2693418

Cant, J.G.H. (1987) Positional behavior of female Bornean orang-utans (*Pongo pygmaeus*). *American Journal of Primatology*. 12:71–90.

Carlsoo, S., Johansson, O. (1962) Stabilization of and load on the elbow joint in some protective movements. *Acta Anatomica* 48:224-231.

Chapman, R.E. (1990) Conventional Procrustes Approaches. In F.J. Rohlf and F.L. Bookstein (Eds.) *Proceedings of the Michigan Morphometrics Workshop.* Michigan: University of Michigan Museum of Zoology. 251-268

Corruccini, R.S. (1978) Comparative Osteometrics of the Hominoid Wrist Joint, with Special Reference to Knuckle-walking. *Journal of Human Evolution*. 7:307-321.

Cortes, C.; Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*. 20:273-297

Crompton, R.H., Vereecke, E.E., Thorpe, S.K.S. (2008) Locomotion and posture from the common hominoid ancestor to fully modern hominins, with special reference to the last common panin/hominin ancestor. *Journal of Anatomy*. 212:501-543.

Diaconsis, P.; Freedman, D. (1984) Asymptotics of Graphical Projection of Pursuit. Annals Stat. 12:793-815

Dodson, P. (1978) On the use of ratios in growth studies. *Systematic Zoology*. 27(1):62-67

Doran, D.M. (1992) The ontogeny of chimpanzee and pygmy chimpanzee locomotor behaviour: a case study of paedomorphism and its behavioural correlates. *Journal of Human Evolution*. 23:139-157.

Doran, D.M. (1993) The comparative locomotor behaviour of chimpanzees and bonobos: the influence of morphology on locomotion. *American Journal of Physical Anthropology*. 91:83-98.

Doran, D.M. (1997) Ontogeny of locomotion in mountain gorillas and chimpanzees. *Journal of Human Evolution* 32:323-344.

Doran, D.M., McNeilage, A. (1998) Subspecific variation in gorilla behavior: the influence of ecological and social factors. In: Robbins, M.M., Sicotte, P., Stewart, K.J. (Eds.) *Mountain Gorillas. Three Decades of Research at Karisoke*. Cambridge: Cambridge University Press. p. 123-149.

Drapeau, M.S.M. (2008) Articular morphology of the proximal ulna in extant and fossil hominoids and hominins. *Journal of Human Evolution*. 55:86-102.

Dryden, I., Mardia, K. (1998) *Statistical Shape Analysis*. New York: John Wiley and Sons.

Dryden, I. (2021) Shapes package. R package version 1.2.6. Available online: https://cran.r-project.org/package=shapes (Accessed: 22/04/2022).

Ferson, S.; Rohlf, F.J.; Koehn, R.K. (1985) Measuring shape variation of two-dimensional outlines, *Systematic Biology*. 34:59-68

Finarelli, J.A., Clyde, W.C. (2004) Reassessing hominoid phylogeny: evaluating congruence in the morphological and temporal data. *Paleobiology*. 30:614–651.

Fleagle, J. G. (1998) *Primate Adaptation and Evolution* (2nd Edition). San Diego: Academic Press.

Fleagle, J.G., Stern, J.T., Jungers, W.L., Susman, R.L., Vangor, A.K., Wells, J.P. (1981) Climbing: a biomechanical link with brachiation and with bipedalism. *Symposium of the Zoological Society of London*. 48:359–375.

Fortunato, S. (2010) Community detection in graphs. *Physics Reports*. 486(3-5):75-174. https://doi.org/10.1016/j.physrep.2009.11.002

Gebo, D.L. (1992) Plantigrady and foot adaptation in African apes: implications for hominid origins. *American Journal of Physical Anthropology*. 89:29–58.

Gebo, D.L. (1996) Climbing, brachiation, and terrestrial quadrupedalism: historical precursors of hominid bipedalism. *American Journal of Physical Anthropology*. 101:55-92.

Gower, J.C., (1975) Generalized Procrustes Analysis, *Psychometrika*. 40:33-50

Gregory, W. (1927) The origin of man from the anthropoid stem- when and where? *Proceedings of the American Philosophical Society*. 66:439-463.

Gregory, W. (1928) Were the ancestors of man primitive brachiators? *Proceedings of the American Philosophical Society*. 67:129-150.

Gunn, S. (1998) Support vector machines for classification and regression. *ISIS Technical Report*. Available from: https://www.svms.org/tutorials/Gunn1998.pdf (Accessed 17/02/2022)

Gunz, P., Mitteroecker, P. (2013) Semilandmarks: a method for quantifying curves and surfaces. *Hystrix, the Italian Journal of Mammalogy*. 24:103-109.

Hallgrimsson, B., Percival, C.J., Green, R., Young, N.M., Mio, W., Marcucio, R. (2015) Morphometrics, 3D Imaging, and Craniofacial Development, *Current Topics in Developmental Biology*. 115:562-597. https://dx.doi.org/10.1016/bs.ctdb.2015.09.003

Hamilton, W.L., Ying, R., Leskovec, J. (2018) Representation Learning on Graphs: Methods and Applications. *IEEE Data Engineering Bulletin*. 40(3):52-74, https://arxiv.org/abs/1709.05584

Henaff, M., Bruna, J., LeCun, Y. (2015) Deep Convolutional Networks on Graph-Structured Data, 1-10, *arXiv,* Available from: https://arxiv.org/abs/1506.05163 (Accessed 17/02/2022)

Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H.; Basu, S., Akoglu, L., et al. (2012) RolX: Structural Role Extraction and Mining in Large Graphs. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 18:1231-1239, https://doi.org/10.1145/2339530.2339723

Hoff, P.D., Raftery, A.E., Handcock, M.S. (2012) Latent space approaches to Social Network Analysis. *Journal of the American Statistical Association*. 97(460):1090-1098

Humphries, J.M., Bookstein, F.L., Chernoff, B., Smith, G.R., Elder, R.L., Poss, S.G. (1981) Multivariate discrimination by shape in relation to size. *Systematic Zoology*. 30(3):291-308

Hunt, K.D. (1991) Mechanical Implications of Chimpanzee Positional Behavior. *American Journal of Physical Anthropology*. 86:521-536.

Hunt, K.D. (2016) Why are there apes? Evidence for the co-evolution of ape and monkey ecomorphology. *Journal of Anatomy*. 228:630-685.

Inouye, S. (1989) Variability of knuckle-walking behaviour in the African apes. *American Journal of Physical Anthropology*. 75:245.

Jenkins, F.A.J. (1973) The functional anatomy and evolution of the mammalian humeroulnar articulation. *American Journal of Anatomy*. 13:281-298.

Jenkins, F.A.J., Fleagle, J.G. (1975) Knuckle-walking and the Functional Anatomy of the Wrists in Living Apes. In: Tuttle, R. (Ed.) *Primate Functional Morphology and Evolution*. The Hague: Mouton. p. 213-227.

Jollife, I. (2002) *Principal Component Analysis*. New York: Springer.

Jungers, W.L., Falsetti, A.B., Wall, C.E. (1995) Shape, Relative Size, and Size-Adjustments in Morphometrics. *Yearbook of Physical Anthropology*. 38:137-161

Kent, J.T. (1994) The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society B*. 56:285-299

Kipf, T., Welling, M. (2017) Semi-supervised classification with graph convolutional networks. *International Conference of Learning Representations*. 1-14, https://arxiv.org/abs/1609.02907

Klingenberg, C.P., Monteiro, L.R. (2005) Distances and directions in multidimensional shape spaces: Implications for morphometric applications. *Society of Statistical Biology*. 54:678-688

Klingenberg, C.P. (2009) Morphometric integration and modularity in configurations of landmarks: tools for evaluating a priori hypotheses. *Evolutionary Development*. 11(4):405-421. https://doi.org/10.1111/j.1525-142X.2009.00347.x

Klingenberg, C.P. (2021) How exactly did the nose get that long? A critical rethinking of the Pinocchio effect and how shape changes relate to landmarks. *Evolutionary Biology*. 48:115-127. DOI: 10.1007/s11692-020-09520-y

Larson, S.G. (1998) Parallel evolution in the hominoid trunk and forelimb. *Evolutionary Anthropology*. 6:87–99.

Lele, S. (1991) Some comments on coordinate-free and scale invariant methods in morphometrics. *American Journal of Physical Anthropology*. 85:407-417

Lele, S., Richtsmeier, J.T. (1992) Euclidean distance matrix analysis: a coordinate-free approach for comparing biological shapes using landmark data. *American Journal of Physical Anthropology*. 86(3):415-427

Leskovec, J. (2019) Graph Neural Networks. *CS224W: Machine Learning with Graphs*. Available from: https://cs224w.stanford.edu [Accessed 17th February, 2022]

Manduell, K.L., Morrogh-Bernard, H.C., Thorpe, S.K.S. (2011) Locomotor behavior of wild orangutans (*Pongo pygmaeus wurmbii*) in Disturbed Peat Swamp Forest, Sabangau, Central Kalimantan, Indonesia. *American Journal of Physical Anthropology*. 145:348–359.

Marzke, M.W. (1971) Origin of the human hand. *American Journal of Physical Anthropology*. 34:61-84.

Marzke, M.W. (1986) Tool use and the evolution of hominid hands and bipedality. In: Else, J.G., Lee, P.C. (Eds.) *Primate evolution*. London: Cambridge University Press. p. 203-209.

Miller, M.I., Younes, L., Trouvé, A. (2014) Diffeomorphometry and geodesic positioning systems for human anatomy, *Technology*. 2(1):36-43.

Mitteroecker, P., Gunz, P. (2009) Advances in Geometric Morphometrics. *Evolutionary Biology*. 36, 235-247.

Morton, D. (1926) Evolution of man's erect posture. *Journal of Morphology*. 43:147-179.

Moyà-Solà, S., Köhler, M. (1996) A Dryopithecus skeleton and the origins of great-ape locomotion. *Nature* 379:156–159.

Moyà-Solà, S., Köhler, M., Alba, D.M., Casanovas-Vilar, I., Galindo, J. (2004) *Pierolapithecus catalaunicus*, a new Middle Miocene great apes from Spain. *Science*. 306:1339–1344.

Murphy, R.L., Srinivasan, B., Rao, V., Ribeiro, B. (2019) Janossy pooling: learning deep permutation-invariant functions for variable-size inputs. *International Conference on Learning Representations*. Available from: https://arxiv.org/abs/1811.01900v3 (Accessed 02/02/2022)

Napier, J.R. (1956) The prehensile movements of the human hand. *Journal of Bone Joint Surgery*. 38B:902–913.

Napier, J.R. (1962) The evolution of the hand. *Scientific American*. 207:56–62.

O'Higgins, P.; Johnson, D.R. (1988) The quantitative description and comparison of biological forms, *Critical Review of Anatomical Sciences*. 1:149-170

O'Higgins, P., Dryden, I.L. (1993) Sexual dimorphism in hominoids: further studies of craniofacial shape differences in *Pan, Gorilla* and *Pongo*. *Journal of Human Evolution*. 24:182-205

Pérez-Criado, L., Rosas, A. (2017) Evolutionary anatomy of the Neandertal ulna and radius in the light of the new El Sidrón sample. *Journal of Human Evolution*. 106:38-53.

Pilbeam, D., Rose, M.D., Badgley, C., Lipschultz, B. (1990) New *Sivapithecus* humeri from Pakistan and the relationship of Sivapithecus and Pongo. *Nature*. 348:237–239.

Pontzer, H., Wrangham, R.W. (2004) Climbing and the daily energycost of locomotion in wild chimpanzees: implications for hominoid locomotor evolution. *Journal of Human Evolution*. 46:317–335.

Profico, A., Bondioli, L., Raia, P., O'Higgins, P., Marchi, D. (2021) morphomap: An R package for long bone landmarking, cortical thickness, and cross-sectional geometry mapping. *American Journal of Physical Anthropology*. 174:129-139.

Rahimi, A., Recht, B. (2007) Random features for large-scale kernel machines. *Proceedings of the International Conference on Neural Information Processing Systems*. 20:1177-1184, DOI: https://doi.org/10.5555/2981562.2981710

Remis, M.J. (1995) Effects of body size and social context on the arboreal activities of lowland gorillas in the Central African Republic. *American Journal of Physical Anthropology*. 97:413-433.

Remis, M.J. (1998) The gorilla paradox: effects of habitat and body size on the positional behavior of lowland and mountain gorillas. In: Strasser, E., Fleagle, J.G., Rosenberger, A., McHenry, H.M. (Eds.) *Primate Locomotion*. Plenum Press. p. 95-106.

Richmond, B.G., Strait, D.S. (2000) Evidence that humans evolved from a knuckle-walking ancestor. *Nature*. 404:382–385.

Richtsmeier, J.T., Cheverud, J.M., Lele, S. (1992) Advances in anthropological morphometrics, *Annual Review of Anthropology*. 21:283-305

Rohlf, F.J., Archie, J.W. (1984) A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae), *Systematic Biology*. 33:302-317

Rohlf, F.J. (1986) Relationships among eigenshape analysis, Fourier analysis, and analysis of coordinates. *Mathematical Geology*. 18:845-854.

Rohlf, J.F.; Slice, D.E. (1990) Extension of the Procrustes method for the optimal superimposition of landmarks. *Systematic Biology*. 39:40-59

Rohlf, F.J., Marcus, L.F. (1993) A revolution in morphometrics, *Trends in Ecology & Evolution*. 8:129-132

Rohlf, F.J. (1996) Morphometric spaces, shape components, and the effects of linear transformations. In: L.F. Marcus, M. Corti, A. Loy, G.J.P. Naylor, D.E. Slice (Eds.) *Advances in Morphometrics*. The Netherlands: Springer. 117-129

Rohlf, F.J. (2000) Statistical Power Comparisons among Alternative Morphometric Methods, *American Journal of Physical Anthropology.* 111:463-478.

Rose, M.D. (1988) Another look at the anthropoid elbow. *Journal of Human Evolution*. 17:193-224

Rose, M.D. (1991) The process of bipedalization in hominids. In: Coppens, Y., Senut, B. (Eds.) *Origine(s) de la Bipédie Chez les Hominidés*. CNRS. p. 37–48.

Rose, M.D. (1993) Functional Anatomy of the Elbow and Forearm in Primates. In: Gebo, D.L. (Ed.) *Postcranial Adaptations of Non-human Primates*. North-Illinois University Press. p. 70-95.

Sarmiento, E.E. (1985) *Functional differences in the skeleton of wild and captive orang-utans and their adaptive significance*. PhD. New York University.

Sarmiento, E.E. (1988) Anatomy of the Hominoid Wrist Joint: Its Evolutionary and Functional Implications. *International Journal of Primatology*. 9:281-345.

Schmitt, D.O., Zeininger, A., Granatosky, M.C. (2016) Patterns, Variability, and Flexibility of Hand Posture during Locomotion in Primates. In: Kivell T., Lemelin P., Richmond B., Schmitt D. (Eds.) *The Evolution of the Primate Hand*. Springer. p. 345-369.

Schölkopf, B., Smola, A., Müller, K.R. (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*. 10(5):1299-1319

Slice, D. (2001) Landmark coordinates aligned by Procrustes Analysis do not lie in Kendall's shape space. *Systematic Biology*. 50(1):141-149. https://doi.org/10.1080/10635150119110

Stern, J.T. (1975) Before Bipedality. *Yearbook of Physical Anthropology.* 19:59-68.

Susman, R.L. (1974) Facultative terrestrial hand postures in an orangutan (*Pongo pygmaeus*) and pongid evolution. *American Journal of Physical Anthropology*. 40:27–37.

Susman, R.L., Badrian, N.L., Badrian, A.J. (1980) Locomotor behavior of *Pan paniscus* in Zaire. *American Journal of Physical Anthropology*. 53:69-80.

Tallman, M. (2012) Morphology of the Distal Radius in extant Hominoids and Fossil Hominins: Implications for the Evolution of Bipedalism. *Anatomical Record*. 295:454-464.

Thomson, N.E. (2021) Correction: The biomechanics of knuckle-walking: 3-D kinematics of the chimpanzee and macaque wrist, hand and fingers. *Journal of Experimental Biology*. 224: jeb242409.

Thorpe, S.K.S., Crompton, R.H. (2005) Locomotor ecology of wild orangutans (*Pongo pygmaeus abelii*) in the Gunung leuser ecosystem, Sumatra, Indonesia: a multivariate analysis using log-linear modelling. *American Journal of Physical Anthropology*. 127:58–78.

Thorpe, S.K.S., Crompton, R.H. (2006) Orangutan positional behavior and the nature of arboreal locomotion in *Hominoidea*. *American Journal of Physical Anthropology*. 131:384-401.

Thorpe, S.K.S., Holder, R.L., Crompton, R.H. (2007) Origin of human bipedalism as an adaptation for locomotion on flexible branches. *Science*. 316:1328–1331.

Thorpe, S.K.S., Holder, R.L., Crompton, R.H. (2009) Orangutans employ unique strategies to control branch flexibility. *Proceedings of the National Academy of Sciences*. 106:12646–12651.

Tuttle, R.H. (1967) Knuckle-walking and the evolution of hominoid hands. *American Journal of Physical Anthropology*. 26:171-206.

Tuttle, R.H. (1969) Quantitative and Functional Studies on the Hands of the *Anthropoidea*. I The *Hominoidea*. *Journal of Morphology*. 128:309-364.

Tuttle, R.H. (1981) Evolution of hominis bipedalism and prehensile capabilities. *Philosophical Transactions of the Royal Society B*. 292:89-94.

Urciuoli, A., Zanolli, C., Almécija, S., Beaudet, A., Dumoncel, J., Morimoto, N., et al. (2021) Reassessment of the phylogenetic relationships of the late Miocene apes *Hispanopithecus* and *Rudapithecus* based on vestibular morphology, *Proceedings of the National Academy of Sciences*. 118(5):e2015215118, https://doi.org/10.1073/pnas.2015215118

Walker, J.A. (2000) Ability of Geometric Morphometric Methods to Estimate a Known Covariance Matrix, *Systematic Biology*. 49(4):686-696

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M. (2019) Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics*. 1(1):1-13

Ward, C.V. (2007) Postcranial and locomotor adaptations of hominoids. In: Henke, W., Tattersall, I. (Eds.) *Handbook of Paleoanthropology Vol. 2: Primate Evolution and Human Origins*. Springer. p. 1363-1386.

Washburn, S.L. (1967) Behaviour and the Origin of Man. The Huxley Memorial Lecture 1967. *Proceedings of the Royal Anthropological Institute*. 1967:21–27.

Wolfe, S.W., Crisco, J.J. Orr, C.M., Marzke, M.W. (2006) The Dart-Throwing Motion of the Wrist: Is It Unique to Humans? *Journal of Hand Surgery*. 31:1429–1437.

Xu, K., Hu, W., Leskovec, J., Jegelka, S. (2019) How powerful are Graph Neural Networks? *International Conference on Learning Representations*. https://arxiv.org/abs/1810.00826v3

Yang, J., Leskovec, J. (2014) Overlapping Communities Explain Core-Periphery Organization in Networks. *Proceedings of the IEEE*. 102(12):1892-1902. https://doi.org/10.1109/JPROC.2014.2364018

Young, N.M. (2003) A reassessment of living hominoid postcranial variability: implications for ape evolution. *Journal of Human Evolution*. 45:441–464.

Zanolli, C., Kaifu, Y., Pan, L., Xing, S., Mijares, A.S., Kullmer, O., et al. (2022) Further analyses of the structural organization of *Homo luzonensis* teeth: evolutionary implications, *Journal of Human Evolution*. 163:103124, https://doi.org/10.1016/j.jhevol.2021.103124

*"...I won't have you cheapen what should be an endless pursuit of perfection."*

John C. McGinley (a.k.a. Dr. Cox)<br>- *Scrubs* -

# Chapter 4

# Conclusions

## 4.1 Conclusions

Geometry, and our subconscious perception of morphology, form a fundamental part of the way we see the world. The shape or form of an element may condition the decisions we make, while our ability to describe our surroundings are often highly dependent on these variables as well. The research detailed in this Doctoral Thesis presents a development of the tools available to describe morphology, proposing a series of precise and high resolution techniques for an empirical characterisation of shape and form. These approaches, and their possible applications, can thus be considered a valuable contribution to science, worthy of further investigation so as to improve their accuracy and transparency.

This study set out to develop new techniques for the detection and modelling of micro- and macroscopic elements, under the premise of reducing subjectivity, improving analytical efficiency and accuracy, while proposing a number of transdisciplinary applications that could benefit multiple fields of science. We have successfully accomplished each of these goals, while taking research a step forward by proposing a new mathematical model that provides a powerful means of extracting and representing morphological data. This study has additionally fulfilled all secondary objectives, with the development of 3 R libraries (*pValueRobust*, *AugmentationMC* and *GraphGMM*), and 2 open-source software (the *TPS Measurement Software* and the *Trampling algorithm*) to ensure the reproducibility of all elements presented within this body of research. Likewise, wherever possible, all data and additional code have been published to ensure transparency.

Considering the original hypotheses proposed at the beginning of this study, we can draw the following conclusions;

**Hypothesis 1. The use of advanced computational resources and robust statistical techniques can greatly improve the accuracy of studies that are typically performed using visual and qualitative criteria.**

As has been seen throughout the course of this Doctoral Thesis, many tasks for the identification or characterisation of elements are traditionally performed visually. From the "U" shaped tooth marks left by carnivores, to the "irregular and asymmetric" skin lesions of a skin-cancer patient, morphology forms an important component of almost any analysis. Here we have shown how tools such as GMMs, or FA, are able to empirically quantify these features, with high statistical differentiation between samples (Courtenay et al., 2020a,b, 2021b, 2022a, 2023, Under Review-a).

Both GMM and FA are very informative methods, presenting not only a means of extracting morphological variables, but also a number of different techniques for the visualisation of morphological changes. Across all studies presented here, both types of analyses have provided a wide array of different means of statistically analysing data, producing very attractive visual representations of data as well. This ensures that results are easy to interpret, and also provides a more empirical way of initially describing a shape or form (Courtenay et al., 2020b, 2021a,b, 2022a, 2023, Under Review-b). This is true of both two-dimensional, and three-dimensional, data, with computer-based analyses proving highly useful for the handling and extraction of data.

We have also shown throughout this study the importance and power of robust statistical approaches (Courtenay et al., 2020a). An in-depth understanding of a dataset is essential prior to any analysis, especially in terms of the data's distribution. As shown throughout this body of research, the correct use of statistical testing is fundamental in order to extract reliable and empirically accurate results (Courtenay et al., 2020a, 2021b). It is notable that for many years analysts have incorrectly used statistical tests, especially in terms of the interpretaion of $p$-Values (see main text and Supplementary Materials of Courtenay et al., 2021a,b). Throughout this body of research, we have proposed a large number of different techniques for the correct and robust evaluation of data, ranging from descriptive statistics, to non-parametric testing both uni- and multivariately, and finally a means of evaluating $p$-Values while proposing $p < 0.003$ as the most empirically reliable $p$-Value threshold (Courtenay et al., 2021b).

The present study has also used robust statistical approaches to fine tune the methodological approaches presented, using these techniques to define as best as possible a means of extracting morphological variables with the least amount of human induced error possible (Courtenay et al., 2020a). Nevertheless, other interesting insights using robust statistics have shown that great care is needed when collecting experimental and referential data, so as to ensure that the conditioning factors being observed are product of the analysis, and not being caused by other confounding or contributing variables (Courtenay et al., 2021a).

**Hypothesis 2. Mathematical tools, that are typically used in biological and anthropological contexts for the quantification of different morphologies, can be useful for the study of non-biological elements as well.**

Both GMMs and FA are techniques traditionally used for the description of biological organisms, however, with the exception of two of the studies included in this body of research (Courtenay et al., Under Review-a, b), the majority of our research successfully applied these techniques to the study of non-anatomical elements, including; tooth marks on bone (Courtenay et al., 2020a, 2021a,b, 2023), cut marks on bone (Courtenay et al., 2020b), as well as skin lesions (Courtenay et al., 2022a).

While not necessarily applied here, many of these techniques are applicable in other contexts as well, including forensic sciences (identification of contemporary BSMs in crime scenes), the study of other types of soft-tissue deformities (e.g. skin ulcers, facial deformities, disabilities), and may even be applicable to material inspection and quality control (e.g. deformation on the surface of different materials, fabrication errors).

An important component of the present Doctoral Thesis is the additional use of applied case studies, beyond the experimental and methodological research included as well. The majority of the presented studies are mostly methodological (Courtenay and González-Aguilera, 2020; Courtenay et al., 2020b,a, 2021a,b, Under Review-a), however, the application of each of these methods to a real-life case study can be considered a strong indicator of the power and effectiveness this line of investigation has (Courtenay et al., 2022a, 2023).

From this perspective, this Doctoral Thesis not only presents a versatile toolset for the description of morphological data, but can also be considered the first case study to identify the activity of an extinct carnivore, based on the marks they leave on bone (Courtenay et al., 2023), as well as the first attempt to quantify skin cancer morphology, while successfully differentiating between NMSC patients in 78.6% of the cases (Courtenay et al., 2022a). From this perspective, the parting hypothesis that "tools typically used in biological and anthropological contexts can be applied to non-biological contexts as well" has been directly proven, not just in experimental settings, but in actual case studies as well.

**Hypothesis 3. Artificially Intelligent Algorithms, especially when fuelled by advanced numeric and categorical simulation techniques, can prove a very powerful tool in the classification of elements based on their morphology.**

AIAs have proven highly successful throughout this Doctoral Thesis, and from multiple perspectives. In general, classification rates of both NNs and SVMs across each study have shown between a 69.0% and 99.9% accuracy, with a median accuracy of $> 90\%$. This is a true testament that AIAs can achieve above human-level performance, especially considering the difficulties in differentiating between cut and trampling marks on bone (Courtenay et al., 2020b), the issues when identifying precise carnivore agents based on their tooth marks (Courtenay et al., 2021b), as well as the low sensitivity rates when diagnosing skin cancer patients (Courtenay et al., 2022a).

We have additionally highlighted the great potential of NNs, SVMs, and a combination of the two in a Neural Support Vector Machine model (NSVM). From this perspective, AIAs can be considered some of the most highly versatile algorithms available, with great potential in multiple applications.

Nevertheless, the true contribution of this investigation in AIA applications, especially in contexts where large datasets are scarce (e.g. archaeology, palaeoanthropology, palaeontology), is the use of unsupervised learning and Monte Carlo based techniques for numeric simulation (Courtenay and González-Aguilera, 2020; Courtenay et al., Under Review-a), alongside the use of robust statistical equivalency testing for the evaluation of synthetic data. Here we have shown the great potential of Machine Teaching (Courtenay and González-Aguilera, 2020; Courtenay et al., 2021b, Under Review-a), whereby AIAs are trained only on synthetic data, and the original data is used for testing, with an increase not only in accuracy, but also in confidence when algorithms are used to make new predictions (Courtenay et al., 2021b). We have additionally been able to prove that bootstrap techniques are problematic and should be avoided in CL applications (Courtenay et al., Under Review-a).

Although not originally hypothesised during the conceptualisation of this Doctoral Thesis, another line of research has become apparent throughout the process of this study. Parallel with the concept of parametric and non-parametric analyses, in Courtenay et al. (Under Review-b) we have been able to present a new mathematical model for the extraction and analysis of morphological data that is not necessarily bound by the linear constraints in typical PCAs. This observation, and proposal of Graph-based GMMs, has consequently increased the resolution of the type of information obtainable from landmark data. Alongside the theoretical reflections provided both in this scientific publication, and those outlined in Appendix A, we believe this to be a valuable contribution to the field of GMMs, and could prove valuable when improving the results of many other applications.

In general, the present Doctoral Thesis presents a strong starting point for future research into transdisciplinary applications of morphological analyses, and hopes to provide a new perspective on the means in which different elements can be studied.

## 4.2 Prospective and Future Research

Overall, this Doctoral Thesis has proven useful for the construction of a methodological framework that can be used in multiple transdisciplinary applications for morphological analyses. Limitations still exist, however, and will be the focus of future research. Augmentation techniques, for example, should never replace real data, as these algorithms cannot invent new variability, only model on the variability that already exists in the sample. Similarly, we wish to stress the importance of correctly using parametric and non-parametric algorithms, whereby analysts should be aware of the limitations that some statistical tests or AIAs have. It is advisable that analysts use the tools suitable for the type of research they are carrying out, and constantly update their knowledge on these methods.

In light of this, we are interested in seeing how the MCMCs from Courtenay et al. (Under Review-a) can be improved, using more complex versions of the algorithm, such as including the implementation of adaptive learning rates, robust step sizes, and other mathematical formulae for the definition of acceptance criterion that may be used to increase the quality of synthetic data.

It would also be interesting to see how the new mathematical model proposed by Courtenay et al. (Under Review-b) affects the study of elements such as tooth pits, cut marks, or other elements that have also been explored in this body of research. Similarly, in Courtenay et al. (Under Review-b) we show how the mathematical and theoretical components of GMMs help simplify the computational complexity of the message passing mechanism. Nevertheless, most graph embedding applications typically include additional parameters that can be learnt in a supervised learning context, typically referred to as Graph Convolutional Networks (GCN). From this perspective, we would be interested in seeing how the use of GCNs could improve classification results in GMMs.

Future research will also focus on how the combination of variables, beyond morphology, may contribute to the accuracy of both statistical results and AIA performance. For example, the shape of a lesion is only one of the variables that are used for skin cancer diagnosis. Other variables, such as colour, texture, and size, are used in combination with border irregularities and symmetry. Courtenay et al. (2022a) present the first quantitative assessment of skin lesion shape, however form could not be calculated due to the lack of a scale bar in each of the photos. In the coming years we will try and rectify this, while assessing how the combination of other types of information (e.g. hyperspectral signatures, *sensu* Courtenay et al., 2021c, 2022b), may improve overall classification rates.

From a different perspective, the present Doctoral Thesis has focused primarily on the processing of morphological variables, however, the available techniques for obtaining 3D models has been mostly omitted. From this point of view, we propose that future research be oriented not only on improving the type of analyses presented here, but also focus on how this information is obtained.

This study has mostly used structured light surface scanning, $\mu$-CT scans, and digital microscopy as a means of obtaining 3D models. Nevertheless, many other techniques exist, and may present more precise means of extracting morphological variables. Current research that has already begun parallel to this Doctoral Thesis, in collaboration with the University of Bordeaux, and the Institut Catalá de Paleoecoloia Humana I Evolució Social (IPHES), is considering robust statistical techniques for the analysis and evaluation of reconstruction error when comparing multiple 3D modelling techniques. In this study, we are considering methods such as Confocal microscopy, Scanning Electron Microscopy, as well as other high-resolution techniques for the study of microscopic elements on bone. We believe it would be interesting to increase this sample, and compare not only the resolution, but also the flexibility and cost of this equipment, so as to propose an optimal workflow for morphological analyses on all levels.

Nevertheless, the primary objective of future research at this moment in time is the diffusion of these techniques in more transdisciplinary applications, so as to assess the true extent to which these methods are applicable. Future research will also consider their limitations, so as to develop more investigation into the creation of a more generalised methodological workflow that can benefit all fields of science.

# Bibliography

Adams, D.C. and Nistri, A. (2010) Ontogenetic convergence and evolution of foot morphology in European cave salamanders (Family: Plethodontidae), *BMC Evolutionary Biology*, 10:216.

Adams, D.C.; Rohlf, F.J.; and Slice, D.E. (2004) Geometric morphometrics: Ten years of progress following the “Revolution”, *Italian Journal of Zoology*, 71(1):5–16.

Albrecht, G.H. (1992) Assessing the affinities of fossils using canonical variates and generalized distances, *Journal of Human Evolution*, 7:49–69.

Allen, J.A. (1877) The influence of physical conditions in the genesis of species, *Radical Review*, 1:108–140.

Andrews, P. and Cook, J. (1985) Natural modifications to bones in a temperate setting, *Man*, 20(4):675–691.

Aramendi, J. (2015) *Facial Ontogenetic Trajectories of Modern Humans and Chimpanzees*, Master’s thesis, The University of Hull and the University of York.

Aramendi, J. (2021) *A new morphological approach to the study of Plio-Pleistocene hominin biomechanics and adaptation*, Ph.D. thesis, Universidad Complutense de Madrid.

Aramendi, J.; Maté-González, M.Á.; Yravedra, J.; Ortega, M.C.; Arriaza, M.C.; González-Aguilera, D.; Baquedano, E.; and Domínguez-Rodrigo, M. (2017) Discerning carnivore agency through the three-dimensional study of tooth pits: Revisiting crocodile feeding behaviour at FLK-Zinj and FLK NN3 (Olduvai Gorge, Tanzania), *Palaeogeography, Palaeoclimatology, Palaeoecology*, 488:93–102, doi:10.1016/j.palaeo.2017.05.021.

Ariza-López, F.J.; Rodríguez-Avi, J.; Gonález-Aguilera, D.; and Rodríguez-Gonzálvez, P. (2019) A new method for positional accuracy control for non-normal errors applied to airborne laser scanning data, *Applied Sciences*, 9:1–18.

Astley, S.J. and Clarren, S.K. (1995) A fetal alcohol syndrome screening tool, *Alcoholism: Clinical and Experimental Research*, 19:1565–1571.

Atchley, W.R. and Anderson, D. (1978) Ratios and the statistical analysis of biological data, *Systematic Zoology*, 27(1):71–78.

Atchley, W.R.; Gaskins, C.; and Anderson, D. (1976) Statistical properties of ratios i. empirical results, *Systematic Zoology*, 25(2):137–148.

Ball, A.D.; Job, P.A.; and Walker, A.L. (2017) SEM-microphotogrammetry, a new take on an old method for generating high-resolution 3d models from SEM images, *Journal of Microscopy*, 267(2):214–226, doi:10.1111/jmi.12560.

Barbero-García, I.; Lerma, J.L.; Marqés-Mateu, Á.; and Miranda, P. (2017) Low-cost smartphone-based photogrammetry for the analysis of cranial deformation in infants, *World Neurosurgery*, 102:545–554.

Barbero-García, I.; Carbelles, M.; Lerma, J.L.; and Marqés-Mateu, Á. (2018) Smarphone-based close-range photogrammetric assessment of spherical objects, *The Photogrammetric Record*, 33(162):283–299, doi: 10.1111/phor.12243.

Barbero-García, I.; Lerma, J.L.; Miranda, P.; and Marqés-Mateu, Á. (2019) Smartphone-based photogrammetric 3d modelling assessment by comparison with radiological medical imaging for cranial deformation analysis, *Measurement*, 131:372–379, doi:10.1016/j.measurement.2018.08.059.

Barbero-García, I.; Lerma, J.L.; and Mora-Navarro, G. (2020) Fully automatic smartphone-based photogrammetric 3d modelling of infant's heads for cranial deformation analysis, *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:268–277, doi:10.1016/j.isprsjprs.2020.06.013.

Barker, M. and Rayens, W. (2003) Partial least squares for discrimination, *Journal of Chemometrics*, 17:166–173.

Bastir, M. and Rosas, A. (2004) Comparative ontogeny in humans and chimpanzees: similarities, differences and paradoxes in postnatal growth and development of the skull, *Annals of Anatomy*, 186:503–509.

Bastir, M.; Rosas, A.; and O'Higgins, P. (2006) Craniofacial levels and the morphological maturation of the human skull, *Journal of Anatomy*, 209:607–654, doi:10.1111/j.1469-7580.2006.00644.x.

Baylac, M.; Villemant, C.; and Simbolotti, G. (2003) Combining geometric morphometrics with pattern recognition for the investigation of species complexes, *Biological Journal of the Linnean Society*, 80:89–98.

Behrensmeyer, A.K.; Gordon, K.D.; and Yanagi, G.T. (1986) Trampling as a cause for bone surface damage and pseudo-cutmarks, *Nature*, 319:768–771.

Bellman, R.E. (1957) *Dynamic programming*, New Jersey: Princeton University Press.

Bellman, R.E. (1961) *Adaptive control processes*, New Jersey: Princeton University Press.

Bello, S. and Galway-Witham, J. (2019) Bone taphonomy inside and out: Application of 3-dimensional microscopy, scanning electron microscopy and micro-computed tomography to the study of humanly modified faunal assemblages, *Quaternary International*, 517:16–32.

Bello, S.M. and Soligo, C. (2008) A new method for the quantitative analyses of cut marks produced by ancient and modern handaxes, *Journal of Archaeological Science*, 36(9):1542–1552.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J.W. (2010) A theory of learning from different domains, *Machine Learning*, 79(1-2):151–175.

Benazzi, S.; Bookstein, F.L.; Strait, D.S.; and Weber, G.W. (2011) A new OH5 reconstruction with an assessment of its uncertainty, *Journal of Human Evolution*, 61(1):75–88, doi:10.1016/j.jhevol.2011.02.005.

Bengio, Y. (2012) Deep learning of representations for unsupervised and transfer learning, *JMLR Workshop and Conference Proceedings*, 27:17–37.

Bengio, Y. and Grandvalet, Y. (2004) No unbiased estimator of the variance of k-fold cross-validation, *Journal of Machine Learning Research*, 5:1089–1105.

Benjamin, D.J. and Berger, J.O. (2019) Three recommendations for improving the use of $p$-values, *The American Statistician*, 73(Sup1):186–191, doi:10.1080/0031305.2018.1543135.

Bergman, C. (1847) Über die verhältnisse der wärmeökonomie der thiere zu ihrer grösse, *Göttinger Studien*, 3(1):595–708.

Binford, L.R. (1981) *Bones: Ancient Men and Modern Myths*, New York: Academic Press Inc.

Bishop, C. (1995) *Neural Networks for Pattern Recognition*, New York: Oxford University Press.

Bishop, C. (2006) *Pattern Recognition and Machine Learning*, Berlin: Springer.

Blackith, R. and Reyment, R.A. (1971) *Multivariate Morphometrics*, New York: Academic Press.

Bland, M. (2015) *An Introduction to Medical Statistics*, Oxford: Oxford University Press.

Blumenschine, R.J. (1995) Percussion marks, tooth marks and the experimental determinations of the timing of hominid and carnivore access to long bones at FLK Zinjanthropus, Olduvai Gorge, Tanzania, *Journal of Human Evolution*, 29:21–51.

Bonaccorso, G. (2019) *Hands-on Unsupervised Learning with Python*, Birmingham: Packt.

Bonhomme, V.; Picq, S.; Gaucherel, C.; and Claude, J. (2014) Momocs: Outline analysis using r, *Journal of Statistical Software*, 56(13):1–24.

Bookstein, F. (1989) Principal warps: thin plate spline and the decomposition of deformations, *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 11(6):567–585, doi:10.1109/34.24792.

Bookstein, F.L. (1978) *The measurement of biological shape and shape change*, Berlin: Springer.

Bookstein, F.L. (1984) A statistical method for biological shape comparisons, *Journal of Theoretical Biology*, 107:475–520.

Bookstein, F.L. (1986a) Size and shape spaces for landmark data in two dimensions, *Statistical Science*, 1(2):181–242.

Bookstein, F.L. (1986b) Size and shape spaces for landmark data in two dimensions (with discussion), *Statistical Science*, 1:181–242.

Bookstein, F.L. (1991) *Morphometric Tools for Landmark Data*, New York: Cambridge University Press.

Bookstein, F.L. (1996) Standard formula for the uniform shape component in landmark data, in: L.F. Marcus; M. Corti; A. Loy; G.P. Naylor; and D.E. Slice (Eds.) *Advances in Morphometrics*, New York: Plenum, 153–168.

Bookstein, F.L. (1997) Landmark methods for forms without landmarks: morphometrics of group differences in outline shape, *Medical Image Analysis*, 1:225–243.

Bookstein, F.L. (2017) A newly noticed formula enforces fundamental limits on geometric morphometric analyses, *Evolutionary Biology*, 44:522–541, doi:10.1007/s11692-017-9424-9.

Bookstein, F.L. (2019) Pathologies of between-groups principal components analysis in geometric morphometrics, *Evolutionary Biology*, 46:271–302, doi:10.1007/s11692-019-09484-8.

Bookstein, F.L.; Chernoff, B.; Elder, R.L.; Humphries, J.M.; Smith, G.R.; and Strauss, R.E. (1985) *Morphometrics in Evolutionary Biology*, Philadelphia: Academy of Natural Sciences Press.

Borsuk, K. (1975) *Theory of Shape*, Warsaw: PWN.

Boschin, F. and Crezzini, J. (2012) Morphometrical analysis on cut marks using a 3d ditigal microscope, *International Journal of Osteoarchaeology*, 22:549–562.

Boulesteix, A.L. (2004) A note on between-group pca, *International Journal of Pure and Applied Mathematics*, 19:359–366.

Brain, C.K. (1967) Bone weathering and the problem of pseudo-tools, *South African Journal of Science*, 63:97–99.

Breiman, L. (1951) Nonparametric discrimination: consistency properties, Technical report, USAF School of Aviation Medicine, University of Berkley, Berkley, California.

Breiman, L. (1996a) Bias, variance, and arcing clasifiers, Technical report, university of California, Statistics Department, Berkley.

Breiman, L. (1996b) Heuristics of instability in model selection, *Annals of Statistics*, 24(6):2350–2383.

Bronstein, A.M.; Bronstein, M.M.; and Kimmel, R. (2005) Three dimensional face recognition, *International Journal of Computer Vision*, 64(1):5–30.

Brownlee, J. (2016) *Deep Learning with Python*, Melbourne: Machine Learning Mastery.

Brownlee, J. (2019) *Better Deep Learning with Python*, Melbourne: Machine Learning Mastery.

Campbell, N.A. and Atchley, W.R. (1981) The geometry of canonical variate analysis, *Systematic Zoology*, 30(3):268–280.

Canny, J. (1986) A computational approach to edge detection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

Cardini, A. and Elton, S. (2007) Sample size and sampling error in geometric morphometric studies of size and shape, *Zoomorphology*, 126:121–134, doi:10.1007/s00435-007-0036-2.

Cardini, A. and Polly, P.D. (2020) Cross-validated Between Group PCA scatterplots: a solution to spurious group separation?, *Evolutionary Biology*, 47:85–95, doi:10.1007/s11692-020-09494-x.

Cardini, A.; Seetah, K.; and Barker, G. (2015) How many specimens do i need? sampling error in geometric morphometrics: testing the sensitivity of means and variances in simple randomized selection experiments, *Zoomorphology*, 134:149–163, doi:10.1007/s00435-015-0253-z.

Cardini, A.; O'Higgins, P.; and Rohlf, F.J. (2019) Seeing distinct groups where there are none: spurious patterns from Between-Group PCA, *Evolutionary Biology*, 46:303–316, doi:10.1007/s11692-019-09487-5.

Carroll, J.D. and Green, P.E. (1997) *Mathematical Tools for Applied Multivariate Analysis*, San Diego: Academic Press.

Caruana, R. (1997) Multitask learning, *Machine Learning*, 28:41–75.

Chambers, J.M.; Freeny, A.; and Heiberger, R.M. (1992) Analysis of variance, designed experiments, in: J.M. Chambers and T.J. Hastie (Eds.) *Statistical Models in S*, Boca Raton: Chapman & Hall/CRC, 145–194.

Chapman, R.E. (1990) Conventional procrustes approaches, in: F.J. Rohlf and F.L. Bookstein (Eds.) *Proceedings of the Michigan Morphometrics Workshop*, Arbor, Michigan: The University of Michigan Museum of Zoology, 251–268.

Chen, B.; Garbatov, Y.; and Soares, C. (2011) Measurement of weld-induced deformations in three-dimensional structures based on photogrammetry technique, *Journal of Ship Product Design*, 27:51–62.

Chollet, F. (2017) *Deep Learning with Python*, Shelter Island: Manning.

Claude, J. (2008) *Morphometrics with R*, New York: Springer.

Cobb, S.N. and O'Higgins, P. (2004) Hominins do not share a common postnatal facial ontogenetic shape trajectory, *Journal of Experimental Zoology*, 302(B):302–321.

Colquhoun, D. (2017) The reproducibility of research and the misinterpretation of p-values, *Royal Society of Open Science*, 4:171085, doi:10.1098/rsos.171085.

Colquhoun, D. (2019) The False Positive Risk: a proposal concerning what to do about *p*-values, *The American Statistician*, 73(Sup1):192–201, doi:10.1080/00031305.2018.1529622.

Coombs, W.T.; Aligna, J.; and Oltman, D.O. (1996) Univariate and multivariate omnibus hypothesis tests to control Type I error rates when population variances are not necessarily equal, *Review of Educational Research*, 66(2):137–179.

Corner, B.D.; Lele, S.; and Richtsmeier, J.T. (1992a) Measuring precision of three-dimensional landmark data, *Journal of Quantitative Anthropology*, 3:347–359.

Corner, B.D.; Lele, S.; and Richtmeier, J.T. (1992b) Measuring precision of three-dimensional landmark data, *Journal of Quantitative Anthropology*, 3:347–359.

Cortes, C. and Vapnik, V. (1995) Support-vector networks, *Machine Learning*, 20:273–297, doi:10.1007/BF00994018.

Costa, A.G. (2010) A geometric morphometric assessment of shape in bone and stone Acheulean bifaces from the Middle Pleistocene site of Castel di Guido, Latlum, Italy, in: S. Lycett and P.R. Chauhan (Eds.) *New perspectives on old stones: analytical approaches to Palaeolithic technologies*, New York: Springer, 23–42.

Courtenay, L.A. (2019) *New methodological advances in the study of taphonomic equifinality in the Lower Pleistocene site of FLK-West (Olduvai Gorge, Tanzania)*, Master's thesis, Universitat Rovira i Virgili.

Courtenay, L.A. and González-Aguilera, D. (2020) Geometric morphometric data augmentation using generative computational learning algorithms, *Applied Sciences*, 10:9133, doi:10.3390/app10249133.

Courtenay, L.A.; Yravedra, J.; Maté-González, M.Á.; Aramendi, J.; and González-Aguilera, D. (2017) 3D analysis of cut marks using a new geometric morphometric methodological approach, *Archaeological and Anthropological Science*, doi:10.1007/s12520-017-0554-x.

Courtenay, L.A.; Yravedra, J.; Huguet, R.; Ollé, A.; Aramendi, J.; Maté-González, M.Á.; and González-Aguilera, D. (2019a) New taphonomic advances in 3d ditigal microscopy: a morphological characterisation of trampling marks, *Quaternary International*, 517:55–66, doi:10.1016/j.quaint.2018.12.019.

Courtenay, L.A.; Yravedra, J.; Huguet, R.; Aramendi, J.; Maté-González, M.Á.; González-Aguilera, D.; and Arriaza, M.C. (2019b) Combining machine learning algorithms and geometric morphometrics: a study of carnivore tooth marks, *Palaeogeography, Palaeoclimatology, Palaeoecology*, 522:28–29, doi: 10.1016/j.palaeo.2019.03.007.

Courtenay, L.A.; Herranz-Rodrigo, D.; Huguet, R.; Maté-González, M.Á.; González-Aguilera, D.; and Yravedra, J. (2020a) Obtaining new resolutions in carnivore tooth pit morphological analyses, *PLoS ONE*, 15(10):e0240328, doi:10.1371/journal.pone.0240328.

Courtenay, L.A.; Huguet, R.; González-Aguilera, D.; and Yravedra, J. (2020b) A hybrid geometric morphometric deep learning approach for cut and trampling mark classification, *Applied Sciences*, 10:150, doi:10.3390/app10010150.

Courtenay, L.A.; Huguet, R.; and Yravedra, J. (2020c) Scratches and grazes: a detailed miscroscopic analysis of trampling phenomena, *Journal of Microscopy*, 277(2):107–117, doi:10.1111/jmi.12873.

Courtenay, L.A.; Herranz-Rodrigo, D. Yravedra, J.; Vázquez-Rodríguez, J.M.; Huguet, R.; Barja, I.; Maté-González, M.Á.; Fernández-Fernández, M.; Muñoz-Nieto, Á.L.; and González-Aguilera, D. (2021a) 3D insigths into the effects of captivity on wolf mastication and their tooth marks; implications in ecological studies of both the past and present, *Animals*, 11:2323, doi:10.3390/anil1082323.

Courtenay, L.A.; Herranz-Rodrigo, D.; González-Aguilera, D.; and Yravedra, J. (2021b) Developments in data science solutions for carnivore tooth pit classification, *Scientific Reports*, 11:10209, doi:10.1038/s41598-021-89518-4.

Courtenay, L.A.; González-Aguilera, D.; Lagüela, S.; del Pozo, S.; Ruiz-Mendez, C.; Barbero-García, I.; Román-Curto, C.; Cañueto, J.; Santos-Durán, C.; Cardeñoso-Álvarez, M.E.; Roncero-Riesco, M.; Hernandez-Lopez, D.; Guerrero-Sevilla, D.; and Rodríguez-Gonzalvez, P. (2021c) Hyperspectral imaging and robust statistics in non-melanoma skin cancer analysis, *Biomedical Optics Express*, 12(8):5107–5127, doi:10.1364/BOE.428143.

Courtenay, L.A.; Barbero-García, I.; González-Aguilera, D.; Rodríguez-Martín, M.; Rodríguez-Gonzalvez, P.; Cañueto, J.; and Román-Curto, C. (2022a) A novel approach for morphological characterisation of non-melanoma skin lesions using elliptic fourier analyses, *Journal of Clinical Medicine*, 11:4392, doi: 10.3390/jcm11154392.

Courtenay, L.A.; González-Aguilera, D.; Lagüela, S.; del Pozo, S.; Ruiz-Mendez, C.; Barbero-García, I.; Román-Curto, C.; Cañueto, J.; Santos-Durán, C.; Cardeñoso-Álvarez, M.E.; Roncero-Riesco, M.; Hernandez-Lopez, D.; Guerrero-Sevilla, D.; and Rodríguez-Gonzalvez, P. (2022b) Deep convolutional neural support vector machines for the classification of basal cell carcinoma hyperspectral signatures, *Journal of Clinical Medicine*, 11:2315, doi:10.3390/jcm11092315.

Courtenay, L.A.; Yravedra, J.; Herranz-Rodrigo, D.; Rodríguez-Alba, J.; Serrano-Ramos, A.; Estaca-Gómez, V.; González-Aguilera, D.; Solano, J.A.; and Jiménez-Arenas (2023) Deciphering carnivoran competition for animal resources at the 1.46 Ma Early Pleistocene site of Barranco León (Orce, Granada, Spain), *Quaternary Science Reviews*, 300:107912, doi:10.1016/j.quascirev.2022.107912.

Courtenay, L.A.; Aramendi, J.; and González-Aguilera, D. (Under Review-a) Recruiting a skeleton crew - methods for simulating and augmenting palaeoanthropological data using Monte Carlo based algorithms, *American Journal of Biological Anthropology*.

Courtenay, L.A.; Aramendi, J.; and González-Aguilera, D. (Under Review-b) A graph based geometric morphometric analysis of primate radii: A new mathematical model for the processing of landmark data, *Journal of Anatomy*.

Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*, London: John Wiley & Sons.

Cox, D.R. and Donnelly, C.A. (2011) *Principles of Applied Statistics*, Cambridge: Cambridge University Press.

Cuthill, J.H.; Guttenberg, S.; Ledger, R.; Crother, R.; and Huertas, B. (2019) Deep learning on butterfly phenotypes tests evolution's oldest mathematical model, *Science advances*, 5:eaaw4967, doi:10.1126/sciadv.aaw4967.

Daver, G.; Berillon, G.; Jacquier, C.; Ardagna, Y.; Yadeta, M.; Maurin, T.; Souron, A.; Blondel, C.; Coppens, Y.; and Boiserrie, J.R. (2018) New hominin postcranial remains from locality OMO 323, Shungura Formation, Lower Omo Valley, southwestern Ethiopia, *Journal of Human Evolution*, 122:23–32, doi:10.1016/j.jhevol.2018.03.011.

del Bove, A.; Profico, A.; Riga, A.; Bucchi, A.; and Lorenzo, C. (2020) A geometric morphometric approach to the study of sexual dimorphoism in the modern human frontal bone, *American Journal of Physical Anthropology*, 173:643–654, doi:10.1002/ajpa.24154.

Demaagd, K.; Oliver, A.; Oostendorp, N.; and Scott, K. (2012) *Practical Computer Vision*, Beijing: O'Reilley.

Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; and Fei-Fei, L. (2009) Imagenet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, doi:10.1109/CVPR.2009.5206848.

Deng, J.; Berg, A.C.; Li, K.; and Fei-Fei, L. (2010) What does classifying more than 10,000 image categories tell us?, in: K. Daniilidis; P. Margos; and N. Paragios (Eds.) *Proceedings of the 11th European Conference on Computer Vision*, Heidelberg: Springer, 71–84.

Détroit, F.; Mijares, A.S.; Corny, J.; Daver, G.; Zanolli, C.; Dizon, E.; Robles, E.; Grün, R.; and Piper, P.J. (2019) A new species of *Homo* from the Late Pleistocene of the Philippines, *Nature*, 568:181–186.

Diaconsis, P. and Freedman, D. (1984) Asymptotics of graphical projection of pursuit, *Annals of Statistics*, 12:793–815.

Diakonikolas, I.; Kamath, G.; Kane, D.M.; Li, J.; Moitra, A.; and Stewart, A. (2017) Being robust (in high dimensions) can be practical, *Proceedings of the International Conference on Machine Learning*, 34:1–10.

Diakonikolas, I.; Kamath, G.; Kane, D.M.; Li, J.; Moitra, A.; and Stewart, A. (2019) Robust estimators in high dimensions with the computational intractability, arXiv: 1604.06443v2.

Diez-Martín, F.; Wynn, T.; Sánchez-Yustos, P.; Duque, J.; Fraile, C.; de Francisco, S.; Urbilarrea, D.; Mabulla, A.; Baquedano, E.; and Domínguez-Rodrigo, M. (2019) A faltering origin for the acheulean? technological and cognitive implications from FLK West (Olduvai Gorge, Tanzania), *Quaternary International*, 526:49–66, doi:10.1016/j.quaint.2019.09.023.

Dobigny, G.; Baylac, M.; and Denys, C. (2002) Geometric morphometrics, neural networks and diagnosis of sibling *Tartellus* species (*Rodentia, Gerbillinae*)., *Biological Journal of the Linnean Society*, 77:319–327.

Dodson, P. (1978) On the use of ratios in growth studies, *Systematic Zoology*, 27(1):62–67.

Domínguez-Rodrigo, M.; Juana, S.; Galán, A.B.; and Rodríguez, M. (2009) A new protocol to differentiate trampling marks from butchery marks, *Journal of Archaeological Science*, 36(12):2643–2654, doi:10.1016/j.jas.2009.07.017.

Domínguez-Rodrigo, M.; Pickering, T.R.; Baquedano, E.; Mabulla, A.; Mark, D.F.; Musiba, C.; Bunn, H.T.; Uribelarrea, D.; Smith, V.; Diez-Martín, F.; Pérez-González, A.; Sánchez-Yustos, P.; Santonja, M.; Barboni, D.; Gidna, A.; Ashley, G.; Yravedra, J.; Heaton, J.L.; and Arriaza, M.C. (2013) First partial skeleton of a 1.34-million-year-old *Paranthropus boisei* from Bed II, Olduvai Gorge, Tanzania, *PLoS ONE*, 8(12):e80347.

Domínguez-Rodrigo, M.; Saldié, P.; Cáceres, I.; Huguet, R.; Yravedra, J.; Rodríguez-Hidalgo, A.; Martín, P.; Pineda, A.; Marín, J.; Gené, C. an Aramendi, J.; and Cobo-Sánchez, L. (2017) Use and abuse of cut mark analyses: The Rorsach effect, *Journal of Archaeological Science*, 86:14–23, doi:10.1016/j.jas.2017.08.001.

Douglas, D. and Peucker, T. (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, *The Canadian Cartographer*, 10(2):112–122, doi:10.3138/FM57-6770-U75U-7727.

Drake, A.G. and Klingenberg, C.P. (2008) The pace of morphological change: historical transformation of skull shape in St Bernard dogs, *Proceedings of the Royal Society: B*, 275:71–76, doi:10.1098/rspb.2007.1169.

Drew, S.J. and Sachs, S.A. (1997) Management of the thalassemia-induced skeletal facial deformity: Case reports and review of the literature, *Journal of Oral Maxillofacial Surgery*, 55:1331–1339, doi:10.1016/s0278-2391(97)90197-x.

Dryden, I.L. and Mardia, K.V. (1998) *Statistical Shape Analysis*, New York: John Wiley & Sons.

Dryden, I.L. and Mardia, K.V. (2016) *Statistical Shape Analysis with Applications in R*, Chichester: Wiley.

Dryden, I.L. and Walker, G. (1999) Highly resistant regression and object matching, *Biometrics*, 55:820–825.

Duchi, J.; Hazan, E.; and Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization, *Journal of Machine Learning Research*, 12:2121–2159.

Edgeworth, M.A. (1885) Methods of statistics, *Journal of the Statistical Society*, 41:181–217.

Elderton, W.P. (1902) Tables for testing the goodness of fit of theory to observation, *Biometrika*, 1:155–163.

Enlow, D.H. and Hans, M.G. (1996) *Essentials of Facial Growth*, Philadelphia: W.B. Saunders.

Fenn, J. and Rasinko, M. (2008) *Mastering the Hype Cycle: How to choose the right innovation at the right time*, Boston: Harvard Business School Publishing.

Ferson, S.; Rohlf, F.; and Koehn, R. (1985) Measuring shape variation of two-dimensional outlines, *Systematic Zoology*, 34(1):59–68.

Fink, M. (2004) Object classification from a single example ustilizing class relevance metrics, *Advances in Neural Information Processing Systems*, 17:449–456.

Finn, C. and Levine, S. (2017) Deep visual foresight for planning robot motion, arXiv: 1610.00696v2.

Finn, C.; Abbeel, P.; and Levine, S. (2017) Model-agnostic meta-learning for fast adaptation of deep networks., In *International Conference on Machine Learning*. 1126-1135. arXiv: 1703.03400v3.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7:179–188.

Foody, G.M. (2008) Harshness in image classification accuracy assessment, *International Journal of Remote Sensing*, 29:3137–3158.

Forbsyth, D.A. and Ponce, J. (2011) *Computer Vision: A Modern Approach*, Boston: Pearson.

Formicola, V. and Franceschi, M. (1996) Regression equations for estimating stature from long bones of early holocene european samples, *American Journal of Physical Anthropology*, 100(1):83–88.

Föstner, W. and Wrobel, B.P. (2016) *Photogrammetric Computer Vision*, Switzerland: Springer.

Franklin, D.; Freedman, L.; Milne, N.; and Oxnard, C. (2006) A geometric morphometric study of sexual dimorphism in the crania of indigenous southern africans., *Southern African Journal of Science*, 102:229–238.

Franklin, D.; O'Higgins, P. amd Oxnard, C.; and Dadour, I. (2007) Sexual dimorphism and population variation in the adult mandible: forensic applications of geometric morphometrics, *Forensic Science Medical Pathology*, 3:15–22.

Freidline, S.E.; Gunz, P.; Janković, I.; Harvati, K.; and Hublin, J.J. (2012) A comprehensive morphometric analysis of the frontal and zygomatic bone of the Zuttiyeh fossil from Israel, *Journal of Human Evolution*, 62:225–241.

Freidline, S.E.; Gunz, P.; and Hublin, J.J. (2015) Ontogenetic and static allometry in the human face: Contrasting khoisan and inuit., *American Journal of Physical Anthropology*, 158:116–131.

Friedman, J. (2002) Stochastic gradient boosting, *Computational Statistics & Data Analysis*, 38(4):367–378.

Friedman, R.J.; Rigel, D.S.; and Kopf, A.W. (1985) Early detection of malignant melanoma: The role of physician examination and self-examination of the skin, *CA: A Cancel Journal for Clinicians*, 35:130–151.

Friedman, R.J.; Gutkowicz-Krusin, D.; Farber, M.J.; Warycha, M.; Schneider-Kels, L.; Papastathis, N.; Mihm, M.C.; Googe, P.; King, R.; Prieto, V.G.; Kopf, A.W.; Polsky, D.; Rabinovitz, H.; Oliviero, M.; Cognetta, A.; Rigel, D.S.; Marghoob, A.; Rivers, J.; Johr, R.; Grant-Kels, J.M.; and Tsao, H. (2008) The diagnostic performance of expert dermoscopists vs a computer-vision system on small-diameter melanomas, *Arch Dermatology*, 144(4):476–482, doi:10.1001/archderm.144.4.476.

Fritzche, D.L. (1961) A systematic method for character recognition, Technical report, The Ohio State University Research Foundation, Department of Electrical Engineering, Columbus, Ohio.

García-Medrano, P.; Ollé, A.; Ashton, N.; and Roberts, M. (2019) The mental template in handaxe manufacture: new insights into Acheulean lithic technological behavior at Boxgrove, Sussex, UK, *Journal of Archaeological Method and Theory*, 26:396–422, doi:10.1007/s10816-018-9376-0.

George, D.; Lehrach, W.; Kansky, K.; Lázaro-Gredilla, M.; Laan, C.; Marthi, B.; Lou, X.; Meng, Z.; Liu, Y.; Wang, H.; Lavin, A.; and Phoenix, D.S. (2017) A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs, *Science*, 358(6368):eaag2612, doi:10.1126/science.aag2612.

Geribàs, N.; Mosquera, M.; and Vergès, J.M. (2010) What novice knappers have to learn to become expert stone toolmakers, *Journal of Archaeological Science*, 37:2857–2870, doi:10.1016/jasc.2010.06.023.

Giardina, C. and Kuhl, F. (1977) Accuracy of curve approximation by harmonically related vectors with elliptical loci, *Computer Graphic Image Processing*, 6(3):277–85.

Gong, Y.; Lin, Z.; Wang, J.; and Gong, N. (2018) Bringing machine intelligence to welding visual inspeciton: development of low-cost portable embedded device for welding quality control, *Electronic Imaging*, 9:1–4, doi:10.2352/ISSN.2470-1173.2018.09.IRIACV-279.

González-Aguilera, D. (2005) *Reconstrucción 3D a partir de una sola vista*, Ph.D. thesis, Universidad de Salamanca.

Goodall, C. (1991) Procrustes methods in the statistical analysis of shape, *Journal of the Royal Statistical Society B*, 53(2):285–339.

Goodfellow, I.; Bengio, Y.; and Courville, A. (2016) *Deep Learning*, Cambridge, Massachusetts: MIT.

Goodfellow, I.J.; Shlens, J.; and Szegedy, C. (2015) Explaining and harnessing adversarial examples, arXiv: 1412.6572v3.

Gower, J.C. (1975) Generalized procrustes analysis, *Psychometrika*, 40:33–51.

Gunz, P. (2005) *Statistical and Geometric Reconstruction of Hominid Crania - Reoncstructing Australopithecine ontogeny*, Ph.D. thesis, university of Wien.

Gunz, P. and Mitteroecker, P. (2013) Semilandmarks: a method for quantifying curves and surfaces, *Hystrix, the Italian Journal of Mammalogy*, 24:103–109, doi:10.4404/hystrix-24.1-6292.

Gunz, P.; Mitteroecker, P.; Bookstein, F.L.; and Weber, W.G. (2005a) Computer aided reconstruction of incomplete human crania using statistical and geometrical estimation methods, *Enter the Past: Computer Applications and Quantitative Methods in Archaeology, Oxford, 2004*, 96-98, BAR International Series 1227, 96-98.

Gunz, P.; Mitteroecker, P.; and Bookstein, F.L. (2005b) Semilandmarks in three dimensions, in: D.E. Slice (Ed.) *Modern Morphometrics in Physical Anthropology*, New York: Springer, 73–98.

Gunz, P.; Mitteroecker, P.; Neubauer, S.; Weber, G.W.; and Bookstein, F.L. (2009) Principles for the virtual reconstruction of Hominin crania, *Journal of Human Evolution*, 57:48–62, doi:10.1016/j.jhevol.2009.04.004.

Hallgrimsson, B.; Percival, C.J.; Green, R.; Young, N.M.; Mio, W.; and Marcucio, R. (2015) Morphometrics, 3d imaging, and craniofacial development, *Current Topics in Developmental Biology*, 115:562–597, doi:10.1016/bs.ctdb.2015.09.003.

Harris, C. and Stephens, M. (1988) A combined corner and edge detector, in: *Alvey Vision Conference*, 147–151.

Hartley, R. and Zisserman, A. (2004) *Multiple View Geometry in Computer Vision*, Cambridge: Cambridge University Press.

Hasan, A.; Pilesjö, P.; and Persson, A. (2011) The use of LIDAR as a data source for digital elevation models - a study of the relationship between the accuracy of digital elevation models and topographical attributes in northern peatlands, *Hydrolic and Earth System Sciences and Discussions*, 8(3):5497–5522, doi:10.5194/hessd-8-5497-2011.

Hawks, J. (2004) How much can cladistics tell us about early hominid relationships?, *American Journal of Physical Anthropology*, 125:207–219.

Haynes, G. (1983) A guide for differentiating mammalian carnivore taxa responsible for gnaw damage to herbivore limb bones, *Paleobiology*, 9(2):164–172.

He, H. and Ma, Y. (2013) *Imbalanced Learning: Foundations, Algorithms and Applications*, Hoboken: John Wiley & Sons.

Held, L. and Ott, M. (2018) On p-Values and Bayes Factors, *Annual Review of Statistics and its Application*, 5(6):1–27.

Henkel, S.; Holländer, D.; Wünsche, M.; Theilig, H.; Hübner, P.; Biermann, H.; and Mehringer, S. (2016) Crack-depth prediction in steel based on cooling rate, *Advances in Materials Science and Engineering*, 0:1016482, doi:10.1155/2016/1016482.

Henning, W. (1966) *Phylogenetic Systematics*, University of Illinois Press: Urbana.

Herbert, J. (2010) Application of a photogrammetry-based system to measure and re-engineer ship hulls and ship parts: an industrial practices-based report, *Computer Aided Design*, 42:731–743.

Herrero-Huerta, M.; Lindenbergh, R.; and Rodríguez-Gonzálvez, P. (2018) Automatic tree parameter extraction by a mobile LiDAR systeam in an urban context, *PLoS ONE*, 13(4):e0196004, doi:10.1371/journal.pone.0196004.

Ho, T.K. (1995) Random decision forests, *Proceedings of the International Conference on Document Analysis and Recognition*, 3:14–16.

Höhle, J. and Höhle, M. (2009) Accuracy measurement of digital elevation models by means of robust statistical methods, *ISPRS Journal of Photogrammetry and Remote Sensing*, 64:398–406.

Hotelling, H. (1951) A generalized T test and measure of multivariate dispersion, in: J. Neyman (Ed.) *Proceedings of the Second Berkley Symposium on Mathematical Statistics and Probability*, Berkley: University of California Press, 23–41.

Hothorn, T.; Hornik, K.; and Zeileis, A. (2006) Unbiased recursive partitioning: a conditional inference framework, *Journal of Computational and Graphical Statistics*, 15(3):651–674, doi:10.1198/10618600 6X133933.

Howells, W.W. (1989) Skull shapes and the map: craniometric analyses in the dispersion of modern *Homo*, *Papers of the Peabody Museum of Archaeology and Ethnology*. Cambridge: Harvard University. No. 79.

Hua, H.; Jin, L.; Xiao, Z.; Tang, Z.; Asundi, A.; and Wanga, Y. (2012) A four-camera video grammetric system for 3-d motion measurement of deformable object, *Optic Lasers Engineering*, 50:800–811.

Humphries, J.M.; Bookstein, F.L.; Chernoff, B.; Smith, G.R.; Elder, R.L.; and Poss, S.G. (1981) Multivariate discrimination by shape in relation to size, *Systematic Zoology*, 30(3):291–308.

Iovita, R.; Tuvi-Arad, I.; Moncel, M.H.; Despriée, J.; Volnchet, P.; and Bahain, J.T. (2017) High handaxe symmetry at the beginning of the European Acheulian: the data from La Noira (France) in context, *PLoS ONE*, 12(5):e0177063, doi:10.1371/journal.pone.0177063.

Jacot, A.; Gabriel, F.; and Hongler, C. (2020) Neural tangent kernel: Convergence and generalization in neural networks, *Neural Information Processing Systems*, 32:1–19.

Jolicoeur, P. (1963) The multivariate generalization of the allometry equation, *Biometrics*, 19(3):497–499.

Jolicoeur, P. and Mosimann, J.E. (1981) Size and shape variation in the painted turtle. a principal component analysis, *Growth*, 24:339–354.

Jollife, I. (2002) *Principal Component Analysis*, New York: Springer.

Jungers, W.L. (1985) *Size and scaling in primate biology*, New York: Plenum.

Jungers, W.L.; Falsetti, A.B.; and Wall, C.E. (1995) Shape, relative size, and size adjustments in morphometrics, *Yearbook of Physical Anthropology*, 38:137–161.

Kaplan, A. and Haenlein, M. (2018) Siri, siri in my hand, who's the fairest in the land? on the interpretations of artificial intelligence, *Buisness Horizon*, 62(1):15–25.

Kendall, D.G. (1977) The diffusion of shape, *Advances in Applied Probability*, 9(3):428–430.

Kendall, D.G. (1983) The shape of poisson-delaunay triangles, in: M.C. Demetrescu and M. Iosifescu (Eds.) *Studies in Probabilities and Related Topics in Honour of Octav Onicescu*, Montreal: Nagard, 321–330.

Kendall, D.G. (1984) Shape, manifolds, procrustean metrics, and complex projective spaces, *Bulletin of the London Mathematical Society*, 16:81–121.

Kendall, D.G. (1985) Exact distributions for shapes of random triangles in convex sets, *Advances in Applied Probability*, 17(2):308–329.

Kendall, D.G. (1989) A survey of the statistical theory of shape, *Statistical Science*, 4(2):87–120.

Kendall, D.G.; Barden, D.; Carne, T.K.; and Le, H. (1999) *Shape and Shape Theory*, Chichester: John Wiley & Sons, LTD.

Kendall, M.G. (1955) *Rank Correlation Methods*, New York: Haffner Publishing Co.

Kennedy-Schaffer, L. (2019) Before $p$<0.05 to beyond $p$<0.05: Using history to contextualize $p$-values and significance testing, *The American Statistician*, 73(Sup1):82–90, doi:10.1080/00031305.2019.1537891.

Kent, J. (1994) The complex bingham distribution and shape analysis, *Journal of the Royal Statistical Society B*, 56:285–299.

Key, A.M. and Lycett, S.J. (2017a) Form and function in the Lower Palaeolithic: history, progress and continued relevance, *Journal of Anthropological Sciences*, 95:1–42.

Key, A.M. and Lycett, S.J. (2017b) Reassessing the production of handaxes versus flakes from a functional perspective, *Archaeological and Anthropological Sciences*, 9:737–753, doi:10.1007/s12520-015-0300-1.

Key, A.M. and Lycett, S.J. (2017c) influence of handaxe size and shape on cutting efficiency: a large-scale experiment and morphometric analysis, *Journal of Archaeological Method and Theory*, 24:514–541, doi: 10.1007/s10816-016-9276-0.

Key, A.M.; Proffitt, T.; Stefani, E.; and Lycett, S.J. (2016) Looking at handaxes from another angle: Assessing the ergonomic and functional importance of edge form in Acheulean bifaces, *Journal of Anthropological Archaeology*, 44:43–55, doi:10.1016/j.jaa.2016.08.002.

Kingdon, J. (2015) *The Kingdon Field Guide to African Mammals*, London: Bloomsbury.

Kingma, D.P. and Ba, J.L. (2015) Adam: A method for stochastic optimization, In *Proceedings of the 3rd International Conference for Learning Representations*. arXiv: 1412.6980.

Kittler, H.; Pehamberger, H.; Wolff, K.; and Binder, M. (2002) Diagnostic accuracy of dermoscopy, *Lancet Oncology*, 3(3):159–165, doi:10.1016/S1470-2045(02)00679-4.

Klingenberg, C.P. (2015) Analyzing fluctuating asymmetry with geometric morphometrics: concepts, methods and applications, *Symmetry*, 7(2):843–934, doi:10.3390/sym7020843.

Klingenberg, C.P. (2021) How exactly did that nose get that long? a critical rethinking of the pinocchio effect and how shape changes relate to landmarks, *Evolutionary Biology*, 48:115–127, doi:10.1007/s116 92-020-09520-y.

Klingenberg, C.P. and Monteiro, L.R. (2005) Distances and directions in multidimensional shape spaces: implications for morphometric applications, *Systematic Biology*, 54(4):678–688.

Kochenderfer, M.J. and Wheeler, T.A. (2019) *Algorithms for Optimization*, Cambridge, Massachusetts: MIT.

Kottner, S.; Ebert, L.C.; Ampanozi, G.; Braun, M.; Thali, M.J.; and Gascho, D. (2017) Virtoscan - a mobile, low-cost photogrammetry setup for fast post-mortem 3d full-body documentations in x-ray computed tomography and autopsy suites, *Forensic Science, Medicine and Pathology*, 13:34–43.

Krizhevsky, A. (2010) Convolutional deep belief networks on cifar-10, Technical report, University of Toronto, Toronto, Canada.

Krizhevsky, A. and Hinton, G. (2009) Learning multiple layers of features from tiny images, Technical report, University of Toronto, Toronto, Canada.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. (2012) Imagenet classification with deep convolutional networks, *Advances in Neural Information Processing Systems*, 25:1–9.

Krogh, A. and Hertz, J.A. (1991) A simple weight decay can improve generalization, *Advances in Neural Information Processing Systems*, 4:950–957.

Kruskal, W.H. and Wallis, W.A. (1952) Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, 47(260):583–621.

Kuhl, F. and Giardina, C. (1982) Elliptic fourier features of a closed contour, *Computer Graphic Image Processing*, 18(3):236–258.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*, Dordrecht: Springer.

Kuow, W.M. and Loog, M. (2019) An introduction to domain adaptation and transfer learning, arXiv: 1812.11806v2.

Lague, M.R.; Chirchir, H.; Green, D.J.; Mbua, E.; Harris, J.K.; Braun, D.R.; Griffin, N.L.; and Richmond, B.G. (2019) Humeral anatomy of the KNM-ER 47000 upper limb skeleton from Ileret, Kenya: Implications for taxonomic identification, *Journal of Human Evolution*, 126:24–38, doi:10.1016/j.jhevol.2018.06.011.

Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data, *Biometrics*, 33:159–174.

Laplace, P.S. (1827) *Traité de Mécanique Céleste, Supplément*, Paris: Duprat.

Lawley, D.N. (1938) A generalization of Fisher's Z-test, *Biometrika*, 30:180–187.

Leakey, L.B. (1959) A new fossil skull from Olduvai, *Nature*, 184(4685):491–493.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. (1998) Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y.; Bengio, Y.; and Hinton, G. (2015) Deep learning, *Nature*, 521:436–444.

Legg, S. and Hutter, M. (2007) A collection of definitions of intelligence, *Frontiers in Artificial Intelligence and Applications*, 157:17–24.

Lele, S. (1991) Some components on coordinate-free and scale-invariant methods in morphometrics, *American Journal of Physical Anthropology*, 85:407–417.

Lepre, C.J.; Roche, H.; Kent, D.V.; Harmand, S.; Quinn, R.L.; Brugal, J.P.; Texier, P.J.; Lenoble, A.; and Feibel, C.S. (2011) An earlier origin for the Acheulian, *Nature*, 477:82–85, doi:10.1038/nature10372.

Lerma, J.L.; Barbero-García, I.; Marqés-Mateu, Á.; and Mirando, P. (2018) Smartphone-based video for 3d modelling: Application to infant's cranial deformation analysis, *Measurement*, 116:299–306, doi: 10.1016/j.measurement.2017.11.019.

Lieberman, D.E.; McBratney, B.M.; and Krovitz, G. (2002) The evolution and development of cranial form in *Homo sapiens*, *PNAS*, 99:1134–1139.

Lombao, D.; Guardiola, M.; and Mosquera, M. (2017) Teaching to make stone tools: new experimental evidence supporting a technological hypothesis for the origins of language, *Scientific Reports*, 7:14394, doi:10.1038/s41598-017-14322-y.

Lord, E.A. and Wilson, C.B. (1984) *The Mathematical Description of Shape and Form*, New York: Wiley.

Lorenz, C.; Ferraudo, A.S.; and Suesdek, L. (2015) Artificial neural network applied as a methodology of mosquito species identification, *Acta Tropica*, 152:165–169.

Lowe, D.G. (1999) Object recognition from local scale-invariant features, *Proceedings of the International Conference on Computer Vision*, 2:1150–1157.

Luhmann, T. (2010) Close range photogrammetry for industrial applications, *ISPRS Journal of Photogrammetry and Remote Sensing*, 65:558–569.

MacKie, R.M. (1986) *An Illustrated Guide to the Recognition of Early Malignant Melanoma*, Glasgow: University of Glasgow.

Mann, H.B. and Whitney, D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, 18(1):50–60.

Marcus, L.F. (1990) Traditional morphometrics, in: F.J. Rohlf and F.L. Bookstein (Eds.) *Proceedings of the Michigan Morphometrics Workshop*, Arbor, Michigan: The University of Michigan Museum of Zoology, 77–122.

Mardia, K.V. (1989) Shape analysis of triangles through directional techniques, *Journal of the Royal Statistical Society B*, 51:449–458.

Mardia, K.V.; Kent, J.T.; and Bibby, J.M. (1997) *Multivariate analysis*, London: Academic Press.

Martin, O. (2018) *Bayesian Analysis with Python*, Birmingham: Packt.

McClelland, J.L. and Rumelhart, D.E. (1986) *Parallel distributed processing*, Cambridge, Massachusetts: MIT.

McMahon, T.A. and Bonner, J.T. (1983) *On size and life*, New York: Scientific American Library.

Mendonca, D.; Naidoo, S.; Skolnick, G.; Skladman, R.; and Woo, A. (2013) Comparative study of cranial anthropometric measurement by traditional calipers to computed tomography and three-dimensional photogrammetry, *Journal of Craniofacial Surgery*, 24(4):1106–1110, doi:10.1097/SCS.0b013e31828dcdcb.

Merlino, G.; Herlyn, M.; Fisher, D.E.; Bastian, B.C.; Flaherty, K.T.; Davies, M.A.; Wargo, J.A.; Curiel-Lewandrowski, C.; Weber, M.J.; Leachman, S.A.; Soengas, M.; McMahon, M.; Harbour, J.W.; Swetter, S.M.; Apline, A.E.; Atkins, M. B. amd Bosenberg, M.W.; Dummer, R.; Gershenwald, A.; Halpern, A.C.; Herlyn, D.; Karakousis, G.C.; Kirkwood, J. M. amd Krauthammer, M.; Lo, R.S.; Long, G.V.; McArthur, G.; Ribas, A.; Shuchter, L.; Sosman, J.A.; Smalley, K.S.; Steeg, P.; Thomas, N.T.; Tsao, H.; Tueting, T.; Weeraratna, A.; Xu, G.; Lomax, R.; Martin, S.; Silverstein, S.; Turnham, T.; and Ronai, Z.A. (2016) The state of melanoma: challenges and opportunities, *Pigment Cell Melanoma Research*, 29(4):404–416, doi:10.1111/pcmr.12475.

Mitchell, T.M. (1997) *Machine Learning*, New York: McGraw and Hill.

Mitteroecker, P. and Bookstein, F. (2011) Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics, *Evolutionary Biology*, 38:100–114, doi:10.1007/s11692-011-9 109-8.

Mitteroecker, P.; Gunz, P.; Bernhard, M.; Schaefer, K.; and Bookstein, F.L. (2004a) Comparison of cranial ontogenetic trajectories among great apes and humans, *Journal of Human Evolution*, 46:679–698.

Mitteroecker, P.; Gunz, P.; Bernhard, M.; Schaefer, K.; and Bookstein, F.L. (2004b) Comparison of cranial ontogenetic trajectories among great apes and humans, *Journal of Human Evolution*, 46:679–698.

Mortenson, P. and Steinbok, P. (2006) Qunatifying positional plagiocephaly: reliability and validity of anthropometric measurements, *Journal of Craniofacial Surgery*, 17(3):413–419.

Muller, A.; Clarkson, C.; and Shipton, C. (2017) Measuring behavioural and cognitive complexity in lithic technology throughout human evolution, *Journal of Anthropological Archaeology*, 48:166–180, doi:10.1 016/j.jaa.2017.07.006.

Mutsvangwa, T.M.; Meintjes, E.M.; Viljoen, D.L.; and Douglas, T.S. (2010) Morphometric analysis and classification of the facial phenotype associated with fetal alcohol syndrom in 5- and 12-year-old children, *American Journal of Medical Genetics*, 152A(1):32–41.

Neyman, J. and Pearson, E.S. (1933a) The testing of statistical hypotheses in relation to probabilities a priori, *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(4):492–510.

Neyman, J. and Pearson, E.S. (1933b) On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society A*, 231:289–337.

Nilsson, N. (1998) *Artificial Intelligence: A New Synthesis*, The Netherlands: Morgan Kauffman.

Nocerino, E.; Menna, F.; Remondino, F.; Toschi, I.; and Rodríguez-Gonzálvez, P. (2017) Investigation of indoor and outdoor performance of two portable mobile mapping systems, *Videometrics, Range Imaging, and Applications XIV*, 10332:103320I, doi:10.1117/12.2270761.

O'Higgins, P. and Dryden, I.L. (1993) Sexual dimorphism in hominoids: further studies of craniofacial shape differences in *Pan*, *Gorilla* and *Pongo*, *Journal of Human Evolution*, 24:182–205.

O'Higgins, P.; Cobb, S.N.; Fitton, L.C.; Gröning, F.; Phillips, R.; Liu, J.; and Fagan, M.J. (2011) Combining geometric morphometrics and functional simulation: an emerging toolkit for virtual functional analyses, *Journal of Anatomy*, 218:3–15, doi:10.1111/j.1469-7580.2010.01301.x.

Olsen, S.L. and Shipman, P. (1988) Surface modification on bone: trampling versus butchery, *Journal of Archaeological Science*, 15(5):535–553.

Otárola-Castillo, E.; Torquato, M.G.; Hawkins, H.C.; James, E.; Harris, J.A.; Marean, C.W.; McPherron, S.P.; and Thompson, J.C. (2017) Differentiating between cutting actions on bone using 3d geometric morphometrics and Bayesian analyses with implications to human evolution, *Journal of Archaeological Science*, 89:56–67, doi:10.1016/j.jas.2017.10.004.

Pante, M.C.; Muttart, M.V.; Keevil, T.L.; Blumenschine, R.J.; Njau, J.K.; and Merritt, S.R. (2017) A new high-resolution 3d quantitative method for identifying bone surface modifications with implications for early stone age archaeological record, *Journal of Human Evolution*, 102:1–11, doi:10.1016/j.jhevol.2016.10.002.

Pearson, K. (1895) Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London*, 58:347–352.

Pearson, K. (1900) X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):151–175.

Peters, R.H. (1983) *The ecological implications of body size*, Cambridge: Cambridge University Press.

Piccolo, D.; Ferrari, A.; Peris, K.; Daidone, R.; Ruggeri, B.; and Chimenti, S. (2002) Dermoscopic diagnosis by a trained clinician vs a clinician with minimal dermoscopy training vs computer-aided diagnosis of 341 pigmented skin lesions: a comparative study, *British Journal of Dermatology*, 147(3):481–486, doi:10.1046/j.1365-2133.2002.04978.x.

Plavacan, J.M. (2012) Sexual size dimorphism, canine dimorphism, and male competition in primates: where do humans fit in?, *Human Nature*, 23:45–67.

Poisson, S.D. (1837) *Recherches sur la probabilité des Jugements en Matière Criminelle et en Matère Civile: Précédées des Règles Générales du Calcul des Probabilités*, Paris: Bachelier.

Pollefeys, M. (2004) Visual modeling with a hand-held camera, *International Journal of Computer Vision*, 59(3):207–232.

Poole, D.; Mackworth, A.; and Goebel, R. (1998) *Computational Intelligence: a Logical Approach*, Reading: Addison-Wesley.

Quenu, M.; Trewick, S.A.; Brescia, F.; and Morgan-Richards, M. (2020) Geometric morphometrics and machine learning challenge currently accepted species limits of the land snail placostylus (pulmonata: Bothriembryontidae) on the list of pines, New Caledonia, *Journal of Molluscan Studies*, 86:35–41, doi:10.1093/mollus/ayz031.

Quinlan, J. (1992) *C4.5: programs for machine learning*, San Mateo: Morgan Kaufmann.

Rahimi, A. and Recht, B. (2007) Random features for large-scale kernel machines, in: *Proceedings of the International Conference of Neural Information Processing Systems*, volume 20, 1–8, doi:10.5555/2981562.2981710.

Ramer, U. (1972) An iterative procedure for the polygonal approximation of plane curves, *Computer Graphics and Image Processing*, 1(3):244–256, doi:10.1016/S0146-664X(72)80017-0.

Rao, C.R. and Suryawanshi, S. (1996) Statistical analysis of shape objects based on landmark data, *PNAS*, 93:12132–12136.

Rao, C.R. and Suryawanshi, S. (1998) Statistical analysis of shape through triangulation of landmarks: a study of sexual dimorphism in hominids, *PNAS*, 95:4121–4125.

Raudseps, J.G. (1965) Some aspects of the tangent-angle vs arc length representation of contours, Technical report, The Ohio State University Research Foundation, Air Force Avionics Laboratory, Columbus, Ohio.

Ravichandiran, S. (2018a) *Hands-on meta learning with Python*, Birmingham: Packt.

Ravichandiran, S. (2018b) *Hands-on Reinforcement learning with Python*, Birmingham: Packt.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. (2016) You only look once: Unified, real-time object detection, arXiv: 1506.02640v5.

Rice, S.H. (1997) The analysis of ontogenetic trajectories: when a change in size or shape is not heterochrony, *Proceedings of the National Academy of Science*, 94:902–912.

Richmond, B.G.; Green, D.J.; Lague, M.R.; Chichir, H.; Behrensmeyer, A.K.; Bobe, R.; Bamford, M.K.; Griffin, N.L.; Gunz, P.; Mbua, E.; Merritt, S.R.; Pobiner, B.; Kiura, P.; Kibunjia, M.; Harris, J.K.; and Braun, D.R. (2020) The upper limb of *Paranthropus boisei* from Ileret, Kenya, *Journal of Human Evolution*, 141:102727, doi:10.1016/j.jhevol.2019.102727.

Richtsmeier, J.T.; Cheverud, J.M.; and Lele, S. (1992) Advances in anthropological morphometrics, *Annual Review of Anthropology*, 21:283–305.

Richtsmeier, J.T.; Deleon, V.B.; and Lele, S.R. (2002) The promise of geometric morphometrics, *American Journal of Physical Anthropology*, 45:63–91.

Robinson, C. and Terhune, C.E. (2017) Error in geometric morphometric data collection: combining data from multiple sources, *American Journal of Physical Anthropology*, 164:62–75, doi:10.1002/ajpa.23257.

Robinson, J. (1960) The affinities of the new Olduvai Australopithecine, *Nature*, 186(4723):456–458.

Rodríguez-Gonzálvez, P. García-Gago, J.; Gomez-Lahoz, J.; and González-Aguilera, D. (2014) Confronting passive and active sensors with non-Gaussian statistics, *Sensors*, 14:13759–13777.

Rodríguez-Gonzálvez, P. and Rodríguez-Martín, M. (2018) Weld bead detection based on 3D geometric features and machine learning approaches, *IEEE Access*, 7:14714–14727, doi:10.1109/ACCESS.2019.2891367.

Rodríguez-Martín, M.; Lagüela, S.; González-Aguilera, D.; and Rodríguez-Gonzálvez, P. (2015) Procedure for quality inspection of welds based on macro-photogrammetric three-dimensional reconstruction, *Optics and Laser Technology*, 73:54–62.

Rodríguez-Martín, M.; Rodríguez-Gonzálvez, P.; Lagüela, S.; and González-Aguilera, D. (2016a) Macrophotogrammetry as a tool for the accurate measurement of three-dimensional misalignment in welding, *Automation in Construction*, 71(2):189–197.

Rodríguez-Martín, M.; Lagüela, S.; González-Aguilera, D.; and Rodríguez-Gonzálvez, P. (2016b) Crack-depth prediction in steel based on cooling rate, *Advances in Materials Science and Engineering*, 0:1016482, doi:10.1155/2016/1016482.

Rodríguez-Martín, M.; Rodríguez-Avi, J.; and de Oña-Crespo, E. Gonález-Aguilera, D. (2019a) Validation of portable mobile mapping system for inspection tasks in thermal and fluid-mechanical facilities, *Remote Sensing*, 11:1–19.

Rodríguez-Martín, M.; Rodríguez-Gonzálvez, P.; Ruiz de Oña, E.; and González-Aguilera, D. (2019b) Validation of portable mobile mapping system for the inspection of tasks in thermal and fluid-mechanical facilities, *Remote Sensing*, 11(19):2205–2219, doi:10.3390/rs11192205.

Rodríguez-Martín, M.; Fueyo, J.G.; González-Aguilera, D.; Madruga, F.J.; García-Martín, R.; Muñoz-Nieto, Á.L.; and Pisonero, J. (2020) Predictive models for the characterization of internal defects in additive materials from thermography sequences supported by machine learning methods, *Sensors*, 20(14):3982, doi:10.3390/s20143982.

Rohlf, F. (1986a) Relationships among eigenshape analysis, fourier analysis, and analysis of coordiantes, *Mathematical Geology*, 18:845–854.

Rohlf, F. and Archie, J. (1984) A comparison of fourier methods for the description of wing shape in mosquitoes (diptera: Culicidae), *Systematic Biology*, 33(3):302–317.

Rohlf, F.J. (1986b) Relationships among eigenshape analysis, Fourier analysis, and analysis of coordinates, *Mathematical Geology*, 18(8):845–854.

Rohlf, F.J. (1990) Fitting curves to outlines, in: F.J. Rohlf and F.L. Bookstein (Eds.) *Proceedings of the Michigan Morphometrics Workshop*, Arbor, Michigan: The University of Michigan Museum of Zoology, 167–177.

Rohlf, F.J. (2000) Statistical power comparisons among alternative morphometric methods, *American Journal of Physical Anthropology*, 111:463–478.

Rohlf, F.J. (2015) The tps series of software, *Hystrix: the Italian Journal of Mammalogy*, 26(1):9–12, doi:10.4404/hystrix-26.1-11264.

Rohlf, F.J. (2021) Why clusters and other patterns can seem to be found in analyses of high-dimensional data, *Evolutionary Biology*, 48:1–16, doi:10.1007/s11692-020-09518-6.

Rohlf, F.J. and Bookstein, F.L. (2003) Computing the uniform component of shape variation, *Systematic Biology*, 52(1):66–69.

Rohlf, F.J.; Loy, A.; and Corti, M. (1996) Morphometric analysis of old world talpidae (mammalia, insectivora) using partial-warp scores, *Systematic Biology*, 45(3):344–362.

Rohlf, J.F. (1993) Shape statistics: Procrustes superimposition and tangent spaces, in: L.F. Marcus; E. Bello; and A. García-Valdecasas (Eds.) *Contributions to Morphometrics*, Madrid: Monografías del Museo Nacional de Ciencias Naturales, volume 8, 131–159.

Rohlf, J.F. (1996) Morphometric spaces, shape components, and the effects of linear transformations, in: L.F. Marcus; M. Corti; A. Loy; G.P. Naylor; and D.E. Slice (Eds.) *Advances in Morphometrics*, New York: Plenum, 117–129.

Rohlf, J.F. (1999) Shape statistics: Procrustes superimposition and tangent spaces, *Journal of Classification*, 16:197–223.

Rohlf, J.F. and Slice, D.E. (1990) Extension of the procrustes method for the optimal superimposition of landmarks, *Systematic Biology*, 39:40–59.

Roussos, P.; Mitsea, A.; Halazonetis, D.; and Sifakakis, I. (2021) Craniofacial shape in patients with beta thalassaemia: a geometric morphometric analysis, *Scientific Reports*, 11:1686, doi:10.1038/s41598-020-80234-z.

Rowland, K. (1966) *Pattern and Shape*, Oxford: Ginn & Co. Ltd.

Ruiz de Oña, E.; Rodríguez-Martín, M.; Rodríguez-Gonzálvez, P.; Mora, R.; and González-Aguilera, D. (2022) WELDMAP: A photogrammetric suite applied to the inspection of welds, *Applied Sciences*, 12(5):2553, doi:10.3390/app12052553.

Rumelhart, D.E.; Hinton, G.E.; and Williams, R.J. (1986) Learning representations by back-propagating errors, *Nature*, 323(6088):533–536.

Russel, S.J. and Norvig, P. (2003) *Artificial Intelligence: A Modern Approach*, New Jersey: Upper Saddle River Press.

Salazar-Gamarra, R.; Seelaus, R.; Lopes da Silva, J.V.; Moreira da Silva, A.; and Dib, L.L. (2016) Monoscopic photogrammetry to obtain 3d models by a mobile device: a method for making facial prostheses, *Journal of Otolaryngology - Head and Neck Surgery*, 45:33, doi:10.11186/s40463-016-0145-3.

Sano, K.; Beyene, Y.; Katoh, S.; Koyabu, D.; Endo, H.; Sasaki, T.; Asfaw, B.; and Suwa, G. (2020) A 1.4-million-year-old bone handaxe from Konso, Ethiopia, shows advanced tool technology in the early Acheulean, *PNAS*, 117(31):18393–18400, doi:10.1073/pnas.2006370117.

Santoro, A.; Barunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. (2016) Meta-learning with memory-augmented neural networks, *Proceedings of the International Conference on Machine Learning*, 33:1–9.

Schmidt-Nielsen, K. (1984) *Scaling: Why is animal size so important?*, Cambridge: Cambridge University Press.

Sellke, T.; Bayarri, M.J.; and Berger, J.O. (2001) Calibration of p values for testing precise null hypotheses, *The American Statistician*, 55(1):62–71.

Serwatka, K. (2015) Bifaces in plain sight: testing elliptical Fourier analysis in identifying reduction effects on Late Middle Palaeolithic bifacial tools, *Litikum*, 3:13–25, doi:10.23898/litikuma0009.

Shannon, C.E. (1948a) A mathematical theory of communication, *The Bell Sytstem Technical Journal*, 27(3):379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.

Shannon, C.E. (1948b) A mathematical theory of communication, *The Bell Sytstem Technical Journal*, 27(4):623–656, doi:10.1002/j.1538-7305.1948.tb00917.x.

Shipman, P. (1988) Actualistic studies of animal resources and hominid activities, in: S. Olsen (Ed.) *Scanning Electrong Microscopy in Archaeology*, Oxford: BAR International Series 452, 261–285.

Siegel, A.F. and Benson, R.H. (1982) A robust comparison of biological shapes, *Biometrics*, 38(2):341–350.

Siegel, A.F. and Pinkerton, J.R. (1982) Robust comparison of three dimensional shapes with an application to protein molecule configurations, Technical report, Department of Statistics, Princeton University, Princeton, New Jersey.

Siegenthaler, M.H. (2015) Methods to diagnose, classify, and monitor infantile deformational plagiocephaly and brachycephaly: a narrative review, *Journal of Chiropractic Medicine*, 14:191–204.

Sing, T.; Sander, O.; Beerenwinkel, N.; and Lengauer, T. (2005) ROCR: Visualizing classifier performance in R, *Bioinformatics*, 21:3940–3941.

Slice, D.E. (1996) Three-dimensional, generalized resistant fitting and the comparison of least-squares and resistant fit residuals, in: L.F. Marcus; M. Cort; A. Loy; G.P. Naylor; and D.E. Slice (Eds.) *Advances in Morphometrics*, New York: Plenum Press, 179–199.

Slice, D.E. (2001) Landmark coordinates aligned by procrustes analysis do not lie in kendall's shape space, *Systematic Biology*, 50(1):141–149.

Slice, D.E. (2005) *Modern Morphometrics in Physical Anthropology*, New York: Plenum.

Soda, K.J.; Slice, D.E.; and Naylor, G.P. (2017) Artificial neural networks and geometric morphometrics methods as a means for classification: A case-sudy using teeth from *Carcharhinus* sp. (*Carcharhinidae*), *Journal of Morphology*, 278:131–141.

Souron, A.; Napias, A.; Lavidalie, T.; Santos, F.; Ledevin, R.; Castel, J.C.; Costamagno, S.; Cusimano, D.; Drumheller, S.; Parkinson, J.; Rosada, L.; and Cochard, D. (2019) A new geometric morphometrics-based shape and size analysis discriminating anthropogenic and non-anthropogenic bone surface modifications on an experimental data set, in: *IMEKO TC4 International Conference on Metrology for Archaeology and Cultural Heritage*, 560–565.

Srivastava, N. (2013) *Improving Neural Networks with Dropout*, Master's thesis, University of Torronto, Canada.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society, Series B (Methodological)*, 36(2):111–147.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):44–47.

Stout, D. and Chaminade, T. (2012) Stone tools, language and the brain in human evolution, *Philosophical Transactions of the Royal Society B*, 367:75–87, doi:10.1098/rstb.2011.0099.

Stout, D.; Toth, N.; Schick, K.; Stout, J.; and Hutchins, G. (2000) Stone tool-making and brain activation: Positron Emission Tomography (PET) studies, *Journal of Archaeological Science*, 27:1215–1223, doi: 10.1006/jasc.2000.0595.

Stout, D.; Toth, N.; and Schick, K. (2006) Comparing the neural foundations of oldowan and acheulean toolmaking: a pilot study using Positron Emission Tomography (PET), in: N. Toth and K. Schick (Eds.) *The Oldowan: Case Studies into the Earliest Stone Age*, Gosport: Stone Age Institute Press, 321–331.

Stout, D.; Toth, N.; Schick, K.; and Chaminade, T. (2008) Neural correlates of Early Stone Age toolmaking: technology, language and cognition in human evolution, *Philosophical Transactions of the Royal Society B*, 363:1939–1949, doi:10.1098/rstb.2008.0001.

Stout, D.; Hecht, E.; Khreisheh, N.; Bradley, B.; and Chaminade, T. (2015) Cognitive demands of Lower Paleolithic tool making, *PLoS ONE*, 10(4):e0121804, doi:10.1371/journal.pone.0121804.

Student (1908) The probable error of a mean, *Biometrika*, 6:1–25.

Su, J.; Vargas, D.V.; and Sakurai, K. (2019) One pixel attack for fooling deep neural networks, arXiv: 1710.08864v7.

Sundberg, P. (1989) Shape and size-constrained principal components analysis, *Systematic Zoology*, 38(2):166–168.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Ehran, D.; Vanhoucke, V.; and Rabinocivh, A. (2014a) Going deeper with convolutions, arXiv: 1409.4842.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. (2014b) Intriguing properties of neural networks, arXiv: 1312.6199v4.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Kiel, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.T.; and Ng, R. (2020) Fourier features let networks learn high frequency functions in low dimensional domains, *Neural Information Processing Systems*, arXiv: 2006.10739v1.

Thompson, D.A. (1917) *On growth and form*, Cambridge: Cambridge University Press.

Tobias, P.V. (1967) *Olduvai Gorge Vol. 2. The Cranium and Maxillary Definition of Australopithecus (Zinjanthropus) boisei*, New York: Cambridge University Press.

Tsao, H.; Olazagasti, J.M.; Cordoro, K.M.; Brewer, J.D.; Taylor, S.C.; Bordeuax, J.S.; Chren, M.M.; Sober, A.J.; Tegeler, C.; Bhushan, R.; and Begolka, W.S. (2015) Early detection of melanoma: reviewing the ABCDEs, *Journal of the American Academy of Dermatology*, 72(4):717–723.

Tukey, J. (1949) Comparing individual means in the analysis of variance, *Biometrics*, 5(2):99–114.

Vargha, A. and Delaney, H.D. (1998) The kruskal-wallis test and stochastic homogeneity, *Journal of Educational Behavioral Statistics*, 23:170–192.

Vestergarrd, M.E.; Macaskill, P.; Holt, P.E.; and Menzies, S.W. (2008) Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting, *British Journal of Dermatology*, 159:669–676, doi:10.1111/j.1365-2133.2008.08713.x.

Walker, J.A. (1996) Principal components of body shape variation with an endemic radiation of threespine stickleback, in: L.F. Marcus; M. Corti; A. Loy; G.P. Naylor; and D.E. Slice (Eds.) *Advances in Morphometrics*, New York: Plenum, 321–334.

Walker, J.A. (2000) Ability of geometric morphometric methods to estimate a known covariance matrix, *Systematic Biology*, 49(4):686–696.

Wasserstein, R.L. and Lazar, N.A. (2016) The ASA statement on $p$-Values: Context, process, and purpose, *The American Statistician*, 70(2):129–133, doi:10.1080/00031305.2016.1154108.

Wasserstein, R.L.; Schirm, A.L.; and Lazar, N.A. (2019) Moving to a world beyond "$p < 0.05$", *The American Statistician*, 73(Sup1):1–19.

Wiering, M.A.; vanr der Ree, M.H.; Embrechts, M.J.; Stollenga, M.F.; Meijster, A.; Nolte, A.; and Schomaker, L.B. (2013) The neural support vector machine, in: *The 25th Benelux Artificial Intelligence Conference*, 257–254.

Wirsansky, E. (2020) *Hands-on Genetic Algorithms with Python*, Birmingham: Packt.

Wu, L.; Clarke, R.; and Song, X. (2010) Geometric morphometric analysis of the early Pleistocene hominin teeth from Jianshi, Hubei province, China, *China Earth SCiences*, 53:1141–1152.

Yezerniac, S.M.; Lougheed, S.C.; and Handford, P. (1992) Measurement error and morphometric studies: statistical power and observer experience, *Systematic Biology*, 41(4):471–482.

Yravedra, J.; García-Vargas, H.; Maté-González, M.Á.; Aramendi, J.; Palomeque-González, J.F.; Vallés-Iriso, J.; Matesanz-Vicente, J.; González-Aguilera, D.; and Domínguez-Rodrigo, M. (2017) The use of micro-photogrammetry and geometric morphometrics for identifying carnivore agency in bone assemblages, *Journal of Archaeological Science: Reports*, 14:106–115, doi:10.1016/j.jasrep.2017.05.043.

Zaczek-Peplinska, J.; Kowalska, M.E.; Malowany, K.; and Malesa, M. (2015) Application of digital image correlation and geodetic displacement measuring methods to monitor water dam behavior under dynamic load, *Journal of Civil Engineering and Architecture*, 9:1496–1505.

Zahn, C. and Roskies, R. (1972) Fourier descriptors for plane closed curves, *IEEE Transactions on Computers*, 21(3):269–281.

Zhang, C. and Ma, Y. (2012) *Ensemble Machine Learning*, The Netherlands: Springer.

Zhang, H. (2004a) The optimality of Naive Bayes, Proc. FLAIRS, 1-6. Available online: `http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf` [Accessed: 07/07/2022].

Zhang, T. (2004b) Solving large scale linear prediction problems using Stochastic Gradient Descent algorithms, In *Proceedings of the 21st International Conference on Machine Learning*. DOI: 10.1145/1015330.1015332.

# Appendix A

# Geometric Morphometrics

## A.1 Procrustes Superimposition Procedures

Sets of landmark configurations can be described as a set of matrices $X_i \in X$, consisting of $n$ individuals, with $p$ landmarks in $k$ dimensions. Two geometrical icons, $X_1$ and $X_2$, can be superimposed and described in terms of *shape* and *form* (see Chapter 2.1). For Geometric Morphometrics to work this superimposition must be congruent by a rigid body transformation (Lord and Wilson, 1984), therefore $X_1 : p \times k$ and $X_2 : p \times k$ must fulfill the following rule: $p_1 \equiv p_2$ and $k_1 \equiv k_2$.

From this perspective, $X_1$ will have the same *form* as $X_2$ if:

$$X_2 = X_1 \Gamma + \vec{1}_p \alpha \tag{A.1}$$

where $\Gamma$ defines the square $(k \times k)$ rotation matrix, and $\alpha$ describes the location parameter containing $k$ values. Additionally, $\vec{1}_p$ describes a $p$-vector of ones.

In extension, $X_1$ and $X_2$ can be considered to have the same *shape* if:

$$X_2 = \beta X_1 \Gamma + \vec{1}_p \alpha \tag{A.2}$$

where $\beta$ describes a scalar parameter. Thus $(\alpha, \Gamma, \beta)$ are used to define the translation, scale and rotation components for the superimposition of $X$.

A number of methods have been proposed for the Procrustes superimposition of two icons, with their eventual extension to include a set of configurations. The most popular methods are typically known to produce two different types of coordinates; Bookstein and Kendall coordinates, named after the authors who initially defined them (Bookstein, 1984; Kendall, 1984; Bookstein, 1986b).

### A.1.1 Bookstein Coordinates

The majority of the work presented by Bookstein (1984, 1986b) was primarily focused on the superimposition of 2D data, however registration can be expanded to 3D data as will be discussed briefly here.

The basis of Bookstein (1984, 1986b)'s approach consists in defining a baseline that can be used as a point of reference for projecting coordinates onto the new coordinate system. The baseline is defined by a pair of landmarks, $p_1$ (or $a$), and $p_2$ (or $b$). Depending on the definition, the objective is to translate coordinates so that points $a$ and $b$ have a fixed set of coordinates, sometimes described as $(0,0)$ and $(1,0)$

(Bookstein, 1991; Rohlf, 1999), or $(-1/2,0)$ and $(1/2,0)$ (Dryden and Mardia, 1998) [1]. These two points are used to calculate the *baseline size* (i.e. the Euclidean distance between $a$ and $b$), which can be used to represent $\beta$ from Equation A.2. Using these two points and the baseline size, Bookstein coordinates define 2D baseline registered configurations $Xb_{:,1}$ and $Xb_{:,2}$ using;

$$Xb_{:,1} = \frac{(b_1 - a_1)(X_{:,1} - a_1) + (b_2 - a_2)(X_{:,2} - a_2)}{\beta} - \frac{1}{2} \tag{A.3}$$

$$Xb_{:,2} = \frac{(b_1 - a_1)(X_{:,2} - a_2) + (b_2 - a_2)(X_{:,1} - a_1)}{\beta} \tag{A.4}$$

$\forall X_i \in X$, configurations can then be considered to be expressed in a $p \times (2-4)k$ space (Claude, 2008).

When extending this concept for coordinates in three dimensions, coordinates $a$ and $b$ are simply developed to be defined as $(-1/2,0,0)$ and $(1/2,0,0)$. Nevertheless, for superimposition along the third dimension, a third basepoint is required, $c$, which is set to be positive on the $x$, $y$ plane, such that $X_c = (Xb_{:,1}, Xb_{:,2}, 0)$.

In three dimensions coordinate registration can be broken down into 3 translation, 1 scaling, and 3 rotation procedures (Bookstein, 1991; Dryden and Mardia, 1998; Claude, 2008). The consequent constructed space is thus defined in $\mathbb{R}^{3\times p-7}$. The first translation ensures that the midpoint between $X_a$ and $X_b$ coincides with $(0,0)$, defining the translated $Xt$ such that;

$$Xt_{:,k} = X_{:,k} - \vec{1}_p \frac{a_k + b_k}{2} \tag{A.5}$$

$Xt$ can then be converted, if required, to a scaled version $Xts$ by dividing by $\beta$.

Finally, rotation is performed calculating a series of rotation matrices defining clockwise rotations around $x$, $y$ and $z$ separately. Thus we define $\theta$ as the rotation along the $z$ axis, $\omega$ as the rotation along the $y$ axis, and $\phi$ as the rotation along the $x$ axis. Rotation matrices are now;

$$\Gamma_z = \begin{pmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{A.6}$$

$$\Gamma_y = \begin{pmatrix} \cos\omega & 0 & -\sin\omega \\ 0 & 1 & 0 \\ \sin\omega & 0 & \cos\omega \end{pmatrix} \tag{A.7}$$

$$\Gamma_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{pmatrix} \tag{A.8}$$

Therefore;

$$X' = \left(\Gamma_x \Gamma_y \Gamma_z Xts^T\right)^T \tag{A.9}$$

are the final $X'$ coordinates projected into the new shape space. Figure A.1 presents an example of the superimposition of two triangles using Bookstein coordinate registration.

[1] The latter will be used in the present document

As a concluding remark, Bookstein coordinates present one distinct disadvantage, being the somewhat subjective definition of basepoints which consequently condition the superimposition procedure. From one perspective, the definition of points to be fixed can be considered problematic in many studies (especially with an elevated number of landmarks), while the quality of superimposition is also dependent on the distance between points; the farther away the reference points, the more distorted the transformation (Rohlf, 1999).

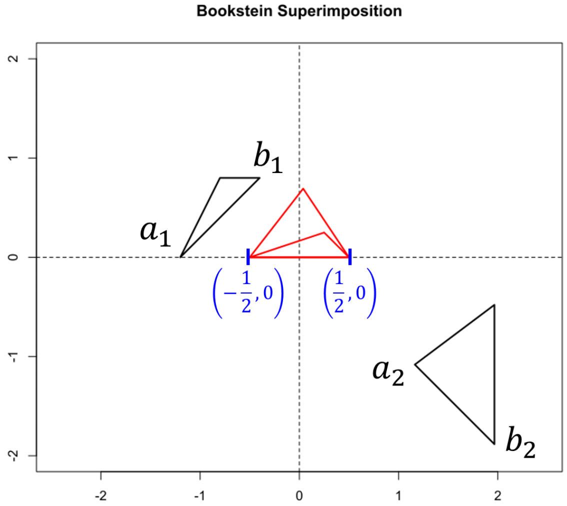


***Figure A.1:*** *Example of two triangles superimposed using Bookstein registration. a and b indicate the basepoints of each triangle used for superimposition.*

### A.1.2 Kendall Coordinates

The methods discussed by Kendall (1984) can be performed using a $(p-1) \times p$ Helmert matrix ($H$). This subset of $H$ (or a sub-Helmert matrix) is premultiplied by the transpose of $X$ to compute a centering matrix. $H$ can be defined as an orthogonal matrix whose first row of elements are equal to $1/\sqrt{p}$. Additionally, row $i$ has $i-1$ elements equal to $-1/\sqrt{i(i-1)}$, followed by one element equal to $(i-1) \times 1/\sqrt{i(i-1)}$, and $p-i$ zeros. The corresponding transformation of $X$ to $X'$ can be considered a configuration lying in *pre-form space*, as only the translation of $X$ has been performed (Rohlf, 1996). The *pre-shape space* ($Z$) and *centred pre-shape* ($Z_C$) space of $X$ can then be obtained by scaling the centred configuration and the Helmertized configuration.

The process of scaling, and essentially defining $\beta$, is performed by calculating the sum of squared deviations of landmarks from the centroid of the configuration (i.e. centroid size). In most traditional GMM literature, the centroid is defined by the mean, however, in this Doctoral Thesis we also consider the possibility of defining the centroid using a more robust metric, such as the median (see Appendix A.5).

An interesting property of the pre-defined spaces is their non-Euclidean nature, defined as the surface of a hypersphere of radius 1 in $(p-1) \times k$ dimensions. In pre-shape space, all possible rotations of a given shape are organised along an orbit called a fiber (Kendall, 1983). A fiber additionally corresponds to a shape in *Kendall's shape space* (the space where translation, rotation and scaling have been removed). To find $\Gamma$ between two shapes, we thus need to find the shortest distance between the two fibers in pre-shape space.

As this space is non-Euclidean in nature, we can define a distance between fibers a number of ways (Gower, 1975; Rohlf and Slice, 1990; Goodall, 1991; Kent, 1994; Rohlf, 1996; Dryden and Mardia, 1998; Slice, 2001). These measurements of distance are known as Procrustes distances. Two main types of Procrustes distance exist; the "chord" ($\Delta$) and the "angular" ($\rho$) distance (eq. A.10 and A.11);

$$\Delta = \sqrt{tr\left(D^T D\right)} \tag{A.10}$$

where $D$ is a $p \times k$ matrix of differences between two configurations. $\Delta$ is related to $\rho$ through;

$$\rho = 2\sin^{-1}\left(\frac{\Delta}{2}\right) \tag{A.11}$$

$\rho$ thus defines the angular distance in radians (See Fig. A.2 in Appendix A.3).

To integrate these notions of distance in non-Euclidean space as a means of estimating $\Gamma$, we can consider solving the following equation;

$$\min \| Z_2 - \beta_1 Z_1 \Gamma \| \tag{A.12}$$

where the objective is to find the smallest distance between two fibers ($Z_i$). We can solve for $\Gamma$ by using the singular-value decomposition $\Gamma = UV^T$, which can be performed on the product of the transpose of $X_1$ and $X_2$ through $X_2^T X_1 = V \Delta U^T$.

The parameter that we wish to optimise to define a best fit can be computed using the following measure of distance;

$$\sqrt{\mathrm{tr}\left(\left(X_2 - \beta X_1 \Gamma\right)^T \left(X_2 - \beta X_1 \Gamma\right)\right)} \tag{A.13}$$

We can thus define Kendall's *form* and *shape* space as the non-Euclidean spaces where configurations have been centred and rotated in the former, as well as scaled in the latter.

## A.2 Generalised Procrustes Analysis

When this procedure is performed on two configurations, as theoretically described above, this procedure is known either as a full or partial Procrustes analysis. The distinction between these two approaches is dependent on how the scaling procedure is performed. From this perspective, Goodall (1991) and (Kent, 1994) define partial scaling as that which uses unit centroid size, while full scaling uses $\cos(\rho)$ (Slice, 2001). When extending this notion to consider the superimposition of more than two configurations, a number of techniques exist. The most common approach, Generalised Procrustes Analyses (GPA), considers minimizing the sum of squared norms of pairwise differences between the shapes;

$$Q = \min\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \| \left(\beta_i X_i \Gamma_i + \vec{1}_p \alpha_i\right) - \left(\beta_j X_j \Gamma_j + \vec{1}_p \alpha_j\right) \|^2\right) \tag{A.14}$$

where $Q$ is the criteria we wish to minimise. GPA is thus performed in three steps; (1) the centred pre-shapes or preforms are calculated, followed by the (2) rotation (and scaling) of the configuration using a measurement of central tendency as a reference point, finally (3) step 2 is iterated until $Q$ can no longer be reduced.

As in the case of calculating a configuration's centroid, the majority of traditional approaches consider using the mean configuration as a measurement of central tendency. Throughout this Doctoral Thesis, we have also considered using the median configuration to adjust for a more robust superimposition (see reflections in Appendix A.5).

Gower (1975) and Rohlf and Slice (1990) offer a means of optimising this approach by adding some additional steps to the algorithm. These authors recommend the rescaling of parameters during every iteration. Additionally, Dryden and Mardia (2016) present an optimised version of the algorithm that is computationally much faster.

## A.3 Non-Euclidean Spaces and a Theoretical Exploration of Shape Space

As previously stated, morphological spaces are not Euclidean in nature, which hinders many statistical tests that can be performed on them. From this perspective, coordinates must first be projected onto a tangent space. Nevertheless, the complexity of morphological space is entirely dependent on the number of landmarks included in our model. We can describe Kendall's shape space for a set of triangles simply as a sphere in $k(p-1)$ dimensions with a radius of 0.5, while the Procrustes hemisphere has a radius of 1.0 (Kendall, 1983; Slice, 2001) (Fig. A.2). It is important to note that, despite $k(p-1) = 4$ for the case of triangles, only three of these dimensions present values greater than 0, and can thus be visualised on the surface of a sphere in 3 dimensions.

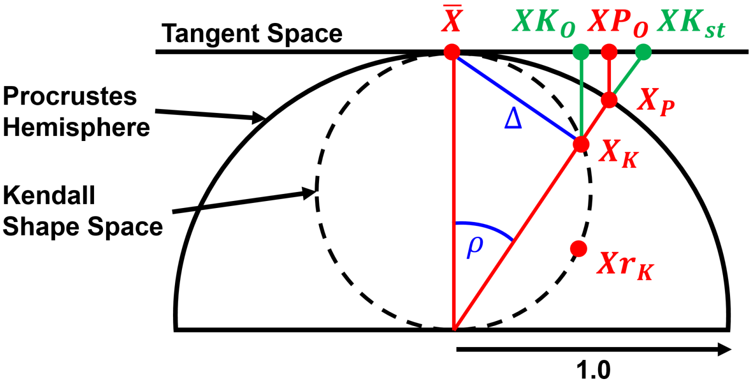


***Figure A.2:*** *A figure describing the geometric relationship between different triangles in different types of shape space. $\bar{X}$ is the central configuration (typically the mean configuration). $X_K$ is a target individual in Kendall's shape space, with its corresponding reflection ($Xr_K$) and the corresponding representation of the same triangle on the Procrustes Hemisphere ($X_P$) (Slice, 2001). $XP_O$ and $XK_O$ describe the Orthogonal projection of Kendall and Procrustes coordinates (respectively) onto tangent space. $XK_{st}$ represents the Stereographic projection. Finally, $\rho$ refers to the full Procrustes distance from the reference shape to the target shape, and $\Delta$ the partial Procrustes distance.*

For the present theoretical explanation, the easiest way to describe and visualise Kendall's shape space is by using triangles. For this purpose, an equilateral triangle was used as a reference shape, adding $\sigma = 0.05$ Gaussian noise to each corner of the triangle to generate 30 random triangles. Theoretically, Gaussian displacements of points will fit a uniform distribution in Kendall's shape space (Kendall, 1985). The coordinates of each triangle on Kendall's sphere can be calculated via;

$$x = \frac{1}{2}\cos\theta \ , \ y = \frac{1}{2}\sin\theta\cos\phi \ , \ z = \frac{1}{2}\sin\theta\sin\phi \tag{A.15}$$

where $\theta$ can be considered analogous with the latitude of the sphere, and $\phi$ the longitude. $(\theta,\phi)$ are 1 for $\bar{X}$, while $x^2+y^2+z^2=\frac{1}{4}$. Similarly, $\phi$ is defined in accordance with Procrustes estimates of the Uniform Shape component (Bookstein, 1989, 1996; Rohlf and Bookstein, 2003). We can derive coordinates through solving the following relationships;

$$\frac{1}{2}\cos\theta = \frac{X_{K_{:,2}}}{1+r^2} \ , \ \frac{1}{2}\sin\theta\cos\phi = \frac{1-r^2}{2\left(1+r^2\right)} \ , \ \frac{1}{2}\sin\theta\sin\phi = \frac{X_{K_{:,1}}}{1+r^2} \tag{A.16}$$

$$X_{K_{:,1}} = \frac{\sin\theta\sin\phi}{1+\sin\theta\cos\phi} \tag{A.17}$$

$$X_{K_{:,2}} = \frac{\cos\theta}{1+\sin\theta\cos\phi} \tag{A.18}$$

Where $r^2 = \left(X_{K_{:,1}}\right)^2 + \left(X_{K_{:,2}}\right)^2$ (Mardia, 1989; Dryden and Mardia, 1998; Kendall et al., 1999).

For the case of our equilateral triangle, we can visualise the corresponding shapes and their projection into shape space in Figure A.3.

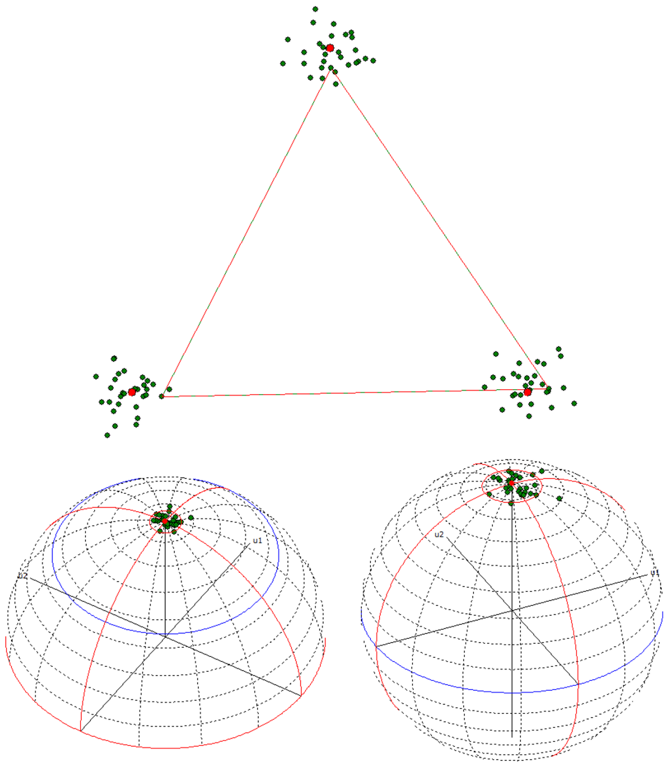


***Figure A.3:*** *A plot of a set of superimposed equilateral triangles, alongside the visualisation of their coordinates on the Procrustes (hyper)hemisphere (bottom left) and Kendall's Hypersphere (bottom right). Figures created using the tpsTri software by Rohlf (2015).*

A study by Slice (2001) describes how, while the geometry of Kendall's shape space is reasonable in theory, practical applications are likely to find landmarks to lie on a shape space not exactly like Kendall's shape space, but in approximation with it (Procrustes Hemisphere, bottom right, Fig. A.2). This is due to how the scaling procedure originally described in Kendall's work performed scaling according to unit centroid size (a Partial Procrustes fit, *sensu* Goodall (1991); Kent (1994)), while most GPA procedures use a Full Procrustes fit for superimposition. Thus, the corresponding surface of a Fully superimposed and a Partially superimposed shape space may be similar on localised regions of the surface, but are not exactly the same. Similarly, large deviations from the pole will imply greater differences between the two shape spaces. This is an important observation, considering how natural biological variation tends to occupy a very small patch on the surface of the manifold (Slice, 2001), and would thus imply a limited difference between the two shape spaces.

Procrustes shape space and Kendall's shape space are directly related for triangles, as the scaling of points on the hemisphere by $\cos(\rho)$ will project points onto the hypersphere (Slice, 2001). For more complex configurations ($p > 3$), however, the relationship between Kendall's hypersphere and the Procrustes (hyper)hemisphere is no longer as simple (*ibid*). This is due to how the equator of the Procrustes (hyper)hemisphere would no longer coincide with that of Kendall's, as it corresponds to a complex projection into a $\mathbb{R}^{p-3}$ space (Kendall, 1984).

As has already been noted, this space is evidently not Euclidean, while the complexity of both Kendall's shape space and the Procrustes hemisphere is much greater when $k > 2$ and $p > 3$. The constructed morphological feature space in these cases can thus be described as a complex manifold, whose surface is also non-Euclidean in nature. In order to project these coordinates into a feature space that is processable via traditional Euclidean based statistical metrics, we project coordinate values onto the tangent space of the manifold. When this projection is performed on Kendall coordinates, the corresponding tangent space is referred to as Kendall's Tangent Space (Rohlf, 1996).

Projections of coordinates can be performed in two ways, either using an orthogonal or stereographic projection. Stereographic projections are computed by dividing the coordinates of the aligned configurations by the cosine of $\rho$ between the shape to be projected and the central configuration reference shape (known typically as the *pole*);

$$XK_{st} = \frac{X_K}{\cos \rho} \tag{A.19}$$

While the orthogonal projection is computed through;

$$XK_O = X_K \left( I_{kp} - \vec{\bar{X}}^T \vec{\bar{X}} \right) \tag{A.20}$$

where $\vec{\bar{X}}$ is a vectorised version of the pole, and $I_{kp}$ is a $kp \times kp$ identity matrix. As would be expected of any projection from a circular to a non-circular cartesian system, this process induces different levels of distorsion that should be taken into consideration. This is especially relevant considering the process of projection is dependent on the pole, or central configuration. Thus shapes farthest away from the pole may be distorted more than those closest.

Stereographic projections are known to accentuate large differences of shapes farthest away from the pole. This is logical considering this type of projection maps points beyond the equator of the hypersphere (Slice, 2001) (Fig. A.2). Orthogonal projections, on the other hand, do not have this problem. Nevertheless, orthogonal projections are known to minimise certain levels of shape distortion and variation. Figure A.3 visualises the difference when projecting into tangent shape space from the Procrustes (hyper)hemisphere for the case of equilateral triangles. Note the present theoretical sample presents low degrees of shape variation ($\sigma = 0.05$), thus points appear to be closely clustered on the surface of the sphere. If simulations are performed with extreme variations from the mean, then these projections would produce large distortions in the reconstructed superimposed shapes in tangent shape space (Fig. A.4).

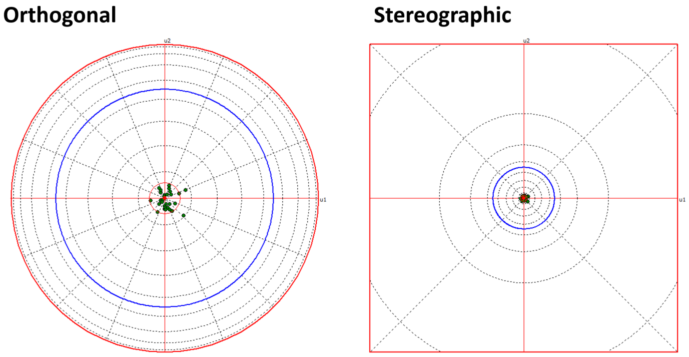


***Figure A.4:*** *Examples of the two types of projection from Kendall's hypersphere to Kendall's tangent shape space. Figures created using the tpsTri software by Rohlf (2015).*

In general, the orthogonal projection is considered to be the most reliable projection of shapes onto their corresponding point in tangent space (Rohlf, 1996, 1999). Likewise, the orthogonal projection from the Procrustes hemisphere has been found to be a much closer approximation to real shape variation, as opposed to coordinates projected from Kendall's shape space (Slice, 2001) (Fig. A.2).

If we theoretically explore and visualise how changes occur across the surface of Procrustes tangent space (Fig. A.5), it can be see how the center of tangent space (i.e. the pole, $\theta = 0$ from the original sphere) represents the equilateral triangle used to generate the theoretical data presented here. As we move to the edge of Procrustes shape space ($\theta = \pi$), which would relatively correspond to the south pole of Kendall's shape space, it can be seen how reflection is represented by this region with the equilateral triangle appearing upside-down. In proximity with the pole (around the northern tropic), shape changes are reduced to slight distortions, with shifts to the right or left resulting in right and left-hand right-angle triangles ($\sin\theta \cos(\phi - (2k\pi/3)) = 1/2$), while shifts downwards generate scalene triangles (in proximity with meridian $\phi = \pi/3$), and shifts upwards describe isosceles triangles ($\phi = \pi$) (Klingenberg, 2015). Shape spaces in proximity with the equator ($\theta = \frac{\pi}{2}$) of the hypersphere represent "splinters", in other words extremely tall or extremely flat triangles (Kendall, 1989; Kendall et al., 1999).

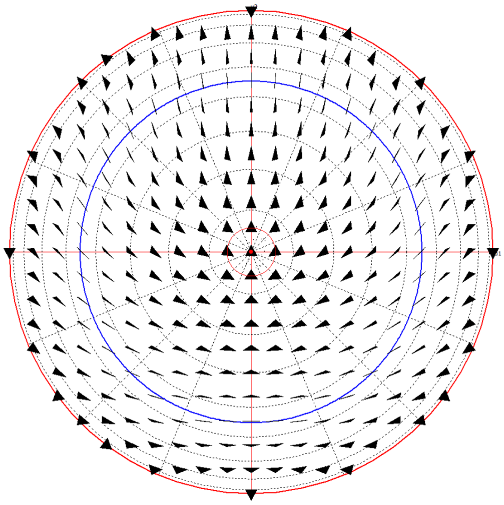


***Figure A.5:*** *Visualisation of shape change across tangent space. The blue line represents the equator from Kendall's hypersphere. Figures created using the tpsTri software by Rohlf (2015).*

If we use this sphere to construct a new set of triangles, we can observe how the two different sets of triangles superimpose (Fig. A.6 & A.7);

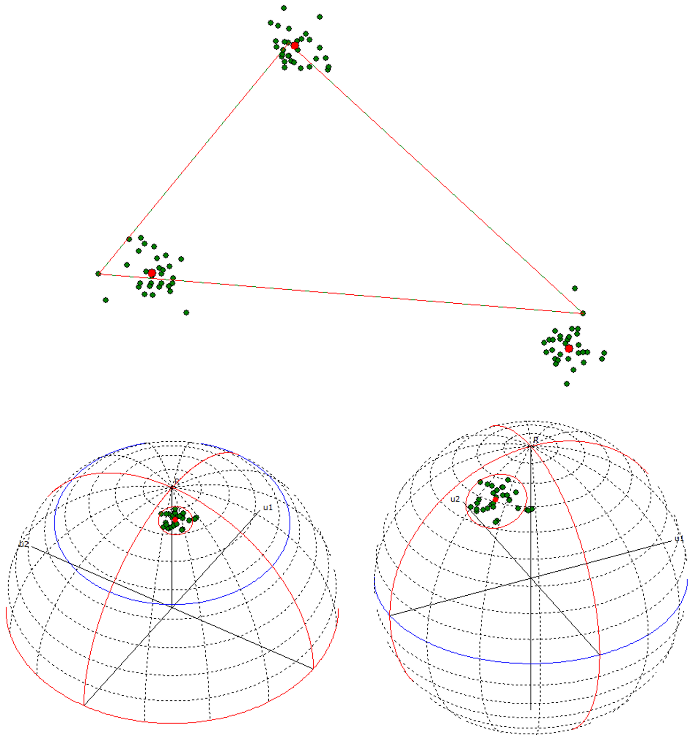


***Figure A.6:*** *A plot of a set of superimposed non-equilateral triangles, alongside the visualisation of their coordinates on the Procrustes (hyper)hemisphere (bottom left) and Kendall's Hypersphere (bottom right). Figures created using the tpsTri software by Rohlf (2015).*

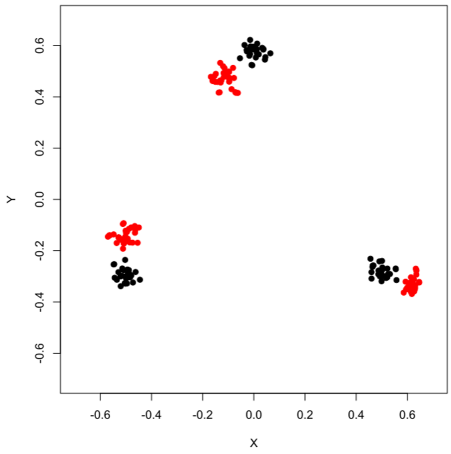


***Figure A.7:*** *Scatter plots of the superimposed coordinates from two sets of triangles in Kendall's tangent space.*

Through calculating a PCA (Fig. A.8), and visualising the variations across PC scores using Thin Plate Splines, we can see how the movement in Kendall's shape space described throughout this chapter results in shape changes we are accustomed to seeing in Geometric Morphometrics.

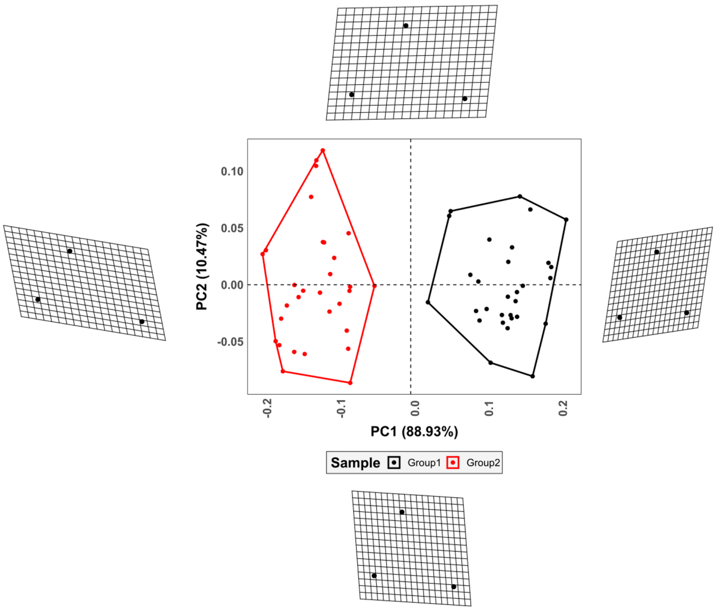


***Figure A.8:*** *PCA derived from the simulated coordinates of triangles expressed throughout this chapter. Thin Plate Spline grid warpings visualise shape changes on the extremity of each PC score.*

## A.4 Thin Plate Splines and Bending Energy

Thin Plate Splines (TPS) are used to mathematically express the deformation of one specimen when mapped onto another (Bookstein, 1989). Building from concepts described by Thompson (1917), where the physical bending energy of deforming a thin metal plate is used as an analogy for shape deformation, Bookstein (1989) defines the formula;

$$U(r) = r^2 \log r^2 \tag{A.21}$$

as the interpolation function for 2D data between two different configurations, where $r$ is the Euclidean distance between two landmark coordinates in the reference configuration. From this function, we define a series of matrices (Bookstein, 1989), $P$, $Q$ and $L$;

$$P = \begin{pmatrix} 0 & U(r_{12}) & \dots & U(r_{1n}) \\ U(r_{21}) & 0 & \dots & (r_{2n}) \\ \dots & \dots & \dots & \dots \\ U(r_{n1}) & U(r_{n2}) & \dots & 0 \end{pmatrix} \forall p \in X \tag{A.22}$$

$$Q = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \dots & \dots & \dots \\ 1 & x_p & y_p \end{pmatrix} \forall p \in X \tag{A.23}$$

$$L = \begin{pmatrix} P & Q \\ Q^T & 0 \end{pmatrix} \tag{A.24}$$

Where $Q$ defines the $x$ and $y$ coordinates of the reference shape. $L_p^{-1}$, being the $p \times p$ upper-left submatrix of $L^{-1}$, defines the *Bending Energy Matrix*. This matrix is already a useful component to calculate coefficients of *non-affine* transformations for all individuals when compared with a reference by simply computing $L_p^{-1} X_i \ \forall \ i \in n$. *Affine* coefficients, on the other hand, are obtained by $L_q^{-1} X_i \ \forall \ i \in n$, where $L_q^{-1}$ is the $p \times 3$ upper-right submatrix of $L^{-1}$.

From $L_p^{-1}$, we can perform eigenvalue decomposition; $L_p^{-1} = E \Lambda E^T$, such that $\Lambda$ defines the diagonal matrix of eigenvalues, and $E$ the eigenvectors. $E_{:,p-k-1}$ is known in Geometric Morphometric literature as the *Principal Warps* (Bookstein, 1989; Rohlf, 1996, 1999), while *Partial Warps* are when these parameters are used to express deformations of shape, rather than in terms of the original coordinates (Bookstein, 1989; Rohlf, 1996). *Partial Warp Scores* are thus obtained by projecting $X$ onto the Principal Warps, resulting in a weighted matrix [2];

$$W = \frac{1}{\sqrt{n}} V \left( I_2 \otimes E \Lambda^{\frac{-\alpha}{2}} \right) \tag{A.25}$$

Where $V$ is the inner product of each row; $V_i = X_i - \bar{X}$. $\alpha$, in this formula, is an important component of analyses as it can be used to fine-tune the type of information represented in subsequent analyses (as will be explained in continuation).

The Singular-Value Decomposition (SVD) of $W$ yields the Relative Warps, $R$, whose definition can vary according to the source (Bookstein, 1991; Rohlf, 1993). In this case, we will define this SVD according to Rohlf (1993), as $W = SDR^T$. Similar to the definition of Equation A.25, we can compute *Relative Warp Loadings* through;

[2] $\otimes$ is the Kronecker tensor product operator, written in the R programming language as **%x%**, and implemented in Python as a function of the Numpy library; **numpy.kron**

$$R' = R\left(I_2 \otimes E\Lambda^{\frac{-\alpha}{2}}\right) \tag{A.26}$$

While the SVD of $W$ can also be visualised in terms of a PCA, typically referred to as a Relative Warp Analysis (RWA) (Bookstein, 1991).

Finally, the Uniform Component of Shape Variation can also be derived from these calculations (among other techniques discussed by Rohlf and Bookstein (2003)), through;

$$N = I_p - E\left(E^T E\right)^{-1} E^T \ , \ : \ E = E_{:,p-k-1} \tag{A.27}$$

A SVD is once again calculated on N in the form of the matrix $LSR^T = V\left(B \otimes I_k\right)$, where $V = X - 1_n\bar{X}$, defining the differences between $X$ and the reference configuration [3]. The product of $LS$ can be used to describe the uniform component of shape differences, by extracting $n$ rows of the first two columns for 2D data, and the first five columns for 3D data. Similarly, $R_{kp,:}$ of the same number of columns provides the uniform component coefficients in the form of linear combinations of $kp$ coordinates (Rohlf and Bookstein, 2003). Nevertheless, the Uniform Component of Shape appears to be a concept of ongoing discussion, with a number of authors proposing different means in which to calculate it (Bookstein, 1989, 1991; Rohlf, 1996; Rohlf and Bookstein, 2003).

The interpretation of this data considers the TPS interpolation function to be a description of a smooth function from one shape to another, such that its deconstruction can be used to define combinations of the affine (uniform) and non-affine (not uniform) deformations between two shapes. Affine transformations include processes such as translation, scaling, and shearing of the overall configuration, without being localised to any particular region of the configuration (subsets of landmarks). Non-affine transformations, on the other hand, refer to the expansion, compression, or bending, of local regions (Rohlf et al., 1996).

Returning to the $\alpha$ value described in Equations A.26 and A.25, this parameter can be used to scale the effects of non-affine variation in consequent analyses (Bookstein, 1991; Rohlf, 1993, 1996). For $\alpha$ values of 0, research has revealed that RWA feature spaces closely approximate a PCA calculated directly on landmark coordinates (Walker, 1996; Rohlf et al., 1996; Rohlf, 1999), with almost perfect correlation between feature spaces. $\alpha$ values approximating 1, however, can be used to enhance global shape variation, highlighting geometrically large-scale variations. $\alpha$ values approximating -1, however, perform the opposite, focusing on local shape variation.

Beyond the use of TPS interpolation functions as descriptors of morphology, TPS are a very powerful visualisation tool, and are more often used for this purpose.

The basis of TPS functions for visualisation builds on the concept of mapping shape changes between two configurations onto an orthogonal grid. In order to map differences between $X$ and a reference configuration, e.g. $\bar{X}$, the TPS warping of the $x$ coordinates of a grid is performed by first defining a column vector $V_x = (X_{:,1}, 0, 0, 0)$, which in turn is transformed to

$$L^{-1}V_x = \left(w_1, w_2, \ldots, w_p, a_1, a_x, a_y\right) \tag{A.28}$$

Using the $a$ coefficients from this vector, we can calculate the transformation function $f_x\left(x, y\right)$ of the $x$-coordiantes of a point on the plane $Z$ through with respect to a landmark $X_{i,:}$ using;

[3] Rohlf and Bookstein describe this formula using the $\bigoplus$ operator in the equation, however, in their description of the formulae they refer to the $\otimes$ operator instead (see Equation 4, Rohlf and Bookstein (2003)). For frame of reference the $\bigoplus$ operator is the "direct sum" operator. At present this operator is not available in Python/R and has to be programmed by hand.

$$f_x(x,y) = a_1 + a_x x + a_y y + \sum_{i=1}^{p} U(\|X_{i,:} Z\|) \tag{A.29}$$

$f_y(x,y)$ is similarly defined as;

$$V_y = (X_{:,2}, 0, 0, 0) \tag{A.30}$$

$$L^{-1} V_y = (w_1, w_2, ..., w_p, a_1, a_x, a_y) \tag{A.31}$$

$$f_y(x,y) = a_1 + a_x x + a_y y + \sum_{i=1}^{p} U(\|X_{i,:} Z\|) \tag{A.32}$$

An example of the resulting product for two simple icons is presented in Figure A.9.

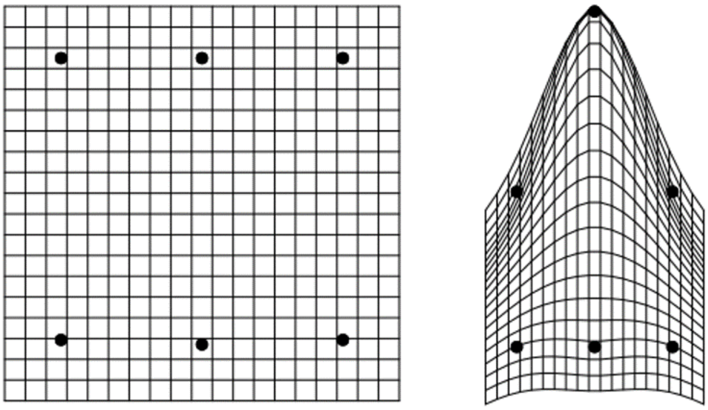


*Figure A.9: Example of a Thin Plate Spline grid warping visualisation with the reference shape in the left panel, and a target shape in the right panel*

## A.5 Some final notes on $\bar{X}$, and proposals for the robust $\tilde{X}$

Common practice in geometric morphometric practice is to use the mean landmark configuration as a reference, not only for the calculation of superimposition procedures, as described above, but also for the computation of TPS interpolation functions. Here we call the mean landmark configuration $\bar{X} : p \times k$ of $X : p \times k \times n$. We can define a landmark $p$ in $\bar{X}$ through the following formulae;

$$\bar{X}_{p,1} = \frac{1}{n} \sum_{i=1}^{n} X_{p,1,i} \; , \; \bar{X}_{p,2} = \frac{1}{n} \sum_{i=1}^{n} X_{p,2,i} \tag{A.33}$$

and for landmark configurations in three dimensions;

$$X_{\mu_{p,3}} = \frac{1}{n}\sum_{i=1}^{n} X_{p,3,i} \tag{A.34}$$

which is essentially the arithmatic mean coordinate of each point $p$. Likewise, the centroid of a single icon is typically calculated using the mean point of landmark coordinates.

Nevertheless, and in light of the observations and research carried out in the present Doctoral Thesis, it can be argued that $\bar{X}$ might be susceptible to the presence of outliers, and thus condition subsquent results. From this perspective, and so as to propose a means of integrating robust statistical measures into Geometric Morphometric analyses, we propose the possibilitiy of substituting $\bar{X}$ with $\tilde{X}$.

$\tilde{X}$ is the median landmark configuration, and can be calculated by first transforming $X$ into an ordered array $\{X_{:,:,(1)},...,X_{:,:,(n)}\}$ where $X_{:,:,(n-1)} \leq X_{:,:,(n)}$. A landmark $p$ in $\tilde{X}$ is now calculated as;

$$\tilde{X}_{p,1} = X_{p,1,(\lceil 0.5n \rceil)} \ , \ \tilde{X}_{p,2} = X_{p,2,(\lceil 0.5n \rceil)} \tag{A.35}$$

and for landmark configurations in three dimensions;

$$\tilde{X}_{p,3} = X_{p,3,(\lceil 0.5n \rceil)} \tag{A.36}$$

While this modification is relatively small, this proposal can be considered to be mathematically and theoretically important when considering the observations made throughout this Doctoral Thesis. Additionally, the supplementary materials of Courtenay et al. (Under Review-b) found the use of not only $\tilde{X}$, but also the use of robust statistics to identify a configuration centroid, to be a much more robust means of performing GPA. When compared with the Generalised Resistant Fit (GRF) algorithm proposed by Siegel and Benson (1982), Siegel and Pinkerton (1982), Rohlf and Slice (1990), Slice (1996), and Dryden and Walker (1999), the Generalised Robust Fit (GRoF) proposed by Courtenay et al. (Under Review-b) was found to not only be equally useful in performing a robust version of GPA, less affected by the *Pinnochio effect*, but is also much more computationally efficient. From this perspective, GRoF was found to fit $\approx$ 200 landmarks in 2.37 seconds on a standard laptop, while GRF took 21:37 hours to converge on a supercomputer.

# Appendix B

# Fourier Shape Descriptors

In response to the limitations presented by landmark based analyses, analysts began developing a different, yet closely linked approach, making use of Fourier descriptors as a function of shape (Rohlf and Archie, 1984; Ferson et al., 1985; Rohlf, 1986a). The principle of Fourier Analysis (FA) is to describe shape as a series of periodic functions along the curvature of an outline (Ferson et al., 1985). These types of analyses are based on prior observations that "look-alike" shapes have a tendency of occupying similar positions in the Fourier descriptor feature space (Fritzche, 1961; Raudseps, 1965). The advantage of calculating shape descriptors in this way is that FA does not require the definition of homologous landmarks; i.e. the number of points along the outline can vary.

Fourier based analyses can be performed in a number of ways; (1) by calculating the distance of any point on the outline to the centroid (Zahn and Roskies, 1972); (2) calculating the variation of the tangent angle for any point (Zahn and Roskies, 1972); or (3), analysing a series of linearly transformed circular coordinates (Giardina and Kuhl, 1977; Kuhl and Giardina, 1982). These approaches are known as Fourier Radius Variation (FRV), Fourier Tangent Angle (FTA), and Elliptic Fourier Analyses (EFA), respectively. While each approach has its advantages and disadvantages, EFA are the best suited for most type of outlines, as EFA is more robust to irregularities along the series (Rohlf, 1986b).

To demonstrate the functionality of these formulae, we will be using a single example of a mosquito wing (Fig. B.1) published by Rohlf and Archie (1984), whose dataset is openly available in the Momocs library (Bonhomme et al., 2014). Each mosquito wing in this dataset is described by 100 points along the outline. For each of the techniques, the reconstruction of the wing's outline will be presented using the cumulative contributions of the first 10 harmonics. All calculations were programmed in the R programming language (v.4.0) by L.A. Courtenay, without the use of external dependencies.

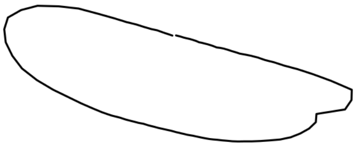


***Figure B.1:*** *An example of a mosquito wing from Rohlf and Archie (1984) that will be used to demonstrate the functionality of FA in the description of shape.*

## B.1 Outline Superimposition

While Fourier Shape Descriptors are not used for the analysis of Procrustes superimposed coordinates, a certain level of superimposition may be required prior to analyses. This can be performed using the left-singular unitary matrix ($U$), obtained from the Single Value Decomposition ($U\Sigma V^T$) of the covariance matrix ($C$), for each set of coordinates ($X$);

$$\sigma(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{B.1}$$

$$C = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix} \tag{B.2}$$

$$C = U\Sigma V^T \tag{B.3}$$

$$X' = XU \tag{B.4}$$

producing aligned coordinates $X'$.

## B.2 Fourier Series

Fourier series are used to describe shapes by decomposing a periodic function into a sum of more simple trigonometric functions, such as *sine* and *cosine* values. The Fourier Series formula can be described as follows;

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx) \tag{B.5}$$

where;

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)dx \tag{B.6}$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)dx \tag{B.7}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)dx \tag{B.8}$$

If we describe an outline as a set of coordinates, $f(x)$ can be converted into $x(t)$ and $y(t)$, representing both the $x$ and $y$ coordinates as functions, and where $t$ is the curvilinear abscissa from 0 to $T$. In this sense, $T$ is defined as;

$$T = \sum_{i=1}^{n} \sqrt{\Delta x_i^2 + \Delta y_i^2} \tag{B.9}$$

which describes the perimeter (period of the signal, or fundamental frequency) of the profile. Following this, the pulse, or angular frequency of the signal, can be defined through;

$$\omega = \frac{2\pi}{T} \tag{B.10}$$

We can thus replace $\pi$ from eq. B.6-B.7 with $T$, and insert $\omega$ to define the formulae;

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\omega t) + \sum_{n=1}^{\infty} b_n \sin(n\omega t) \tag{B.11}$$

$$y(t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos(n\omega t) + \sum_{n=1}^{\infty} d_n \sin(n\omega t) \tag{B.12}$$

$$a_n = \frac{2}{T}\int_0^T x(t)\cos(n\omega t)dt \tag{B.13}$$

$$b_n = \frac{2}{T}\int_0^T x(t)\sin(n\omega t)dt \tag{B.14}$$

$$c_n = \frac{2}{T}\int_0^T y(t)\cos(n\omega t)dt \tag{B.15}$$

$$d_n = \frac{2}{T}\int_0^T y(t)\sin(n\omega t)dt \tag{B.16}$$

## B.3 Fourier Radius Variation Coefficients as Descriptors of Morphology

Fourier Radius Variation formulae for a given curve can be defined as the distance from the centroid to the outline, i.e. the radius $r$. We can describe this as a function of $t$ through;

$$r(t) = \mu_0 + \sum_{i=1}^{\infty} (a_i \cos(nt) + b_i \sin(nt)) \tag{B.17}$$

where

$$\mu_0 = \frac{1}{2\pi}\int_0^{2\pi} r(t)dt \tag{B.18}$$

$$a_n = \frac{1}{\pi}\int_0^{2\pi} r(t)\cos nt\, dt \tag{B.19}$$

$$b_n = \frac{1}{\pi} \int_0^{2\pi} r(t) \sin nt \, dt \tag{B.20}$$

If we describe the curve as a set of points $p$ along the outline, we can calculate the radius $r$, being a measurement of the distance from $p_i$ to the centroid (Fig. B.2), as a periodic function, and thus expand the previous formulae such that;

$$r(\theta) = \frac{1}{2}a_0 + \sum_{n=1}^{p} a_n \cos(\omega_n t) + b_n \sin(\omega_n t) \tag{B.21}$$

where $n$ is the number of harmonics describing the series. Using this expansion of the original formulae, $a_n$ and $b_n$ are now defined as;

$$a_n = \frac{2}{p} \sum_{i=1}^{p} r_i \cos(nt_i) \tag{B.22}$$

$$b_n = \frac{2}{p} \sum_{i=1}^{p} r_i \sin(nt_i) \tag{B.23}$$

and we substitue $\mu_0$ for $a_0$ as;

$$a_0 = \sqrt{\frac{2}{p} \sum_{i=1}^{p} r_i} \tag{B.24}$$

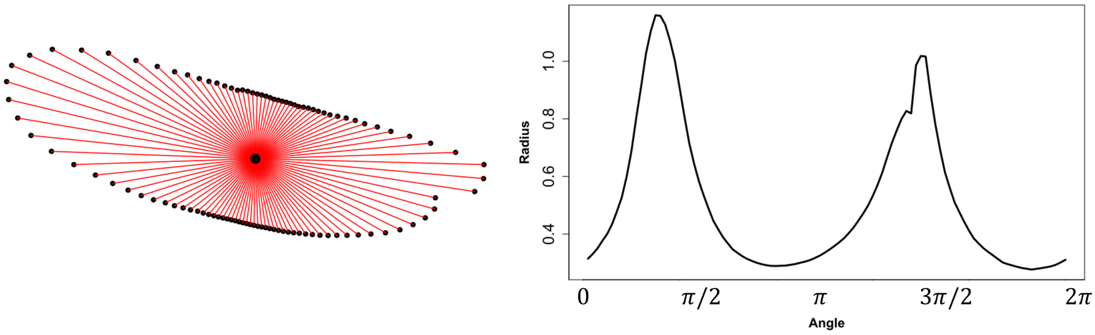


***Figure B.2:*** *A diagram representing the means in which outlines are described using Fourier Radius Variation*

Figure B.3 demonstrates how FRV harmonics can be used to reconstruct the shape of the mosquito wing presented in Figure B.1.

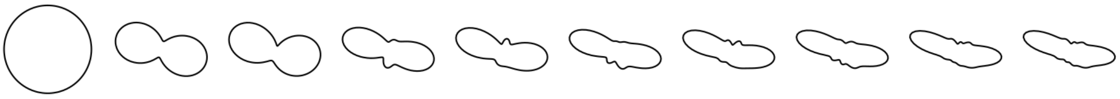


***Figure B.3:*** *Diagram presenting the reconstruction of a mosquito wing's outline based on the cumulative contributions of the first 10 Fourier Radius Variation harmonics*

## B.4 Fourier Tangent Angle Coefficients as Descriptors of Morphology

Fourier Tangent Angles as descriptors of morphology are very similar to the previously defined Radius Variation descriptors. From this perspective, most formulae remain the same, with the exception that instead of calculating the distance from a point to the centroid, FTA formulae take a scaled outline with range $[0,2\pi]$, and express change through a function of $\phi(t)$. $\phi(t)$ is expressed as the angle from a point at a certain distance $(t)$ along the line $(\theta_t)$ to the starting point $(\theta_0)$, and is calculated as $\phi(t) = \theta_t - \theta_0 - t$ (Fig. B.4). We can substitute this measure into the FRV formulae (Appendix B.4) such that coefficients are calculated through;

$$a_0 = \sqrt{\frac{2}{p}\sum_{i=1}^{p}\phi(t)} \tag{B.25}$$

$$a_n = \frac{2}{p}\sum_{i=1}^{p}\phi(t)\cos(n\theta_i) \tag{B.26}$$

$$b_n = \frac{2}{p}\sum_{i=1}^{p}\phi(t)\sin(n\theta_i) \tag{B.27}$$

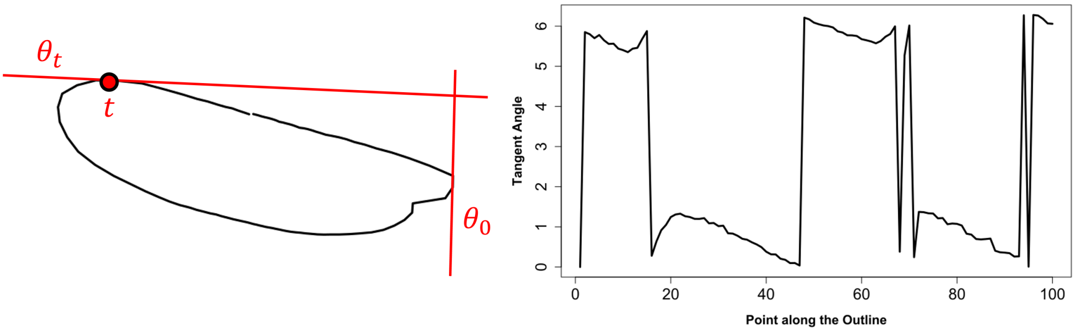


***Figure B.4:*** *A diagram representing the means in which outlines are described using Fourier Tangent Angles*

Figure B.5 demonstrates how FTA harmonics can be used to reconstruct the shape of the mosquito wing presented in Figure B.1.

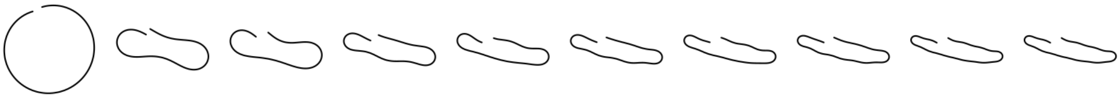


***Figure B.5:*** *Diagram presenting the reconstruction of a mosquito wing's outline based on the cumulative contributions of the first 10 Fourier Tangent Angle harmonics*

## B.5 Elliptic Fourier Coefficients as Descriptors of Morphology

Shapes and outlines can be described by a set of Fourier series coefficients, $a_n$, $b_n$, $c_n$ and $d_n$, where $n$ is the number of harmonics used to describe the series (Ferson et al., 1985), and $k$ is the number of points along the outline ($i$). Fourier transforms are thus used to analyse the displacement of $x$ and $y$ coordinates along the outline (Fig. B.6). Each harmonic $n$ can be calculated through;

$$a_n = \frac{T}{2\pi^2 n^2} \sum_{i=1}^{p} \frac{\Delta x_i}{\Delta t_i} \left( \cos \frac{2\pi n t_i}{T} - \cos \frac{2\pi n t_{i-1}}{T} \right) \tag{B.28}$$

$$b_n = \frac{T}{2\pi^2 n^2} \sum_{i=1}^{p} \frac{\Delta x_i}{\Delta t_i} \left( \sin \frac{2\pi n t_i}{T} - \sin \frac{2\pi n t_{i-1}}{T} \right) \tag{B.29}$$

$$c_n = \frac{T}{2\pi^2 n^2} \sum_{i=1}^{p} \frac{\Delta y_i}{\Delta t_i} \left( \cos \frac{2\pi n t_i}{T} - \cos \frac{2\pi n t_{i-1}}{T} \right) \tag{B.30}$$

$$d_n = \frac{T}{2\pi^2 n^2} \sum_{i=1}^{p} \frac{\Delta y_i}{\Delta t_i} \left( \sin \frac{2\pi n t_i}{T} - \sin \frac{2\pi n t_{i-1}}{T} \right) \tag{B.31}$$

These coefficients can be directly used for multivariate analyses, however, in the case where coefficients are to be used as descriptors of shape, as opposed to form, values are additionally normalized and scaled (Ferson et al., 1985). This normalization process is computed using;

$$\begin{bmatrix} A_n & B_n \\ C_n & D_n \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix} \tag{B.32}$$

Here $\lambda$ defines the “semimajor axis of the best fitting ellipse” (Ferson et al., 1985);

$$a' = A_1 \cos\theta + B_1 \sin\theta \tag{B.33}$$

$$c' = C_1 \cos\theta + D_1 \sin\theta \tag{B.34}$$

$$\lambda = \sqrt{a'^2 + c'^2} \tag{B.35}$$

$\theta$ is the orientation of said ellipse in radians;

$$\theta = \arctan\left(\frac{c'}{a'}\right), \quad 0 \leq \theta < 2\pi \tag{B.36}$$

and $\psi$ is the rotation of the starting point from the end of the ellipse;

$$\psi = \frac{1}{2} \arctan \frac{2(a_1 b_1 + c_1 d_1)}{a_1^2 + c_1^2 - b_1^2 - d_1^2}, \quad 0 \leq \theta < \pi \tag{B.37}$$

The result of this transformation is the cancelling out of coefficients $a_1$, $b_1$ and $c_1$, as well as the elimination of size.

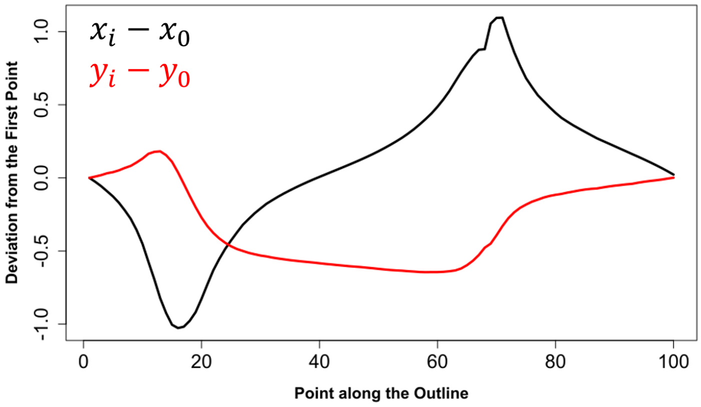


***Figure B.6:*** *A diagram representing the means in which outlines are described using Elliptic Fourier Analyses. For the purpose of describing morphology, coefficients are based on x and y coordinates along the outline, for this reason only the oscilloscope has been displayed in this figure.*

Figure B.7 demonstrates how EFA harmonics can be used to reconstruct the shape of the mosquito wing presented in Figure B.1.

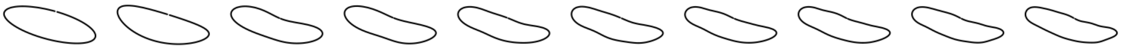


***Figure B.7:*** *Diagram presenting the reconstruction of a mosquito wing's outline based on the cumulative contributions of the first 10 Elliptic Fourier harmonics*

## B.6 Defining an Optimal Number of Harmonics

As can be seen throughout this section, the use of Fourier Descriptors is dependent on the definition of $n$. The first few harmonics, regardless of the technique, describe morphology as very abstract and vague representations of the overall shape, generally represented as a circle or an ellipsis. As harmonics are added, the cumulative contribution of harmonics begin to resemble the original shape.

Little agreement exists on the precise number of harmonics to be used (Claude, 2008), while in general higher level harmonics are mostly described by noise of minor deviations in outline morphologies. Nevertheless, we can calculate the descriptive power ($P$) of FRV and FTA harmonics through;

$$P_n = \frac{A_n^2 + B_n^2}{2} \tag{B.38}$$

and we can calculate the descriptive power of EFA harmonics through;

$$P_n = \frac{A_n^2 + B_n^2 + C_n^2 + D_n^2}{2} \tag{B.39}$$

# Appendix C

# *p*-Value Evaluation

In the October of 2017, the American Statistical Association (ASA) held a 2 day symposium to discuss the use of *p-Values and Statistical Significance*. Product of said symposium can be summarised in the 73rd Volume (Issue Sup1) of *The American Statistician*; a collection of 43 detailed articles on the use and misuse of *p-Values*, where the term *"Statistically significant"*, alongside recommendations on how to proceed in a world beyond $p$<0.05 (Wasserstein et al., 2019), are discussed in great detail. While this collection builds on the original statements and principles proposed by the ASA after the symposium of October, 2015, and published in 2016 (Wasserstein and Lazar, 2016), the more extensive special issue of 2019 strongly discourages the use of the term *"significant"*. Under this premise, careful evaluation of each of the proposals has led to the conclusion that 0.05 is no longer an acceptable threshold. Throughout the present Doctoral Thesis, arguments are proposed to substitute 0.05 with 0.003, as will be explained in continuation.

## C.1 A Brief Overview of *p*-Value Historiography

The $p$-value has a long history, arguably dating back to the XVIIIth century, while popularised and standardised during the beginning of the XXth century (Kennedy-Schaffer, 2019). Among the many possible authors, notable early usage of $p$-values include some of the most relevant names in statistics; Laplace (1827), Poisson (1837), and Pearson (1900), *inter alia*. Nevertheless, few authors actually make an attempt at clarifying an acceptable "threshold" or norm from which future analysts can base their work on. While as early as Edgeworth (1885)'s article we can find a possible reference to a certain threshold of importance;

> "*by far the greater proportion (namely 0.995)... Accordingly, if out of a set of [...] N statistical numbers which fulfil the law of error, we take one at random, it is exceedingly improbable that it will differ from the Mean to the extent of twice, and* à fortiori *thrice, the modulus... Such is the nature of the law of error*" - Edgeworth (1885), Page 185

To some, at the time, 0.005 was considered excessive and not adaptable to all situations.

Building on this, the most notable and widely-accepted contributions in this field of research, therefore, are not found until the later works of Pearson (1900), Elderton (1902) and Student (1908) (whose real name was William Sealy Gosset). From each of these studies, we find valuable insights into early statistical reasoning, namely with the definition of probability distributions that laid down a framework from which $p$-value calculations would be standardised and developed ($\chi^2$, $t$ & $z$ distributions). Product of these studies are the first detailed publications of $p$-values, with notable contributions by R.A. Fisher. Alongside a series of articles, Fisher published his 1925 book titled *Statistical Methods for Research Workers* (Fisher, 1925), containing the referential passage;

> "*The value for which p = 0.05... is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.*" - Fisher (1925), Page 47

Here it is important to point out specifically the phrase *"twice the standard deviation"*, from which it can be seen that 0.05 was not an arbitrarily chosen value, but more theoretically derived from $1-2\sigma \approx 0.05$. Nevertheless, it is also important to point out that Fisher did not only use 0.05, and in some cases we find his tables including a large number of different p-values, with reference to 0.01 as another threshold for different circumstances.

Nevertheless, Fisher's work was not published without some debate. In a series of studies by Neyman and Pearson (1933a,b), we find a slightly different approach to defining, not only $p$-Value calculations in hypothesis testing, but also the reasoning behind using 0.05 as a threshold;

> "*...the testing of statistical hypotheses cannot be treated as a problem of estimation*" (pg. 492) "*$H_0$ is accepted when some alternative $H_i$ is true, the conesquences that follow will depend upon the nature of $H_i$ and its difference from $H_0$... while it is the 'size' of $p_I(w)$ that matters, it is what may be termed the quality of errors contributing to the risk $P_{II}(w)$, that must be taken into account*" - Neyman and Pearson (1933a), Page 497

> "*We deal first with the class of alternatives for which $\alpha > \alpha_0$. If $\varepsilon = 0.05$... we shall proabbly decide to accept the hypothesis $H_0$ as far as this class of alternatives is concerned*" - Neyman and Pearson (1933b), Page 303

Once again, we also find the authors contemplating a more robust threshold of 0.01 depending on the type of probability distributions at hand (Neyman and Pearson, 1933b).

Here, Neyman and Pearson approach the concept of hypothesis testing from a different perspective, stating that the acceptance of a hypothesis should not just depend on the Null Hypothesis' ($H_0$) distribution, but also be contextualised with the Alternative Hypothesis' ($H_a$) distribution (Neyman and Pearson, 1933a). From this perspective, these authors search for a means of reducing possible error, with Type I ($P_I$) and Type II ($P_{II}$) statistical errors being False Positive and False Negative errors accordingly. In sum, Neyman and Pearson thus discuss a significance level $\alpha$ as 0.05 or 0.01 for $P_I$ risk, while $\beta$ = 1 - *statistical power* as the $P_{II}$ risk.

While the $p$-Value threshold between the two approaches does not differ greatly, the means of deriving this threshold is quite different. Under this premise, it is possible to argue that this constant mention of 0.05 across multiple sources, even if derived from different calculations, is what has generated such confusion in the use of $p$-Values over the years. As 0.05 appears to resonate throughout history of applied statistics, with few questions regarding the reliability of this threshold, many authors misuse the concept of $p < 0.05$ in hypothesis testing without understanding the true origin of this value.

While the present body of research will refrain from declaring a preference for one approach or the other, we feel it is important to note both approaches have their advantages and disadvantages, and a general understanding of their origin is necessary in order to correctly evaluate $p$-value calculations. If we are to consider some of the more recent research into fields of robust statistics, it may be considered that $2\sigma$ is not always the best reflection of reality when dealing with non-symmetrical empirical probability distributions.

While the $t$ and $z$ distributions are more robust to a number of these cases, presenting longer tails and a means of making general inferences regarding the population, limitations are still present and the definition of a single $p$-Value threshold is not always as straightforward.

## C.2 p-Value Calibrations and the use of Bayes Theorem

In order to define the methods that can be used to evaluate $p$-Values, we think it's important to define what $p$-Values and hypothesis testing refer to in the XXIst century. For this purpose we quote Wasserstein and Lazar's 2016 ASA statement;

> "*a p-Value is the probability under a specified statistical model that a statistical summary of the data [...] would be equal to or more extreme than its observed value.*" - Wasserstein and Lazar (2016), Page 131

Among the many papers presented by the ASA's 2019 Special Issue, an important point raised by most authors builds on the exclusively Frequentist perspective most scientists have upon approaching hypothesis testing. In as such, most authors who contributed to the Special Issue raised the point that one of the best means of overcoming $p$-Value misuse would be to find a compromise between Bayesian and Frequentist approaches. As pointed out by Colquhoun (2019), however, this is not an easy task and many analysts are likely to try and avoid these approaches.

Part of the Bayesian-Frequentist conflict stems from the distrust many non-statisticians have with the concept of Bayesian Priors (Martin, 2018; Colquhoun, 2019). While in many fields of science, a typical argument against Bayesian approaches consider that not enough is known about certain phenomena in order to construct truly informative priors for Bayesian analysis, in many cases this does not mean that Bayesian approaches should be discarded. As pointed out by most Bayesian practitioners, the use of priors is only *one* component of the theorem, and while informative priors can be greatly beneficial to the quality of results, weakly-informative priors, or even diffuse priors, are still an option if little is known about the subject matter (Martin, 2018). Similarly, weakly-informative priors are a popular technique for regularisation in many Bayesian learning strategies.

When using Bayes' theorem for hypothesis testing, we can define the following formula (eq. C.1);

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \therefore \frac{P(H_a|x)}{P(H_0|x)} = \frac{P(x|H_a)}{P(x|H_0)} \cdot \frac{P(H_a)}{P(H_0)} \tag{C.1}$$

where $x$ is our data, $H_a$ is the alternative and $H_0$ the null hypothesis. Observing this equation, it can immediately be seen how Bayesian approaches have the distinct advantage of meeting the requirements set forth by Neyman and Pearson (1933a,b), whereby the equation considers the test's results not only with regards to $H_0$ but also in context with $H_a$. Similarly, Bayesian statistics employ the use of prior and posterior probabilities to quantify levels of uncertainty within a statement (Martin, 2018), providing statisticians with a means of defining their certainty or uncertainty about $H_a$. Furthermore, this not only implies that the reliability of a $p$-Value threshold can be empirically assessed, but is also adaptable to the specific data at hand. In Bayesian statistics, this is usually seen by how the final output of a model is not a single value, but a probability distribution of all the possible values (Martin, 2018).

The Bayesian alternative to $p$-Values in hypothesis testing are known as Bayes Factors (BFs). A BF is the likelihood ratio of the marginal likelihoods of $H_a$ and $H_0$. These values are usually interpreted in ranges, with BF 1 to 3 being weak, 3 to 10 being moderate, 10 to 30 being substantial, 30 to 100 being strong, and $>100$ being very strong evidence in favour for $H_a$ (Held and Ott, 2018). In a similar sense, different authors have made attempts at categorizing $p$-Values in this manner, with 0.1 to 0.05 implying weak, 0.05 to 0.01 moderate, 0.01 to 0.001 strong, and <0.001 very strong evidence for $H_a$ (Cox and Donnelly, 2011; Bland, 2015). Nevertheless, in some senses this still assumes a one-size-fits-all $p$-Value, regardless of specific $H_a$ and $H_0$ distributions a study may have. Likewise, this can also create an illusion that the $p$-Value is a calculation of the probability of something being true, while being namely conditioned on the $H_0$ (Held and Ott, 2018).

To directly confront the differences between Frequentist derived $p$-Values and Bayesian metrics, multiple efforts have been made to propose a means of "calibrating" p-value with BFs (Sellke et al., 2001; Held and Ott, 2018; Benjamin and Berger, 2019). Nevertheless, not all produce the same results (Benjamin and Berger, 2019). In part this can be explained by some of the issues in which p-values have been defined, whereby the conversion of $p$-values according to Fisher's definition require a different formula (eq. C.2) (Fisher, 1925), to those proposed by Neyman and Pearson (1933a,b) (eq. C.3);

$$B(p) = -e\,p\log(p) \tag{C.2}$$

$$\alpha(p) = \left(1 + [-e\,p\log(p)]^{-1}\right)^{-1} \tag{C.3}$$

Equation C.2 is defined as the lower bound on the odds provided by the data for $H_0$ and $H_a$. Equation C.3, on the other hand, is defined as the posterior probability of $H_0$ based on equation C.2, combined with the assumption of an equal prior probability for $H_a$ and $H_0$ (i.e. 0.5) (Sellke et al., 2001). Under this premise, we can see how, once again, $\alpha(p)$ fulfills Neyman and Pearson's criteria on defining probability distributions for both $H_a$ and $H_0$.

In the 2019 ASA Special Issue, we find an interesting development of these equations by Benajmin and Berger, providing 3 recommendations for improving the use of $p$-Values (Benjamin and Berger, 2019). The authors; (1) argue that 0.05 should be replaced by 0.005, with values between 0.05 and 0.005 being referred to as "suggestive" evidence, rather than "significant"; (2) suggest that analysts report a Bayes Factor Bounds (BFB) value that supports the reported $p$-Value; and (3) suggest that analysts report the posterior odds of $H_1$ to $H_0$.

Firstly, with regards to point (2), a BFB is the upper bound, or confidence interval, with regards to traditional BFs. Benjamin and Berger thus define BFB as an indication for "*the highest possible BF consistent with the observed p-Value*" (Benjamin and Berger, 2019). Adapting equation C.2 to this statement, BFB can thus be defined as (eq. C.4);

$$BF \leq BFB \equiv \frac{1}{-e\;p\;\log(p)} \tag{C.4}$$

BFBs are generally reported *at most* as odds against the $H_0$ (Benjamin and Berger, 2019). Their interpretation, therefore, can consider 0.05, whose corresponding BFB value is 2.44, indicating 2.44:1 odds that $H_a$ is correct. For those who find odds hard to interpret, and would thus prefer a % value, BFBs can easily be converted to percentages via the following formula (eq. C.5);

$$P^U\left(H_a|p\right) = \frac{BFB(p)}{1 + BFB(p)} \tag{C.5}$$

The third recommendation presented by Benjamin and Berger requires these values be presented as posterior odds. This is simply carried out by multiplying the BFB value by the prior odds.

One key disadvantage to the use of BFs is that they require the specification of a prior probability distribution, a topic that frequently produces disagreements, or is too complex to define. In cases where the prior probabilities are unknown, this value can be derived from other variables (see: http://fpr-calc.ucl.ac.uk/ Colquhoun (2017)), however, as suggested by Colquhoun (2019), one fair assumption to make is that these probabilities are equal (50:50, a.k.a. random). This is considered a valid diffuse prior for most hypothesis testing. In light of this, posterior odds for a $p$-value of e.g. 0.05, would have a posterior probability of $BFB(0.05) \cdot 0.5$.

For ease of comparisons, Table C.1 (located towards the end of this Appendix) contains a number of $p$-values calibrated with BFB, $P^U\ (H_a|p)$ and their corresponding posterior odds.

The final consideration made regarding $p$-Values in the present body of research employed the use of the False Positive Risk (FPR) approach proposed by Colquhoun (2019).

The FPR is a powerful notion developing a number of the components seen up to this point. FPR attempts to quantify the probability of declaring an effect is real, when it is, in fact, not. This is very similiar to Neyman and Pearson's notions of the $\alpha$ threshold and $P_I$ risk, considering both refer directly to the False Positive Rate of a prediction. In order to calculate this, Colquhoun uses Bayes' theorem to define the likelihood ratio in favour of $H_a$ as equation C.6, where FPR can then be calculated through equation C.7 and in the case where equal prior probabilities for $H_a$ and $H_0$ are defined we can simplify this equation to equation C.8;

$$L_{10} = \frac{P(x|H_a)}{P(x|H_0)} \tag{C.6}$$

$$FPR = \frac{1}{1 + L_{10}\frac{P(H_a)}{1-P(H_a)}} \tag{C.7}$$

$$FPR = \frac{1}{1 + L_{10}\frac{0.5}{0.5}} = \frac{1}{1 + L_{10}} \tag{C.8}$$

Throughout his article, Colquhoun provides three main formulae for defining $L_{10}$, each with their own peculiarities. For the purpose of simplicity, however, and so as to provide a means of calibrating Benjamin and Berger's BFB calculations with FPR, this Doctoral Thesis has employed the Sellke-Berger approach, defined in Appendix A2 of Colquhoun (2019), using equation C.4 as $L_{10}$. Interestingly, using BFB in this sense forces FPR to be the inverse of $P^U\ (H_a|p)$ (see Fig. C.2), creating a complementary metric that calculates the probability of Type I errors being present alongside the upper-bound probability of $H_a$ being true.

In continuation, the following appendix provides a series of tables and figures that can be used as a frame of reference for both the BFB and FPR approaches. Figures C.1 & C.2 represent a graphical representation of the trend between $p$-Values and each of the Bayesian-based metrics. One characteristic of Figure C.1, that is of notable importance, is the difference between Bayes values when using $\alpha(p)$ and $B(p)$, especially at $p = 0.05$. From here it can be seen that only when $p \approx 0.005$, or closer to $3\sigma$, do the two values coincide, likely indicating that 0.005 or 0.003 be a more robust threshold to consider a $p$-Value as "significant" or not in the eyes of Benjamin and Berger (2019)'s Recommendation 0.1.

When consulting plots of curves for each of these equations, interesting patterns emerge, namely a non-symmetric U-shaped curve (Fig. C.1 & C.2), indicating BFB or FPR values to not be unique to one single $p$-Value. This can consequently be interpreted as a means for reversing the focus from $H_a$ to $H_0$, with values falling on one side of the curve supporting either one of these hypotheses. To find the limit between

$p(H_a)$ and $p(H_0)$, a simple experiment was devised, calculating the point of maximum curvature for each curve using steps of 0.0001 across $p$-Values. This experiment found $p = 0.3681$ as the limit between each hypothesis.

Finally, to observe how different priors influence FPR and $P^U(H_a|p)$ curves, Figure C.3 shows how priors affect the morphology of these curves.

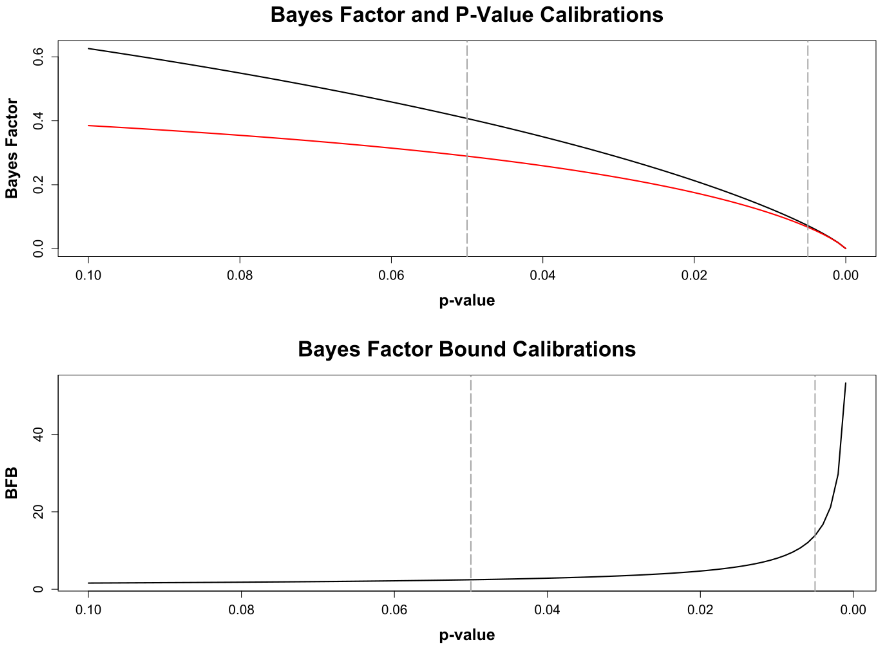


***Figure C.1:*** *p-Value calibration curves for p-Values between 0.1 and 0.0001. Top Panel: Visualisation of calibration curves for different p-Values using the $B(p)$ (black) and $\alpha(p)$ (red) formulas described in equations C.2 & C.3 respectively. Bottom Panel: Visualisation of the Bayes Factor Bound (BFB) calibration curve for different p-Values. Vertical grey dotted lines mark p = 0.05 and p = 0.005 respectively. Figure created using base-R.*

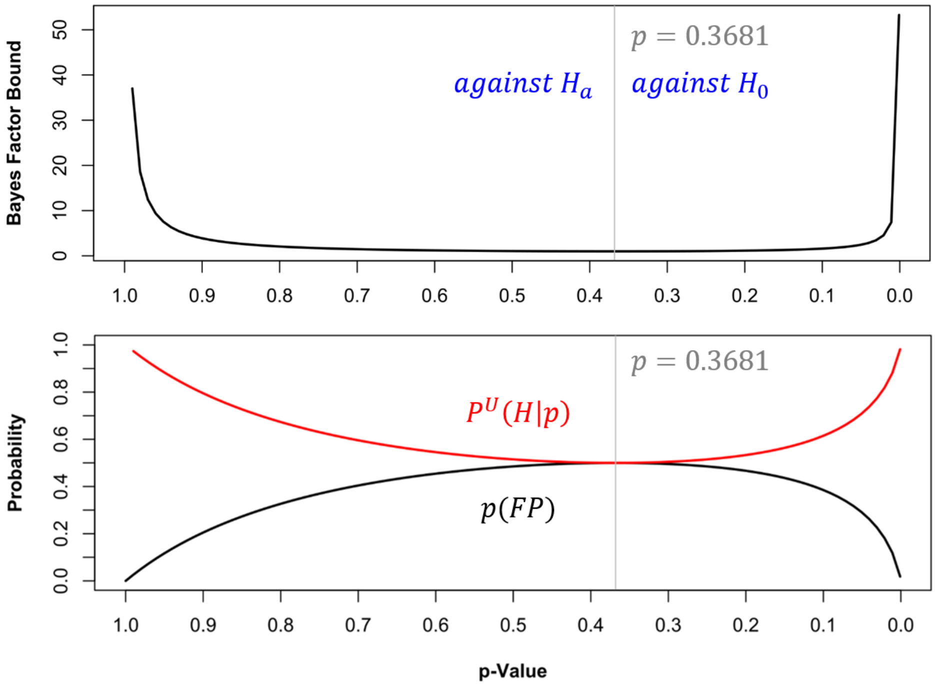


***Figure C.2:*** *p-Value calibration curves for p-Values between 1 and 0.001. Top Panel: Bayes Factor Bound calibration curves (eq. C.4). Bottom Panel: Upper bound probability supporting each hypothesis (red; eq. C.5) and probability of False Positives (black; eq. C.7). Grey line marks the point of maximal curvature along the lines. Figure created using base-R.*

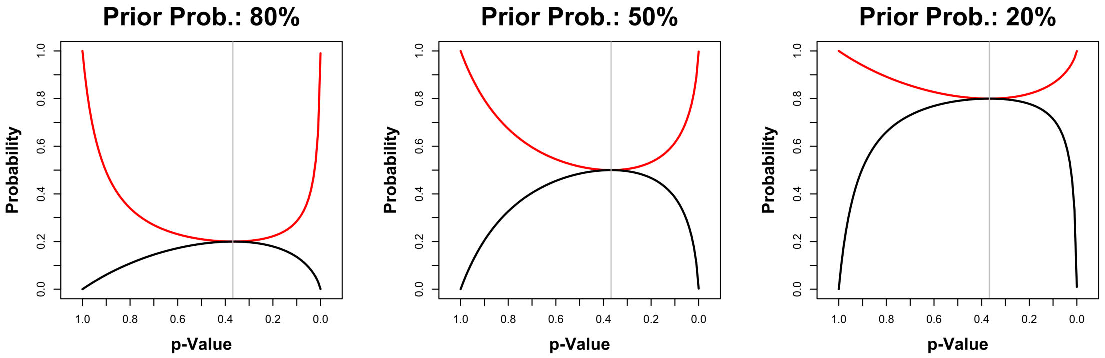


***Figure C.3:*** *Plots describing different variations of Fig. C.2's Bottom Panel using different Priors. Grey line marks the point of maximal curvature along the lines. Figure created using base-R.*

Table C.1: *Bayes Factor Bounds (BFBs), Posterior Probability of $H_a$ values ($P^U(H_a|p)$), Posterior Odds of $H_a$ to $H_0$ values ($H_a$:$H_0$), and False Positive Risk (FPR) values, for a number of their corresponding p-Values. Prior odds for $H_a$:$H_0$ and FPR calculations were set using a a number of different prior odds*

| prior | $p$ | 0.368 | 0.100 | 0.050 * | 0.010 | 0.005 | 0.003 † | 0.001 | 0.0001 |
|---|---|---|---|---|---|---|---|---|---|
| | BFB | 1.000 | 1.598 | 2.456 | 7.988 | 13.89 | 21.11 | 53.26 | 399.4 |
| | $P^U(H_a\|p)$ | 0.500 | 0.615 | 0.710 | 0.889 | 0.933 | 0.955 | 0.982 | 0.998 |
| 0.2 | $H_a$:$H_0$ | 0.200 | 0.320 | 0.491 | 1.600 | 2.778 | 4.222 | 10.65 | 79.88 |
| 0.2 | FPR | 0.800 | 0.715 | 0.620 | 0.334 | 0.224 | 0.159 | 0.070 | 0.010 |
| 0.5 | $H_a$:$H_0$ | 0.500 | 0.799 | 1.220 | 3.994 | 6.943 | 10.55 | 26.63 | 199.7 |
| 0.5 | FPR | 0.500 | 0.385 | 0.289 | 0.111 | 0.067 | 0.045 | 0.018 | 0.002 |
| 0.8 | $H_a$:$H_0$ | 0.800 | 1.278 | 1.964 | 6.391 | 11.11 | 16.89 | 42.60 | 319.5 |
| 0.8 | FPR | 0.200 | 0.135 | 0.092 | 0.030 | 0.177 | 0.012 | 0.005 | 0.001 |

* *p*-Values analogous with Fisher (1925)'s definition of "deviations exceeding twice the standard deviation"; † adaption of Fisher (1925)'s definition using the third standard deviation.

As evidenced by the numerous equations, tables and figures provided within this Appendix, part of the reasons behind the choice of $p < 0.005$ as being significant, and where possible our recommendation of $p < 0.003$, is based on how $p = 0.05$ only implies a 71% probability in favour for our Alternative Hypothesis when our prior probabilities are 1:2, leaving a 28.9% risk for Type I errors, while $p = 0.005$ lowers this risk to 6.74%, and $p = 0.003$ lowers the risk even further to 4.5 % of Type I errors.

Finally, considering observations by Courtenay et al. (2021b) and subsequent studies by Courtenay et al. (2021c,a), another means of calibrating *p*-Values, especially those above the 0.3681 threshold, is to calculate the probability of $H_0$ being true after having accepted $H_a$. From this perspective, values below 0.3681 will simply be the FPR, while values of $p > 0.3681$ are calibrated using the inverse of the FPR as follows;

$$IFPR = \frac{1}{1 + L_{10}\left(1 - \left(\frac{P(H_a)}{1-P(H_a)}\right)\right)} \tag{C.9}$$

defining $P(H_0)$ as;

$$P(H_0) = \begin{cases} FPR(p), & p \leq 0.3681, \\ 1 - IFPR(p), & p > 0.3681 \end{cases} \tag{C.10}$$

The following table presents some $p$ to $P(H_0)$ calibrations;

Table C.2: $P(H_0)$ for a number of their corresponding *p*-Values using different priors

| $p$ | 0.999 | 0.800 | 0.500 | 0.368 | 0.100 | 0.050 | 0.010 | 0.005 | 0.003 |
|---|---|---|---|---|---|---|---|---|---|
| 0.2 Prior | 0.993 | 0.892 | 0.809 | 0.800 | 0.715 | 0.620 | 0.334 | 0.224 | 0.159 |
| 0.5 Prior | 0.974 | 0.673 | 0.515 | 0.500 | 0.385 | 0.289 | 0.111 | 0.067 | 0.045 |
| 0.8 Prior | 0.902 | 0.340 | 0.210 | 0.200 | 0.135 | 0.092 | 0.030 | 0.018 | 0.012 |

# Appendix D

# The Neural Network "Black Box"

The objectives of this Appendix are to simply visualise the internal properties of a Neural Network. As has already been presented in Courtenay et al. (2020b), a Neural Network (NN), in its simplest form, is a series of mathematical operations that try to replicate actual biological neural functionality. These networks can sometimes consist of hundreds of neurons, organised in successive layers. The simplest of NNs are represented by a simple *input layer*, where every variable is introduced into the network by a neuron in the input layer, which is connected to a *hidden layer* of neurons, before being mapped out to the *output*. The number of neurons in the output depend on the type of output required. NNs are highly versatile and can be used for many different tasks from regression to classification, among many others.

We can define a NN as;

$$h_{1,i} = f_1\left(\sum_{j=1}^{n_0} w_{i,j}^{(1)} x_j + b_i^{(1)}\right) \quad : \quad i = 1, ..., n_1 \tag{D.1}$$

$$h_{2,i} = f_2\left(\sum_{j=1}^{n_1} w_{i,j}^{(2)} h_{1,j} + b_i^{(2)}\right) \quad : \quad i = 1, ..., n_2 \tag{D.2}$$

$$\hat{y}_i = f_3\left(\sum_{j=1}^{n_2} w_{i,j}^{(3)} h_{2,j} + b_i^{(3)}\right) \quad : \quad i = 1, ..., n_3 \tag{D.3}$$

where each of the inputs $x_i$ are mapped to the output $\hat{y}_i$ by a series of weights that pass information from neuron to the next. Each neuron is additionally connected to a bias neuron $b$. $f()$ is known as an activation function, which controls the flow of information from one neuron to the next, and can be tuned to impose non-linearity on the general functionality of the NN. This is a slightly more complete definition of a NN than that presented in Courtenay et al. (2020b).

Numerous activation functions exist, however the most common include;

$$\text{Linear} \quad f(x) = x \tag{D.4}$$

$$\text{Rectified Linear Unit (ReLU)} \quad f(x) = max(0, x) \tag{D.5}$$

$$\text{Tanh} \quad f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{D.6}$$

$$\text{Sigmoid} \quad f(x) = \frac{1}{1+e^{-x}} \tag{D.7}$$

$$\text{Softmax} \quad f(\vec{x}) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{D.8}$$

The simplest of activation functions is the Linear function (eq. D.4), with range $(-\infty,\infty)$, which is typically used when NNs are trained to perform tasks such as linear regression. ReLU (eq. D.5) is a non-linear transformation of this function, with range $[0,\infty)$, and is seemly the most popular activation function. The Tanh activation function (eq. D.6), which is bound between $[-1,1]$ is also a commonly used activation function, and is used by Courtenay and González-Aguilera (2020) for the augmentation of scaled values between $[-1,1]$. Finally, both Sigmoid (eq. D.7) and Softmax (eq. D.8) are commonly used as activation functions for the output of NNs. Sigmoid is bound between $[0,1]$, and is thus frequently used for binary classification tasks, e.g. Courtenay et al. (2020b). Softmax, on the other hand, has the additional property of accepting a vector of $x$ values, $\vec{x}$, of length $K$ (eq. D.8). This ensures that the vector produced by Softmax consists of values bound between $[0,1]$, while the sum of this vector is 1. Softmax is thus useful for non-binary classification problems, such as those presented in Courtenay et al. (2021b) and Courtenay et al. (2023).

An increasingly popular development of equation D.5 also exists combining its properties with that of equation D.7, known as the Self-Gated “Swish” activation function (eq. D.9);

$$\text{Swish} \quad f(x) = x \cdot \frac{1}{1+e^{-x}} \tag{D.9}$$

This is a version of ReLU whose first and second derivatives are much smoother than that of ReLU. This function can be calculated to be bound between $(\approx -0.3,\infty)$, and has been used in Courtenay et al. (2021b) and Courtenay et al. (2023).

The mathematics behind training a NN are complex and go far beyond the scope of this brief demonstration. For further reading on this topic consult Goodfellow et al. (2016), or the reference list of Courtenay et al. (2020b).

## D.1 A simple Neural Network

For the purpose of this demonstration, a simple NN was designed for the classification of the different types of trampling bone surface modifications described by Courtenay et al. (2020c). In the aforementioned study, Courtenay et al. (2020c) define grazes to be short ($\approx$3.75mm) and wide ($\approx$0.35mm) naturally produced abrasions on the surface of bone, commonly produced when the bone is dry. Scratches, on the other hand, are long ($\approx$5.67mm) and thin ($\approx$0.22mm) marks that are more commonly produced when the bone is fresh.

The present demonstration took the measurements of 251 trampling marks provided in the supplementary materials of Courtenay et al. (2020c), and augmented them (x1000) using a Markov Chain Monte Carlo algorithm as described by Courtenay et al. (Under Review-a). A simple 4 hidden layer neural network was then trained on this data. This hidden layer consists of 2 input neurons (length & width), followed by a set of hidden neurons ($n = [15,10,5,3]$), before being mapped out to a single output neuron for binary classification (scratch (1) or graze (0)). All hidden layers used ReLU activation functions (eq. D.5), while the output layer was activated using the Sigmoid function (eq. D.7). Dropout layers were inserted in between each of the layers with $\vec{p} = [0.1,0.1,0.1,0.1]$ (Srivastava, 2013), while a kernel $l2 = 1\times10^{-05}$ regulariser was used (Krogh and Hertz, 1991). The network was trained for 1000 epochs with a batch size of 128. Loss was calculated using the binary crossentropy function (Bishop, 1995, 2006), while back propagation was performed using the RMSprop optimizer (Kochenderfer and Wheeler, 2019). Training was performed using a 70:30 *Training:Validation* split, with 30% of the sample separated prior to this for testing, as is typical in Deep Learning common practice (Goodfellow et al., 2016).

A visualisation of the neural network, and representation of the training results, have been detailed in the Figure D.1 and Table D.1;

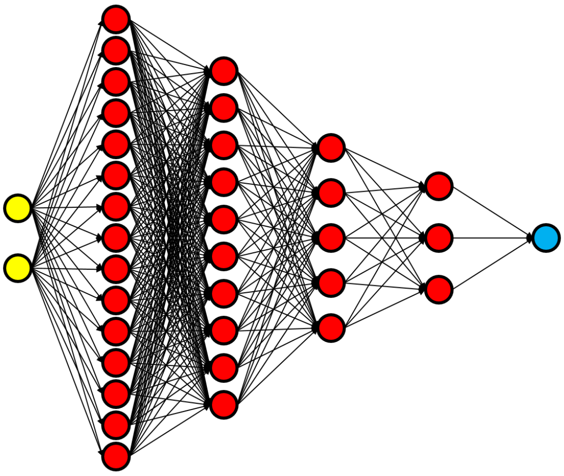


***Figure D.1:*** *Visualisation of the Neural Network used in this demonstration. Yellow neurons represent input, blue output with sigmoid activation, red neurons represent hidden neurons. Note how no Bias neurons have been included in this visualisation.*

Table D.1: *Evaluation matrices for the training of a neural network on the Courtenay et al. (2020c) dataset. AUC = Area Under the receiver operator characteristic Curve*

| Metric | |
|---|---|
| Training Accuracy | 0.94 |
| Validation Accuracy | 0.97 |
| Training Loss | 0.16 |
| Validation Loss | 0.12 |
| Test Accuracy | 0.97 |
| Test Loss | 0.13 |
| Test AUC | 0.99 |
| Test Sensitivity | 0.95 |
| Test Specificity | 0.98 |
| Test Kappa | 0.93 |

## D.2 Inside the Black Box

Figure D.2 presents the results for this demonstration, visualising how certain regions of the network activate or disactivate depending on the type of information they are provided with as input. The visualisation of activation frequency was performed using a balanced in-out degree, typical of applications in graph theory. Both the in and out-degrees were calculated, and then normalised according to all the possible in or out connections they may have. Red neurons indicate a high balanced in-out degree, while white neurons represent a low balanced in-out degree. The edges of the graph (weighted connections) are also visualised according to the frequency with which information passes through them.

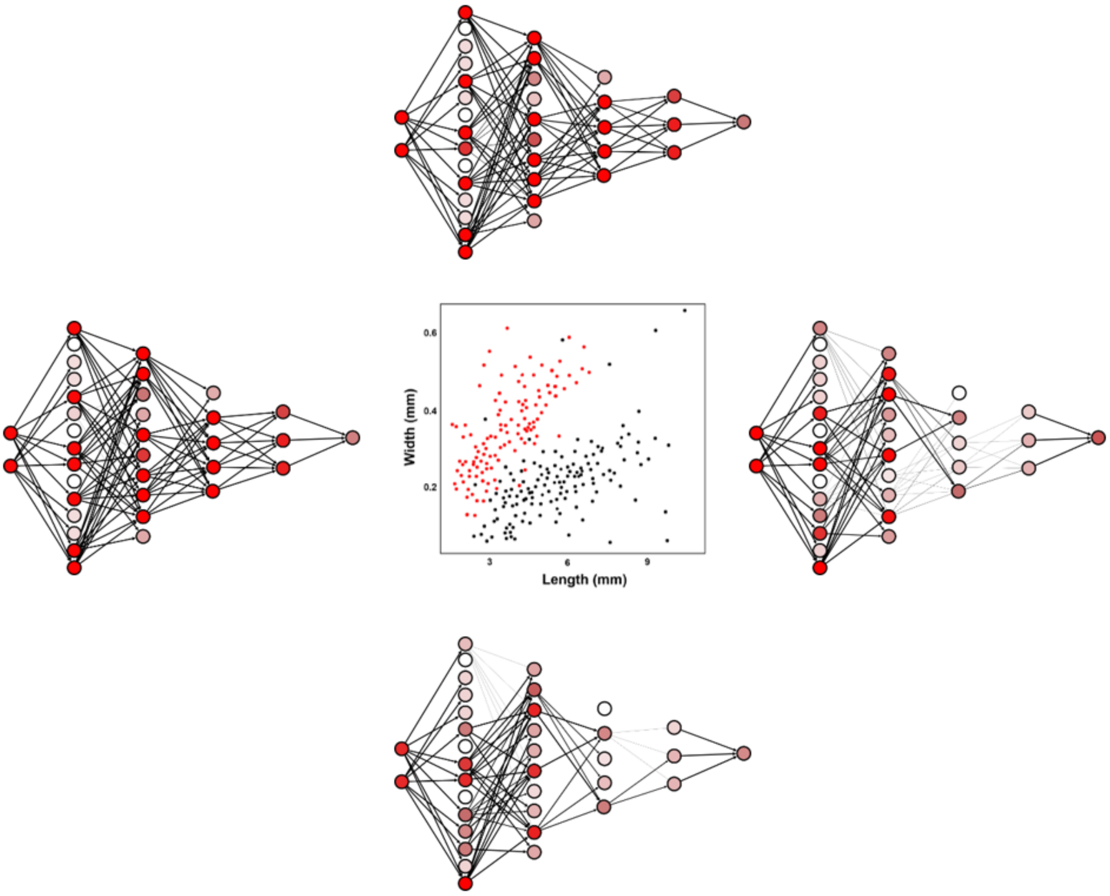


***Figure D.2:*** *Visualisation of the balanced degree of each neuron in the Neural Network used for this demonstration. Neural Network visualisations are presented at the extremity of each axis of the graph, defining the type of input provided to them. Red neurons are those that are activated the most, white neurons are activated the least. Edge darkness indicates how often information passes through them.*

## D.3 How should Computational Learning algorithms be evaluated?

The evaluation of classification algorithms in computational learning is of the upmost importance, however can often be considered misleading. In most contexts, the words "80% classification" would be seen as a powerful algorithm, however this is often not the case. In contexts such as medicine, for the construction of diagnostic tools, more weight is given to evaluation metrics such as sensitivity and specificity. In other cases, such as economics and supervised anomaly detection in bank fraud, precision and recall are more reliable. In general, precision/recall values are considered metrics to be used for imbalanced classification, while sensitivity and specificity are more sensitive to imbalance, thus being more reliable in general classification problems (He and Ma, 2013). The term imbalance here refers to where one of the labels we wish to classify is underrepresented in the dataset.

We can calculate the balance ($B$) of a dataset using an adaptation of Shannon's entropy theory (Shannon, 1948a,b);

$$B = \frac{-\sum_{i=1}^{g} \frac{c_i}{n} \log \frac{c_i}{n}}{\log g} \tag{D.10}$$

where $c$ is the count of individuals in group $g$, and the sum of all values in vector $c$ is $n$ (i.e. the total sample size). The $B$ index is bound between $[0,1]$, with values of 1 indicating balance and values of 0 indicating imbalance.

In bank fraud, for example, there may be 100 frauds out of 10,000 examples. In this case, the imbalance is described as 1:100 ($B = 0.08$). In medicine, this would be the case of classifying a rare disease. A balanced problem, on the other hand, would be in the case where we have the same number of healthy individuals as unhealthy individuals in a clinical trial (1:1, $B = 1$). Finally, it is not the same for slight imbalance to exist (1:10, $B = 0.43$), as slightly more extreme cases (1:100, $B = 0.08$), or very extreme cases (1:1000, $B = 0.01$). In each of these cases different evaluation metrics will be more or less reliable.

From this perspective, each of the metrics are defined for a particular reason, however all carry out very similar functions. Evaluation metrics can thus be defined as a means of not only calculating classification accuracy, but also a means of evaluating how often an algorithm is right, wrong, not wrong, or right in saying the opposite. While this can sometimes be confusing, consider the following situation; it is not the same for a doctor to diagnose a patient with an illness and they are actually healthy, as it is to diagnose a patient as healthy and they are in fact ill. In the former case, this could lead to a law suit, while in the latter this could even lead to death.

In order to perform these evaluations, a Confusion Matrix $M$ is defined;

$$M = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

where $P$ refers to positive, $N$ to Negative, $T$ to True and $F$ to False. In the case of the aforementioned patient, a False Positive would be where we diagnose the patient as ill when they are in fact not. A True Positive would be if we diagnose the patient as ill and they are in fact ill.

For simplicity, as well as the purpose of demonstrating how model evaluation works, each of the three matrices that will be tried and tested in this Appendix are defined in the context of binary classification (ill or healthy, fraud or not-fraud). Evaluation metrics for binary classification in this context can thus be defined as;

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{D.11}$$

$$Sensitivity = \frac{TP}{TP+FN} = Recall \tag{D.12}$$

$$Specificity = \frac{TN}{FP+TN} \tag{D.13}$$

$$Precision = \frac{TP}{TP+FP} \tag{D.14}$$

With F-Measure ($F$):

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{D.15}$$

and Kappa ($\kappa$);

$$P_e = P_{pos} + P_{neg} \tag{D.16}$$

$$P_{pos} = \frac{TP+FP}{TP+FP+FN+TN} \cdot \frac{TP+FN}{TP+FP+FN+TN} \tag{D.17}$$

$$P_{neg} = \frac{FN+TN}{TP+FP+FN+TN} \cdot \frac{FP+TN}{TP+FP+FN+TN} \tag{D.18}$$

$$P_o = \frac{TP+TN}{TP+FP+FN+TN} \tag{D.19}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{D.20}$$

For ease of calculation, the Area Under Curve (AUC) of either Receiver Operating Characteristic (ROC), or Precision-Recall curves, were excluded from this appendix. AUC however can be seen as a powerful balance between Sensitivity, Specificity and Precision (Sing et al., 2005).

For the purpose of this demonstration we have defined three confusion matrices; A is an example of a very poor classifier with a misleadingly high accuracy; B is a slightly more evident case of a poor classifier, with relatively high accuracy; and C is a relatively good classifier with a more balanced accuracy. Matrices A and C both have 80% classification accuracy, a threshold that would normally be considered "good". Matrix B, however, has 70% accuracy, a rate that is not particularly good but could be worse.

$$A = \begin{bmatrix} 80 & 20 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 60 & 20 \\ 10 & 10 \end{bmatrix}, \quad C = \begin{bmatrix} 50 & 10 \\ 10 & 30 \end{bmatrix}$$

The evaluation results for all three matrices can be consulted in the following table;

Table D.2: *Evaluation matrices for Confusion Matrices A, B and C.*

| Metric | Matrix A | Matrix B | Matrix C |
|---|---|---|---|
| Accuracy | 0.80 | 0.70 | 0.80 |
| Sensitivity | 1.00 | 0.86 | 0.83 |
| Specificity | 0.00 | 0.33 | 0.75 |
| Precision | 0.80 | 0.75 | 0.83 |
| Recall | 1.00 | 0.86 | 0.83 |
| F-Measure | 0.88 | 0.80 | 0.83 |
| Kappa | 0.00 | 0.21 | 0.58 |

As can be seen in the table, the most obvious observation is that 80% accuracy is still possible even if an algorithm is very weak. In this case, "80% accuracy" is completely unreliable. Likewise, for Matrix B, while the accuracy falls 10% in comparison with other metrics for Matrix A, it is still a much more powerful classifier. This appears increasingly more obvious as Kappa is calculated. If we can consider how the number of True Negatives is 0 in Matrix A, the first fraction in eq. D.18 will equate to 0. A scalar 0 multiplied or divided by anything will condition the formula to reach 0, thus lowering the $\kappa$ statistic drastically. Likewise, considering how Matrix A presents evaluation metrics higher than Matrix B, the low TN value still conditions $\kappa$ to be much lower. Finally, even in Matrix C, where most values (sensitivity, precision and recall) are above the acceptable 0.8 threshold, $\kappa$ still proves to be a strict statistic determining Matrix C to be a moderate classifier (Landis and Koch, 1977; Foody, 2008; Kuhn and Johnson, 2013).

The next interesting component to consider is how, even in matrices that have been specifically designed to be bad (A and B), both the Precision, Recall and F-Measure values are very high. This is quite misleading, and could raise the question: are these really reliable? To simply answer this question, consider the following matrix;

$$D = \begin{bmatrix} 85000 & 20 \\ 30 & 110 \end{bmatrix}$$

Matrix D represents a classification task with an extreme imbalance of $B = 0.016$, typical of anomaly detection. The resulting precision and recall in this case study comes to 0.99976 and 0.99964 respectively. This indicates that the algorithm that produced matrix D was a very powerful classifier for both the smaller and larger class. Specificity, however, would not truly reflect this power with a value of only 0.85.

From this perspective, it can be seen that different evaluation metrics are designed for different purposes, and the choice of metrics in a study *should* reflect this.

Finally, another popular means of evaluating an algorithm is by assessing the model's loss. Loss is defined as the error produced when making predictions. There are two very popular formulae for the classification of loss;

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{D.21}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{D.22}$$

These are known as the Mean Squared Error (MSE) and the square root of MSE (RMSE), respectively. If we consider, for example, a binary classification problem, where an algorithm is designed to detect whether a patient has cancer (1) or not (0), and the algorithm predicts a label of 0.9, then the error for this particular case is $1.0 - 0.9$. Evidently, an optimal algorithm will have the least amount of error possible, however should remain consistent.

In the training of neural networks, measurements of loss are fundamental for the diagnosis of over and underfitting (Fig. D.3). In cases where training and validation loss are high, this can typically be used to identify an underfit NN. When training loss is low and the validation loss is high, this is often a sign of overfitting. When both the training loss and validation loss are low and reach an optimal value at the same time, this is an indication that an algorithm is well trained.

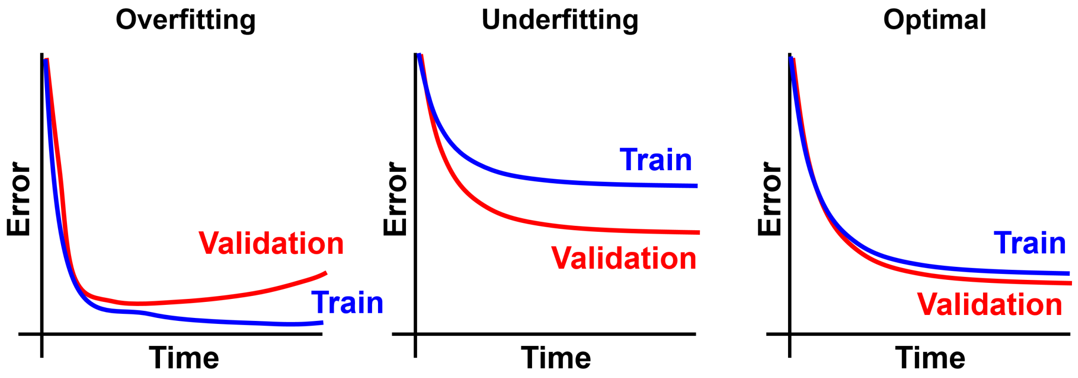


***Figure D.3:*** *Examples of neural network learning curves in different situations.*

# Appendix E

# Software and Computational Resources

All analyses and applications described throughout the present Doctoral Thesis relied heavily on the development and use of multiple computer software, all of which are openly available through multiple repositories on the author's GitHub page (`https://github.com/LACourtenay`), as well as in forks on the TIDOP research group's GitHub page (`https://github.com/TIDOP-USAL`)

For this purpose, the applications presented through this Doctoral Thesis were developed in 4 primary programming languages (in order of relevance); R (v.4.0), Python (v.3.7), JavaScript (v.1.8), and C++ (v.17).

The main contributions included are the development of 3 R libraries for; (1) the calibration of $p$-Values and use of robust statistical tools (Courtenay et al., 2020a, 2021a,b); (2) augmentation of both numerical and categorical data using Monte Carlo based algorithms (Courtenay et al., Under Review-a), as well as the evaluation of this data (Courtenay and González-Aguilera, 2020); and (3) the implementation of Graph-based geometric morphometric tools (Courtenay et al., Under Review-b). In addition to these R libraries, a software was developed in C++ for the processing of linear measurement data extracted from tooth scores, following the work of Courtenay et al. (2021a), that can also be applied to the study of cut marks following the work of other authors, e.g. (Bello and Soligo, 2008). Finally, an additional application for the classification of trampling marks based on width and length measurements was also developed to gain a better understanding on how Neural Networks work and can be implemented into a GUI. For this application, all 4 programming languages were used.

The following subsections of this appendix will describe in detail the different computer software developed.

## E.1 R Libraries

### E.1.1 pValueRobust

The **pValueRobust** library was developed to implement the multiple calculations detailed in Appendix C, so as to provide tools for the calibration and evaluation of $p$-Values (Courtenay et al., 2020a, 2021a,b).

This is a simple library, written purely in the base R programming language without any external dependencies, and is the sole property of L. A. Courtenay (©2022, MIT License). This library was written following the functional programming paradigm, without the use or implementation of any classes. Most functions are simple, with only a single input required (a $p$-Value), while some accept an optional prior probability parameter. Table E.1 describes the functions available in this library for $p$-Value calibrations;

| Function Name | Description |
|---|---|
| BFB | Bayes Factor Bound (BFB) calculation for a given $p$-Value. BFB is the upper bound confidence interval with regards to traditional Bayes Factors, i.e. the highest possible Bayes Factor consistent with the observed $p$-Value. BFBs are generally reported as *at most* the odds against $H_0$ |
| posterior_odds | The posterior odds of BFB, provided a given prior probability. The default prior probability is 0.5. |
| p_BFB | The equivalent of BFB represented as a percentage value. This function can be provided with a prior probability. The default prior is 0.5. |
| FPR | False Positive Risk (FPR) calculation for a given $p$-Value. The FPR is the probability of an observation being a False Positive, or Type I Statistical Error, having accepted $H_a$. This implementation of the formula uses BFB as a means of defining the Likelihood ratio in favour of $H_a$. This function can be provided with a prior probability. The default prior is 0.5. |
| p_H0 | The Probability of the Null-Hypothesis ($H_0$) having accepted $H_a$. This function can be provided with a prior probability. The default prior is 0.5. |
| calibration_curve | A simple function that can be used to calculate callibration curves for BFB values, FPR values, $pH_0$ values, and a comparison of p_BFB and FPR curves. This function accepts prior probabilities. The default prior is 0.5. |

Table E.1: Functions for the calibration and evaluation of $p$-Values. $H_0$ is the Null Hypothesis, while $H_a$ is the alternative hypothesis.

In addition to functions oriented towards the calibration of $p$-Values, the **pValueRobust** library also contains a set of miscellaneous functions for the computation of robust descriptive statistical measures. Each of these functions accept vectors as input. Table E.2 describes each of these functions briefly.

| Function Name | Description |
|---|---|
| median_absolute_deviation | A function for the calculation of the Median Absolute Deviation (MAD) or Normalised MAD (NMAD) for a univariate distribution. By default the NMAD value is returned, using the constant 1.4826 for normalisation. *nmad* = *FALSE* can be used to return the non-normalised MAD value. |
| biweight_midvariance | A function for the calculation of the Biweight Midvariance (BWMV) or Square-Root of the BWMV ($\sqrt{BWMV}$) for a univariate distribution. By default this function calculates $\sqrt{BWMV}$ |
| quantile_CI | A function that calculates either the quantile value of a given univariate distribution, or Confidence Intervals for this distribution. |

Table E.2: Additional functions for the calculation of robust statistical metrics

### E.1.2 AugmentationMC

The **AugmentationMC** library was developed for the implementation of Monte Carlo based algorithms, under the premise that these functions can be used for the augmentation and simulation of both numeric and categorical data (Courtenay and González-Aguilera, 2020; Courtenay et al., Under Review-a). This library is written in the base R programming language, and only contains a single dependency, being the **abind** library for tensor operations. This library is the sole property of L. A. Courtenay (©2022, MIT License).

**AugmentationMC** is written using a mixture of functional and object-oriented programming. For object-oriented programming purposes, a single S4 class was designed and developed, named the **MCMC_Trace** class. The **MCMC_Trace** class is the base of all Markov Chain Monte Carlo algorithms related to this library. Associated with this class are 5 different methods, as well as one generic method that accepts input in the form of vectors and matrices as well. Finally, all implementations of Monte Carlo based algorithms are programmed using functional programming.

The following table describes each of the functions included in the *AugmentationMC*, their input type, class as well as a description of the function itself;

| Function Name | Class | Input Type | Description |
|---|---|---|---|
| numericMC | MC | Numerical | Monte Carlo based algorithm for the simulation of numeric (univariate or multivariate) data. |
| categoricalMC | MC | Categorical | Monte Carlo based algorithm for the simulation of categorical (univariate or multivariate) data. |
| bivariate_multimodalMC | MC | Both | Monte Carlo based algoritm for the simulation of bivariate datasets consisting of one categorical and one numeric variables. |
| conditionalMC | MC | Both | Monte Carlo based algorithm for the conditional simulation of variables. Provided a dataset of one categorical variable and numerous numeric variables, the algorithm will first simulate an instance of the categorical variable, and then calculate the most probable numeric values that are associated with it. Given a single numeric variable, and a set of categorical variables, the algorithm will do the same; first simulate an instance of the numeric variable, and then predict the most probable categorical values associated with it. |
| multimodalMC | MC | Both | Monte Carlo based algorithm for the simulation of both numeric and categorical variables. |
| MCMC | MCMC | Numerical | Metropolis-Hastings 1st Order Variant of the Monte Carlo Markov Chain algorithm. This function produces an S4 class **MCMC_Trace** object containing the trace of the Marov Chain over the number of iterations defined by the user. The user may also define a step size for the Markov chain. |
| burn_in | MCMC | MCMC_Trace | A function used to define the burn-in period of the Markov chain, thus eliminating a percentage of the chain as defined by the user. |
| trace_plot | MCMC | MCMC_Trace | A visualisation of the MCMC trace. |
| autocorrelation | MCMC | MCMC_Trace | A plot of the autocorrelation function, or correlogram, that defines the serial correlation of a MCMC trace. Powerful chains present correlogram curves that approximate 0 quickly. |
| effective_sample_size | MCMC | MCMC_Trace | A function used to calculate the effective sample size for mean estimation from a trace. The smaller the ESS value, the poorer the chain. For univariate chains, values above 1000 are recommendable, for multivariate chains values above 100 are recomendable per dimension. |

| Function Name | Class | Input Type | Description |
|---|---|---|---|
| sample_from_trace | MCMC | MCMC_Trace | A function used to sample the final augmented dataset from the **MCMC_trace**. The user must define the sample size they wish to obtain, and the algorithm will extract *n* number of observations from the posterior distribution. |
| TOST | Both | Numerical | A "Two One-Sided Test" of equivalency to compare a simulated dataset and the original dataset. The **robust** parameter allows the user to define whether Welch's *t*-statistic or Yuen's robust *t*-statistic will be used for homogeneous and non-homogeneous distributions respectively. |

Table E.3: A detailed description of the functions from the **AugmentationMC** library. MC = Monte Carlo based algorithm. MCMC = Markov Chain Monte Carlo

### E.1.3 GraphGMM

The **GraphGMM** library was developed for multiple purposes oriented towards the field of Geometric Morphometrics, namely the integration of Graph-based tools, following the mathematical models proposed by Courtenay et al. (Under Review-b). This library is written in the R programming language, however has multiple dependencies, including;

- **shapes**. A geometric morphometric library for some basic geometric morphometric functions.
- **Morpho**. A geometric morphometric library for some basic geometric morphometric functions.
- **igraph**. A library for network analyses and graph theory.
- **abind**. A library for tensor operations
- **RTriangle**. A library with functions for point cloud triangulation.
- **ggplot2**. A visualisation library.
- **philentropy**. A library for distance calculations.
- **Rtsne**. A library for t-Distributed Stochastic Neighbour Embeddings
- **rgl**. A library for 3D visualisation.
- **dplyr**. A library for database management.
- **sm**. A library for the visualisation of probability distributions.
- **robustbase**. A library with computationally efficient implementations of a colMedian function, as opposed to colMeans.
- **MASS**. Statistical library and toolset.
- **Rvcg**. A library with functions for the triangulation of point clouds.
- **RColorBrewer**. A miscellaneous library with functions for the creation of colour palattes.
- **car**. Statistical library and toolset.
- **magrittr**. A library containing useful tools for handling data in R, especially the pipeline %>% operator.

This library is the sole property of L. A. Courtenay (©2022, MIT License).

**GraphGMM** is written using a functional programming approach, and contains 4 main families of functions. The first group of functions consists in functions for basic Geometric Morphometric analyses (Table E.4). These include an implementation of Generalised Procrustes Analyses, with ability to perform Generalised Robust Procrustes Analyses, as well as functions for Generalised Resistant Fit, Relative Warp Analyses, and other common tools in Geometric Morphometric research.

The second family includes functions directly associated with Graph-based Geometric Morphometrics (Table E.5), including the graph embedding function, as well as a set of functions for the computation of graphs based on the spatial distribution of landmarks in the configuration. This set of functions also includes some miscellaneous and experimental functions for in-depth analyses of a landmark graph.

The third family of functions are oriented towards dimensionality reduction (Table E.6), including the implementation of Principal Component Analyses functions that plot not only PCA feature space but also PCA-biplots, as well as functions for non-linear dimensionality reduction, including the t-Distributed Stochastic Neighbour Embedding algorithm, and non-linear kernel PCA algorithms.

The final family are a simple set of miscellaneous functions (Table E.7).

Future improvements and updates to the **GraphGMM** library have the intention of improving the way the library is designed, including an update from a functional programming paradigm to an object-oriented paradigm with S3 and S4 classes.

A detailed description and manual of the **GraphGMM** library was included in the supplementary materials of Courtenay et al. (Under Review-b). We strongly recommend consulting these documents.

| Function Name | Input Type | Description |
|---|---|---|
| GPA | LM Array | A function that can be used to perform a Generalised Procrustes Fit (GPA). This function is adaptable for both shape and form analyses, with additional functions for robust implementations of the algorithm. Using *proc_method = "LS"* and *robust = "TRUE"*, this function can also be used to perform a Generalised Robust Fit. The *proc_method = "optimalLS"* performs GPA according to Rohlf and Slice (1990) and Dryden and Mardia (2016) |
| generalised_resistant_fit | LM Array | A function that can be used to perform a Generalised Resistant Fit, following Siegel and Benson (1982) and Slice (1996). This function is adaptable for both shape and form analyses, with additional functions for robust implementations of the algorithm. |
| relative_warp_analysis | Proc. LM Array | A functon for the calculation of Relative Warp Analyses, and their corresponding Principal Component Analysis plots. |
| orthogonal_projection | Kendall Coords. | A function for the orthogonal projection of superimposed coordinates from a non-Euclidean space to their Euclidean tangent space. |
| stereographic_projection | Kendall Coords. | A function for the stereographic projection of superimposed coordinates from a non-Euclidean space to their Euclidean tangent space. |
| calc_central_morph | Proc. LM Array | A function that can be used to calculate the mean or median landmark configuration from a landmark array |
| mahalanobis_dist_matrix | PC Scores | A function for the calculation of inter- and intra-group mahalanobis distances and $p$-Values given a set of PC Scores |
| procD_comparisons | Proc. LM Array | A function for the calculation of inter- and intra-group Procrustes distance $p$-Values |
| morphological_predictor | Proc. LM Array & PC Scores | A function used for the prediction of landmark coordinates using Least Square Regression given a set of morphological variables. |
| tps_visualisation | Proc. LM Array & PC Scores | A function for the visualisation of shape deformations across a given feature space using Least Square Regression and Thin Plate Splines |

Table E.4: A detailed description of the functions from the **GraphGMM** library designed for basic Geometric Morphometric analyses. Proc. LM Array = Procrustes superimposed landmark array. LM Array = Non-superimposed landmark array. Kendall Coords. = Kendall Shape Coordinates

| Function Name | Input Type | Description |
|---|---|---|
| graph_embeddings | Proc. LM Array & Edges | The primary function of the **GraphGMM** library, used for the computation of landmark embeddings according to the edges connecting each landmark. |
| similarity_matrix | Proc. LM Array | A function that can be used to calculate the distance between landmark pairs, either using the cosine, procrustes, or Chevyshev distance. |
| knn_graph | $\bar{X}$ | A function to compute the spatial relationships between landmarks in a given configuration using the KNN nearest neighbour algorithm. |
| triangulate2d | $\bar{X}$ in 2D | A function to compute the spatial relationships between landmarks in a given configuration using the Delaunay 2D and 2.5D triangulation algorithm. |
| triangulate3d | $\bar{X}$ in 3D | A function to compute the spatial relationships between landmarks in a given configuration using the ball-pivoting triangulation algorithm. |
| as_edge_list | $2 \times 2$ Matrix | A simple function to convert computed edges into an *edge_list* for graph computations. |
| plot_landmark_graph | $\bar{X}$ & Edges | A function that can be used to plot and visualise a landmark graph. |
| graph_configuration_statistics | $\bar{X}$ & Edges | A function that can be used to calculate graph properties, including; the clustering coefficient, graph density, landmark degree centrality, eigenvector centrality, and betweenness centrality. |
| plot_landmark_stats | $\bar{X}$ & Edges | A function that can be used to plot and visualise a landmark graph incorporating the statistical data provided by the *graph_configuration_statistics* function. |
| create_landmark_graph | $\bar{X}$ & Edges | A function that can be used to create an **igraph** object for more detailed analyses with the **igraph** library |
| landmark_modularity | $\bar{X}$ & Edges | A function that can be used to calculate landmark modules based on the properties of the landmark graph |

Table E.5: A detailed description of the functions from the **GraphGMM** library designed for basic Graph Embeddings and the analysis of Landmark Graphs. Proc. LM Array = Procrustes superimposed landmark array. LM Array = Non-superimposed landmark array. $\bar{X}$ = Central configuration for a landmark array.

| Function Name | Input Type | Description |
|---|---|---|
| pca_plot | A dataset | A function used to perform Principal Component Analyses |
| pca_biplot | A dataset | A function used to perform Principal Component Analyses and plot the PCA bi-plot |
| calculate_optimal_pc_scores | A dataset | A function that permutes the number of Principal Components to calculate the optimal number of PC Scores required for statistical analyses |
| kernel_pca | Matrix | A function to perform a non-linear Principal Component Analysis based on the specified kernel function to be used. |
| kernel_pca_biplot | Matrix | A function to perform a non-linear Principal Component Analysis, based on the specified kernel function to be used, and plot the PCA bi-plot |
| kernel_cauchy | Matrix | A function for the non-linear transformation of data prior to dimensionality reduction, using the Cauchy kernel. |
| kernel_gaussian | Matrix | A function for the non-linear transformation of data prior to dimensionality reduction, using the Gaussian kernel. |
| kernel_laplace | Matrix | A function for the non-linear transformation of data prior to dimensionality reduction, using the Laplacian kernel. |
| kernel_poly | Matrix | A function for the non-linear transformation of data prior to dimensionality reduction, using the Polynomial kernel. |
| kernel_rbf | Matrix | A function for the non-linear transformation of data prior to dimensionality reduction, using the Radial Basis Function kernel. |
| kernel_spline | Matrix | A function for the non-linear transformation of data prior to dimensionality reduction, using the Spline kernel. |
| tsne_plot | Matrix | A function for the computation of t-Distributed Stochastic Neighbour Embeddings for a given dataset as a means for non-linear dimensionality reduction. |

Table E.6: A detailed description of the functions from the **GraphGMM** library designed for Dimensionality Reduction. Proc. LM Array = Procrustes superimposed landmark array.

| Function Name | Input Type | Description |
|---|---|---|
| separate_samples | LM Array | A function that can be used to separate landmark samples into separate groups. |
| vector_from_landmarks | Proc. LM Array | A function for the flattening of a tensor containing landmark coordinates. |
| write_morphologika_file | LM Array & labels | A function for the creation of a morphologika file given a set of landmarks and the labels describing what samples they come from. |

Table E.7: A description of the miscellaneous functions from the **GraphGMM** library. Proc. LM Array = Procrustes superimposed landmark array. LM Array = Non-superimposed landmark array.

## E.2 TPS Measurement Software

The TPS Measurement Software, available from `https://github.com/LACourtenay/Measurement_TPS_GUI`, is an application written in C++, using the Qt (v.6.0.4) framework for the Graphical User Interface (GUI). This application was written and compiled using the Microsoft Visual Studio integrated development environment (2019 release), in Windows 10. The creation of an installer for this application was performed using the Inno Setup Compiler (v.6.2.0) software.

This software was created for the purpose of loading a folder containing multiple .tps files, each of which contain 7 landmark coordinates extracted from a bone surface modification profile. The software proceeds to extract the coordinate data, scale this data, and either (1) calculate the measurements needed for further processing, or format these coordinates into a morphologika file. Importantly, the TPS measurement software takes into consideration the mathematical and statistical properties described by Courtenay et al. (2021a), with regards to the variable "opening angle". From this perspective, the application calculates not only the opening angle in degrees, but also performs a linear transformation of this variable for multivariate statistical purposes.

This software was programmed using both functional and object-oriented programming paradigms. The *Landmark.h* header file contains a simple structure, or “struct”, defining the properties of a single landmark in a configuration. This struct additionaly has a static *scale* attribute that can be used to scale the entire configuration at once.

This application is the sole property of L. A. Courtenay (©2021, MIT License).

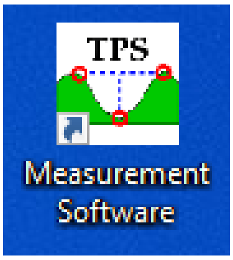
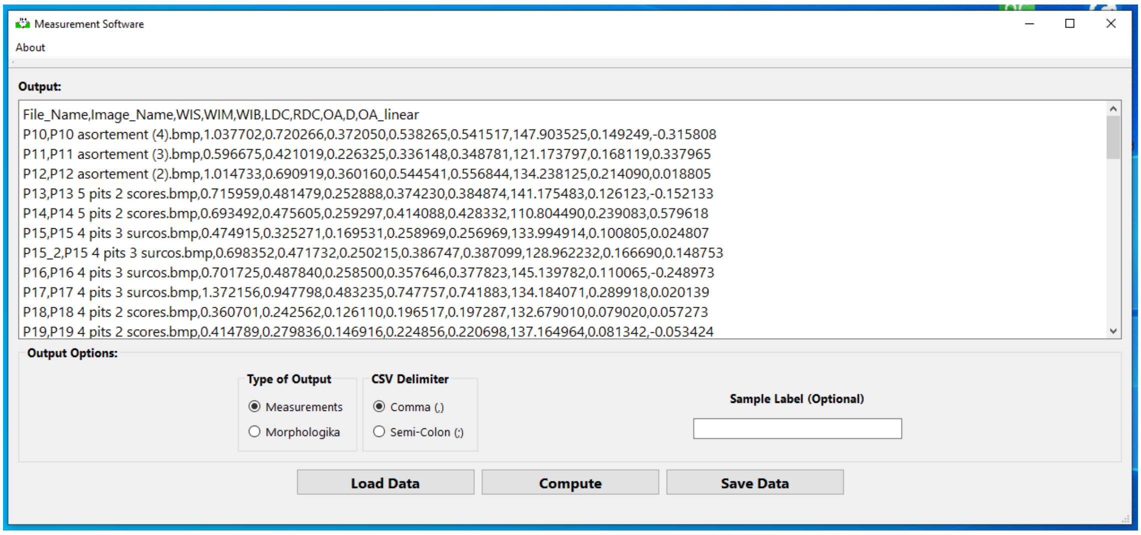


***Figure E.1:*** *TPS Measurement Software Icon*



```
File_Name,Image_Name,WIS,WIM,WIB,LDC,RDC,OA,D,OA_linear
P10,P10 asortement (4).bmp,1.037702,0.720266,0.372050,0.538265,0.541517,147.903525,0.149249,-0.315808
P11,P11 asortement (3).bmp,0.596675,0.421019,0.226325,0.336148,0.348781,121.173797,0.168119,0.337965
P12,P12 asortement (2).bmp,1.014733,0.690919,0.360160,0.544541,0.556844,134.238125,0.214090,0.018805
P13,P13 5 pits 2 scores.bmp,0.715959,0.481479,0.252888,0.374230,0.384874,141.175483,0.126123,-0.152133
P14,P14 5 pits 2 scores.bmp,0.693492,0.475605,0.259297,0.414088,0.428332,110.804490,0.239083,0.579618
P15,P15 4 pits 3 surcos.bmp,0.474915,0.325271,0.169531,0.258969,0.256969,133.994914,0.100805,0.024807
P15_2,P15 4 pits 3 surcos.bmp,0.698352,0.471732,0.250215,0.386747,0.387099,128.962232,0.166690,0.148753
P16,P16 4 pits 3 surcos.bmp,0.701725,0.487840,0.258500,0.357646,0.377823,145.139782,0.110065,-0.248973
P17,P17 4 pits 3 surcos.bmp,1.372156,0.947798,0.483235,0.747757,0.741883,134.184071,0.289918,0.020139
P18,P18 4 pits 2 scores.bmp,0.360701,0.242562,0.126110,0.196517,0.197287,132.679010,0.079020,0.057273
P19,P19 4 pits 2 scores.bmp,0.414789,0.279836,0.146916,0.224856,0.220698,137.164964,0.081342,-0.053424
```

Type of Output: ● Measurements ○ Morphologika

CSV Delimiter: ● Comma (,) ○ Semi-Colon (;)

*Figure E.2: The Graphical User Interface of the TPS Measurement Software*

## E.3 Trampling Algorithm

The Trampling Algorithm is a simple software implemented in 4 different programming languages; R, Python, JavaScript, and C++. This software is available from https://github.com/LACourtenay/Trampling_Algorithm. The purpose of this software is to implement a simple Neural Network, trained as described in Appendix D. Once trained, the weights of the Neural Network were exctracted and programmed into a GUI for ease of use. Additional analyses observed how the internal functioning of the Neural Network works so as to have a better understanding on how they can be implemented and improved in future. This toy example is simply to show how Neural Networks can be trained and inserted into a GUI for use by a non-specialist.

The Trampling Algorithm was originally trained and designed in the Python programming language (v.3.7), using the TensorFlow (v.2.0) library. Once the algorithm was trained, a GUI was designed and implemented using the PyQt5 framework, while all linear algebra calculations were performed using the Numpy (v.1.17.3) library.

For the R release, the neural network was programmed in the base R programming language with no use of dependencies or libraries. Due to restrictions with R, no GUI framework was implemented.

The JavaScript release used HTML and CSS for the creation and design of the GUI, with JavaScript underneath for linear algebra applications.

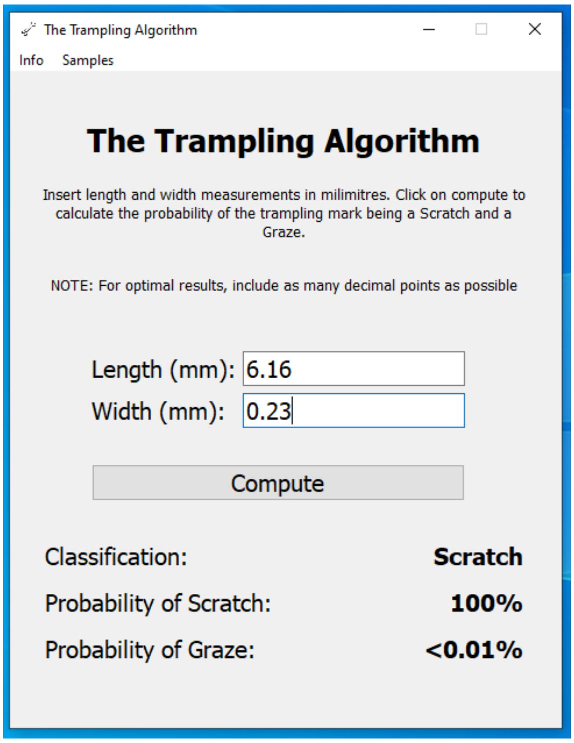
Finally, the C++ release was written using the Eigen library for linear algebra applications, and the Qt (v.6.0.4) framework for the GUI. Code was compiled originally using Qt Creator (v.4.14.2) in windows. The creation of an installer for this application was performed using the Inno Setup Compiler (v.6.2.0).

All 4 releases of this algorithm were designed and written using a functional programming approach, and only used object-oriented programming where necessary for the creation of GUIs (namely when working with the Qt framework).

The advantages and disadvantages of each of these releases are simple; C++ in general is much faster than any of the other languages, however if the user wishes to run this application on an apple product then a macOS specific compilation may be necessary. Python and R are much slower, while the Python implementation also requires the installation of additional libraries (Numpy and PyQt5) in order to work.

Finally, the release using JavaScript is likely to be the easiest to use regardless of the user's operating system, yet is not as fast as C++.

***Figure E.3:*** *Trampling Algorithm Software Icon*



The Trampling Algorithm

Insert length and width measurements in milimitres. Click on compute to calculate the probability of the trampling mark being a Scratch and a Graze.

NOTE: For optimal results, include as many decimal points as possible

Length (mm): 6.16

Width (mm): 0.23

Classification: **Scratch**

Probability of Scratch: **100%**

Probability of Graze: **<0.01%**

***Figure E.4:*** *The Graphical User Interface of the Trampling Algorithm Software*

## E.4 Additional Applications

Wherever applicable, all code related to each of the articles have been published as supplementary materials to the article, or directly linked to the author's GitHub page (`https://github.com/LACourtenay`), ensuring that the methodology presented in each of the scientific articles is as easy as possible to implement.

For Courtenay et al. (2020a), all code used to perform statistical analyses have been included at `https://github.com/LACourtenay/GMM_Measurement_Accuracy_Tools`. This mostly consists of R code for morphometric and statistical analyses, however, also includes Python code for unsupervised pattern recognition applications. In addition, a video tutorial on how to replicate the landmark model and methodological approach were also included in association with this article at https://vimeo.com/409256777.

In the case of Courtenay and González-Aguilera (2020), Python code for all 4 types of Generative Adversarial Networks have been included at https://github.com/LACourtenay/GMM_Generative_Adversarial_Networks.

Both Courtenay et al. (2021b) and Courtenay et al. (2023) use the same code, available at; https://github.com/LACourtenay/Carnivore_Tooth_Pit_Classification. This repository contains R code for Support Vector Machine applications, Python code for Neural Support Vector Machines, and also contains JavaScript, HTML and CSS code for the purpose of some data visualisations. In the case of Courtenay et al. (2023), this study used the earliest versions of the **AugmentationMC** library as well. Likewise, the python code used to prepare, train and implement neural networks from Courtenay et al. (2020b) is available at https://github.com/LACourtenay/Deep-Neural-Network-for-Cut-Mark-Classification.

As for (Courtenay et al., 2021a, Under Review-a, b), the software used for each of these studies has already been described above.

Finally Courtenay et al. (2022a) includes the code used for Elliptic Fourier Analyses as supplementary files.

## E.5 Computational Resources

Three different computer systems were used and experimented with throughout the course of this Doctoral Thesis. The main computer system used was a Dell Precision T1700 desktop computer located in the TIDOP lab of the Polytechnic School of Ávila, University of Salamanca. This computer is equipped with an Intel Xeon E3-1240 v3 CPU processor (4 cores, 3.40 GHz operating Frequency), 8GB of RAM, and a single NVIDIA Quadro K600 GPU. The second system consists in a ASUS X550VX portable laptop, equipped with a Intel Core i5-6300HQ CPU processor (4 cores, 2.30 GHz), 8 GB of RAM, and a NVIDIA GTX 950 GPU.

Finally, access to a supercomputational system was provided by the Supercomputation Center of Castilla y León (SCAYLE). SCAYLE provides services to researchers throughout the autonomous community, and is accessed remotely using a remote computer software and the Secure Shell SSH cryptographic network protocol. For this study the MobaXterm software was used, accessing a server in SCAYLE with a Broadwell architecture. The Broadwell architecture is specifically designed for Computational Learning processes, while SCAYLE provides access to 2 Intel Xeon E5-2695 v4 CPU processors (18 cores each, 2.10 GHz), 384 GB of RAM, and 8 NVidia Tesla v100 GPUs.

# Appendix F

# Scientific Impact

This doctoral thesis presents a total of nine scientific publications, seven published and two under review, in a mixture of different indexed journals across multiple fields of science. The following Appendix describes in detail the impact each of the journals have, sourcing the statistics about a journal's performance from the Journal Citation Reports (JCR) database(`jcr.clarivate.com`), released by Clarivate Analytics, and integrated into the Web of Science. This is an international and prestigious database that contextualises journals from all fields of science. The citation metrics reported throughout this Appendix take into account the status of the journal on the date that each of the articles were published. JCR is annually updated and released in June.

## F.1 Journal Impact Metrics

A journal's performance is typically measured by its Journal Impact Factor (JIF), the percentile and quartile which the journal is ranked in their respected field, as well as the Journal Citation Indicator (JCI);

- The JIF is a measure of the volume of publications released by the journal in relation to the number of citations these publications receive. This metric can be calculated by dividing the number of citations in a year by the total number of articles published in the two previous years. A JIF of 1 thus implies that articles published in the present year have been cited an average of one time. A corrected version of this metric can also be calculated by removing self-citations.
- Using JIF, journals are ranked in their respected fields according to percentiles, which are then used to establish quartiles (Q numbers). A journal ranking as the top 15th journal in a field containing 100 journals will thus have a percentile of 85.0.
- Quartiles describe a journal's position in a field in terms of four different groups; Q1 journals are those that rank in the top 75th percentile of a field, Q2 journals rank between the top 75th and 50th percentile, Q3 journals rank between the top 50th and 25th percentiles, while Q4 journals are those journals that fall into the top 25th percentile of a field. Additionally, a journal can be ranked D1 if it falls in the top 10 percent.
- The JCI is a measure of a journal's citation impact that is normalised to consider the work published by a journal over a recent three year period in a particular field. An average JCI for a particualr field is 1, with a JCI of 1.5, for example, indicating that the journal has 50% more impact than the average.

Beyond these metrics, a journal can be assessed by considering the overall volume of citations over time. This can be performed by first analysing how many citations a journal receives a year, while a distribution graph can also be computed to assess the frequency of publications in a year. A citation distribution graph thus details the population of articles that are cited in a year and thus helps assess the contribution of citations to a JIF. High ranking journals are more likely to present an even distribution across the graph, implying all articles to have the same weight on the journals JIF. More skewed distributions indicate a handful of articles proving highly influential.

## F.2 Scientific Publications and Journals

Table F.1: *Summary of articles published in the present Doctoral Thesis in order of their appearance in the main text. Number of citations are documented using the Scopus database (*`https://www.scopus.com/`*, as of the 13th of December, 2022)*

| | Journal | Month | Year | Open-Access | JIF | Percentile | Q | Citations |
|---|---|---|---|---|---|---|---|---|
| 1 | PLoS ONE | October | 2020 | Yes | 3.240 | 64.58 | Q2 | 12 |
| 2 | Animals | August | 2021 | Yes | 3.231 | 79.84 | Q1 | 4 |
| 3 | Appl. Sci. | December | 2020 | Yes | 2.679 | 58.33 | Q2 | 3 |
| 4 | Under Review | | | | | | | |
| 5 | Appl. Sci. | December | 2019 | Yes | 2.474 | 65.38 | Q2 | 16 |
| 6 | Sci. Rep. | May | 2021 | Yes | 4.380 | 77.08 | Q1 | 13 |
| 7 | Quat. Sci. Rev. | December | 2022 | No | 4.596 | 76.37 | Q1 | 0 |
| 8 | J. Clin. Med. | August | 2022 | Yes | 4.964 | 68.90 | Q1 | 0 |
| 9 | Under Review | | | | | | | |

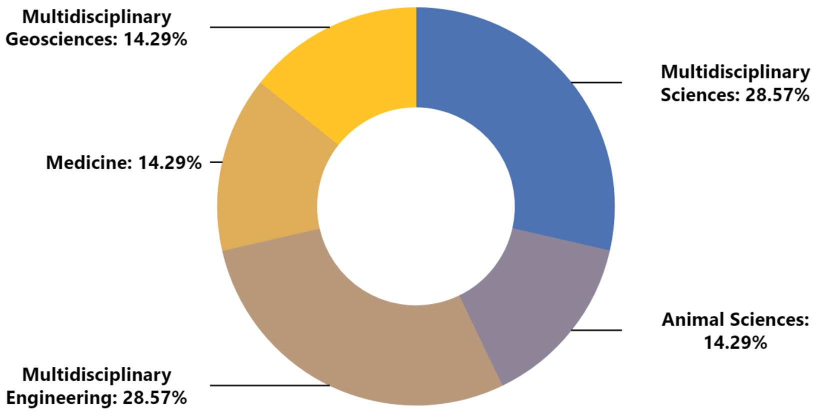


***Figure F.1:*** *Doughnut diagram presenting the different fields of science where articles were published throughout this Doctoral Thesis.*

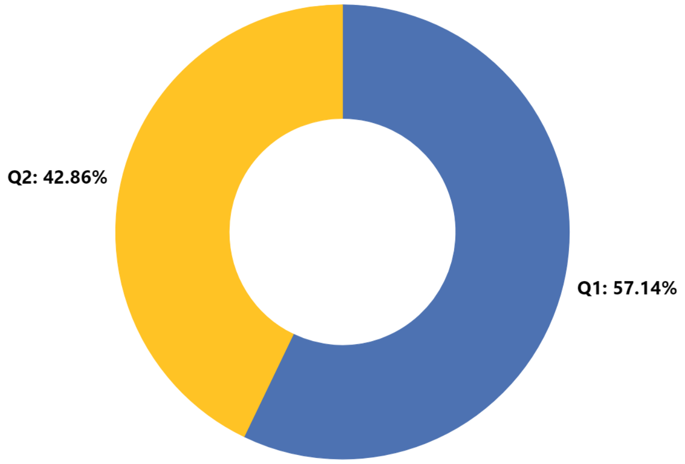


***Figure F.2:*** *Doughnut diagram presenting the quartile ranking of each journal where articles were published throughout this Doctoral Thesis.*

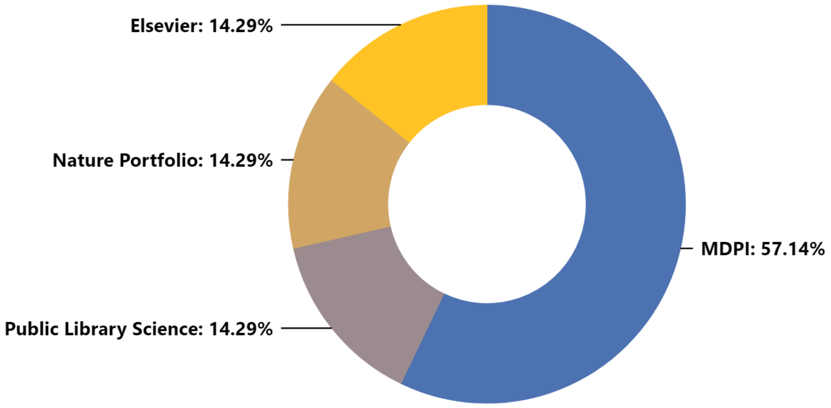


***Figure F.3:*** *Doughnut diagram presenting the editorials where each of the articles were published throughout this Doctoral Thesis.*

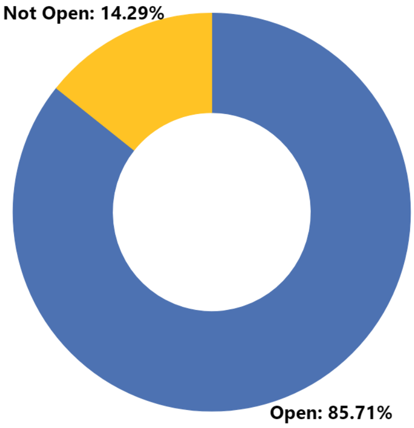


***Figure F.4:*** *Doughnut diagram presenting the proportion of scientific articles that were published as Open-Access for this Doctoral Thesis.*

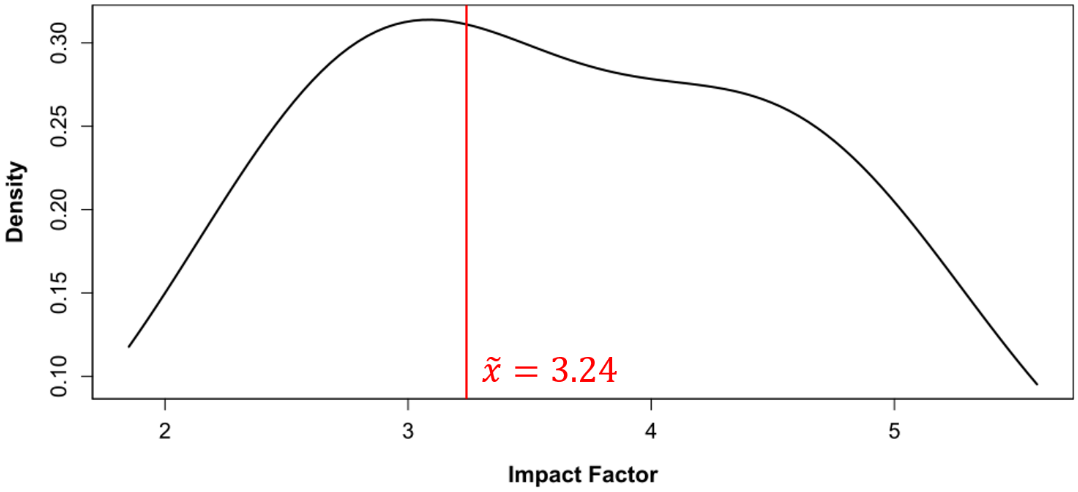


***Figure F.5:*** *Density plot of the JIF metrics associated to each of the articles published in this Doctoral Thesis, resulting in a median Impact Factor of 3.24*

## F.2.1 Animals

*Animals* is an open-access journal affiliated with the World Association of Zoos and Aquariams and the European College of Animal Welfare and Behavioural Medicine. The journal's scope focuses on research that; *offers substantial new insights into any field of study that involves animals, including zoology, ethnozoology, animal science, animal ethics and animal welfare*.

| | |
|---|---|
| Abbreviation | Animals |
| Editorial | MDPI |
| Address | St Alban-Anlage 66, CH-4052 Basel, Switzerland |
| First Index Year | 2018 |
| Current Fields (Rank, Quartile) | Agriculture, Dairy & Animal Science (13 / 62, Q1) |
| | Veterinary Sciences (16 / 144, Q1) |
| Current JIF | 3.231 |
| Current JIF (no self-citations) | 2.551 |
| Current JCI | 1.34 |
| Nº PhD Publications | 1 |

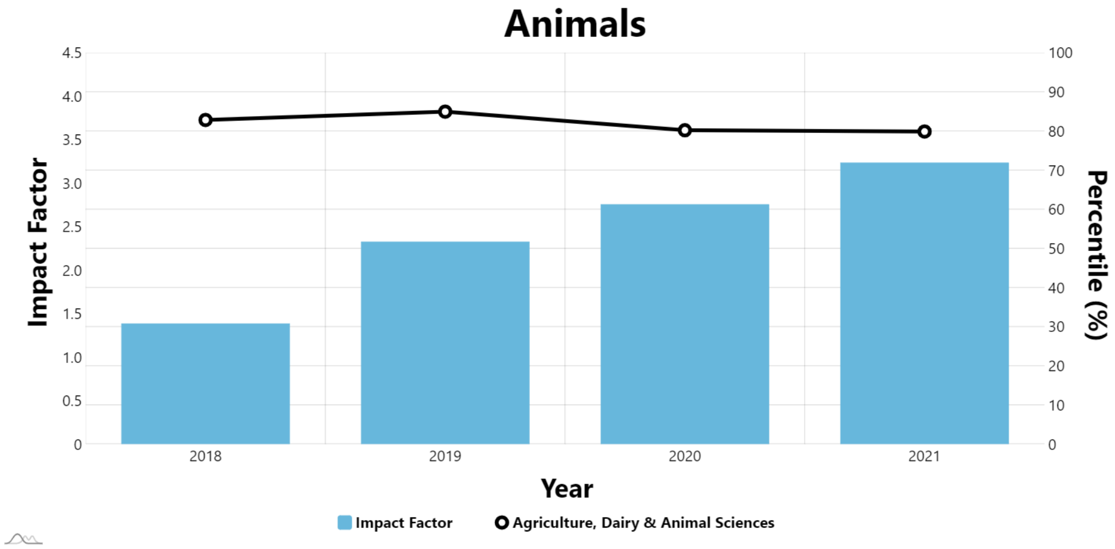


***Figure F.6:*** *Journal Impact Factor and percentile ranking of the MDPI journal "Animals".*

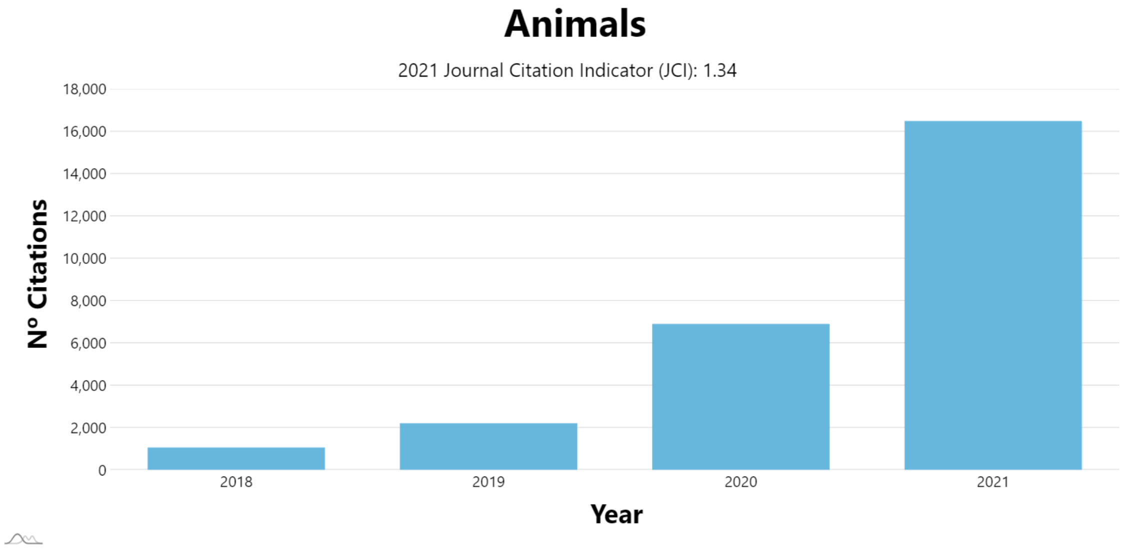


***Figure F.7:*** *Number of yearly citations for the MDPI journal "Animals".*

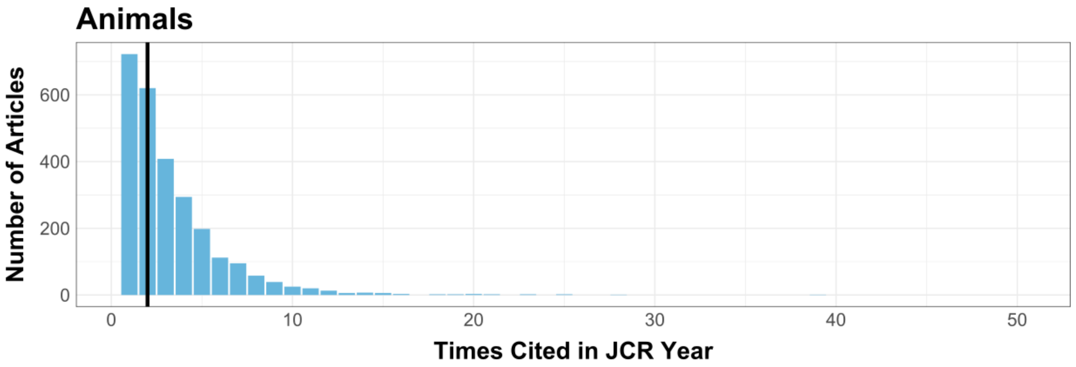


***Figure F.8:*** *Citation Distribution graph describing the number of articles that are cited in a year in the MDPI journal "Animals". The median article citation is marked by the vertical black line ($\tilde{x} = 2$)*

## F.2.2 Applied Sciences

*Applied Sciences* is an open-access journal whose scope focuses on research that focuses on all aspects of applied natural sciences, with the additional aim of; *encouraging scientists to publish their experimental and theoretical results in as much detail as possible*.

| | |
|---|---|
| Abbreviation | Appl. Sci. |
| Editorial | MDPI |
| Address | St Alban-Anlage 66, CH-4052 Basel, Switzerland |
| First Index Year | 2014 |
| Current Fields (Rank, Quartile) | Multidisciplinary Engineering (39 / 92, Q2)<br>Multidisciplinary Chemistry (100 / 179, Q3)<br>Multidisciplinary Materials Science (218 / 345, Q3)<br>Applied Physics (76 / 161, Q2) |
| Current JIF | 2.838 |
| Current JIF (no self-citations) | 2.468 |
| Current JCI | 0.59 |
| Nº PhD Publications | 2 |

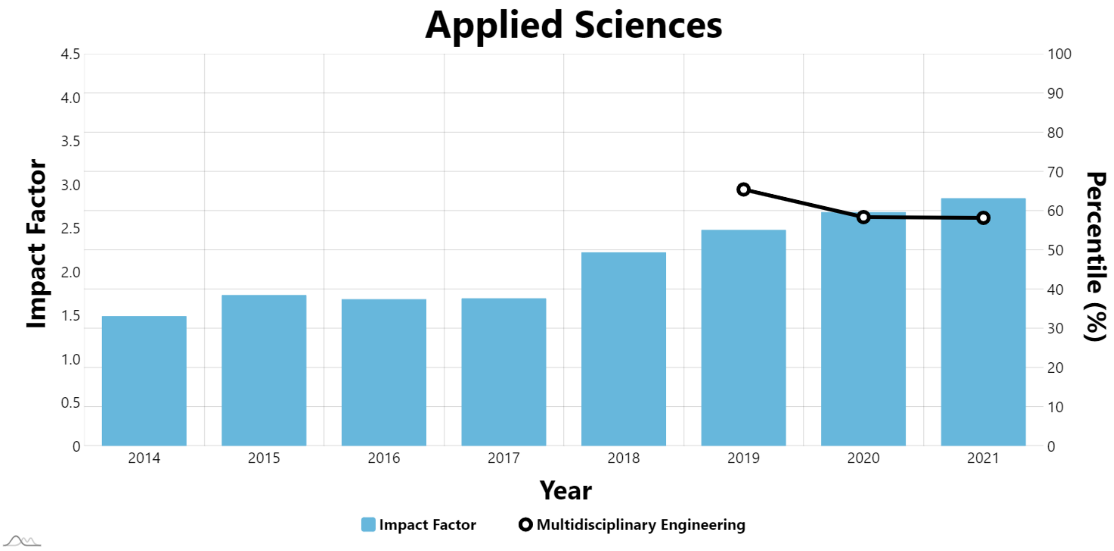


***Figure F.9:*** *Journal Impact Factor and percentile ranking of the MDPI journal "Applied Sciences".*

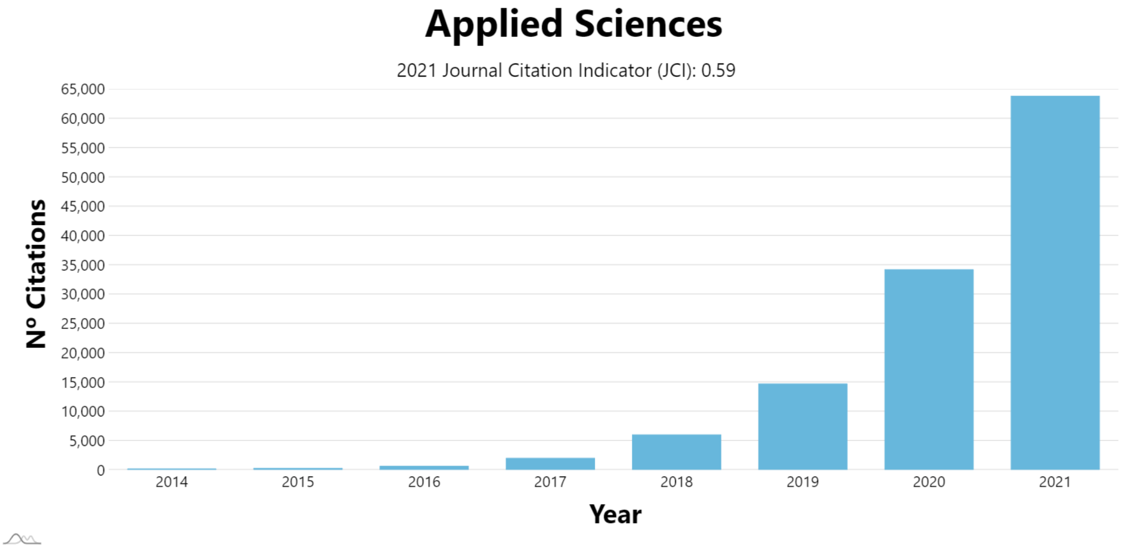


*Figure F.10: Number of yearly citations for the MDPI journal "Applied Sciences".*

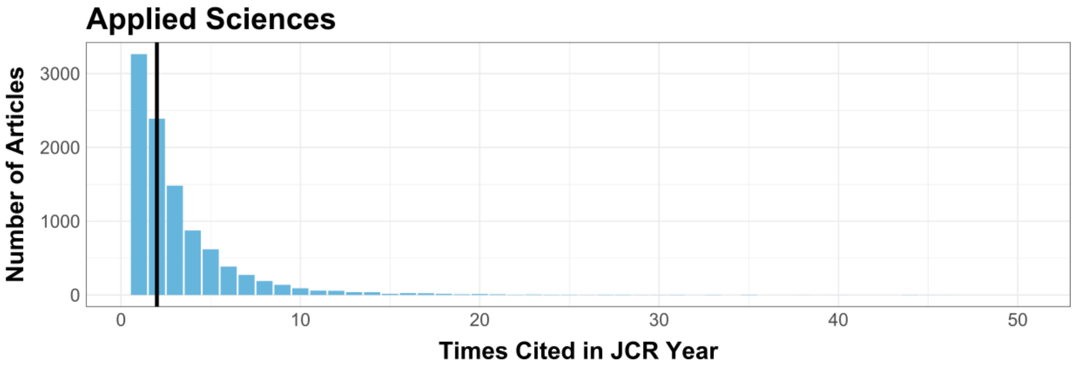


*Figure F.11: Citation Distribution graph describing the number of articles that are cited in a year in the MDPI journal "Applied Sciences". The median article citation is marked by the vertical black line ($\tilde{x} = 2$)*

### F.2.3 Journal of Clinical Medicine

*Journal of Clinical Medicine* is an open-access journal affiliated with the International Bone Research Association, Italian Resuscitation Council, Spanish Society of Hematology and Hemotherapy, Japan Association for Clinical Engineers, and the European Independent Foundation in Angiology/Vascular Medicine, among others. The journal's scope focuses on both clinical and pre-clinical research.

| | |
|---|---|
| Abbreviation | J. Clin. Med. |
| Editorial | MDPI |
| Address | St Alban-Anlage 66, CH-4052 Basel, Switzerland |
| First Index Year | 2017 |
| Current Fields (Rank, Quartile) | General & Internal Medicine (50 / 329, Q1) |
| Current JIF | 4.964 |
| Current JIF (no self-citations) | 4.722 |
| Current JCI | 1.01 |
| Nº PhD Publications | 1 |

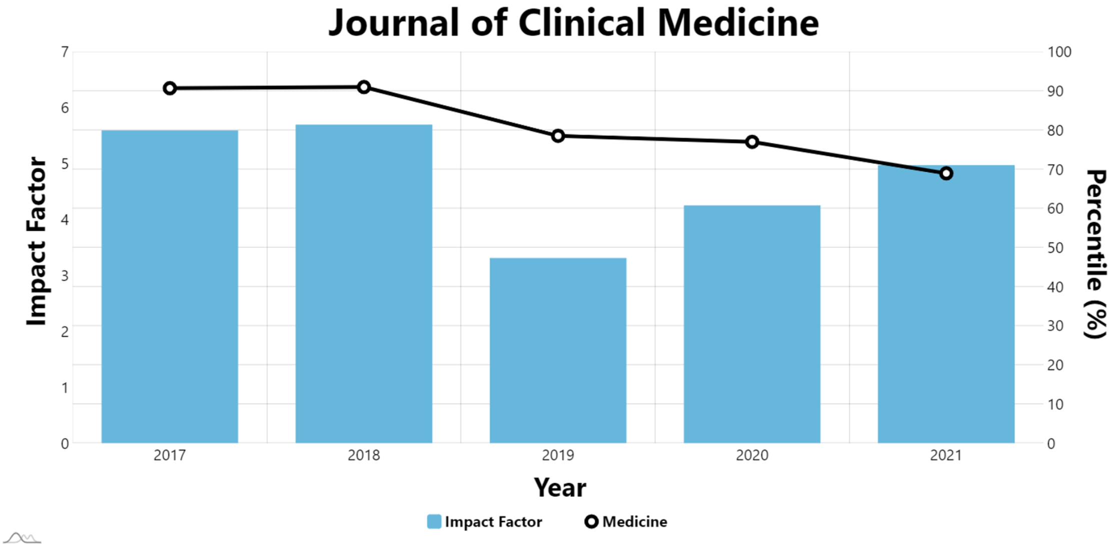


***Figure F.12:** Journal Impact Factor and percentile ranking of the MDPI journal "Journal of Clinical Medicine".*

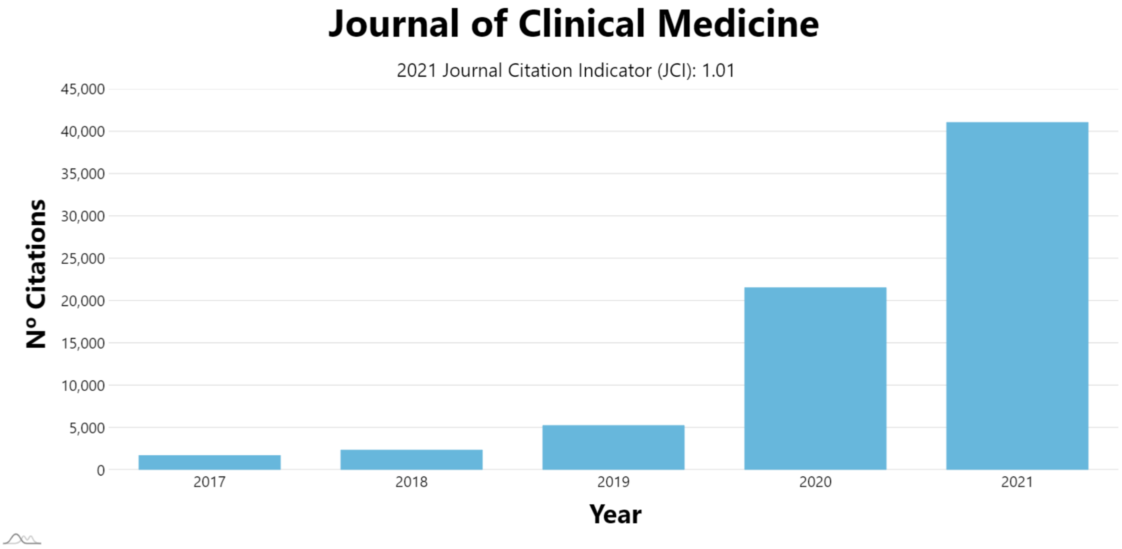


*Figure F.13:* Number of yearly citations for the MDPI journal "Journal of Clinical Medicine".

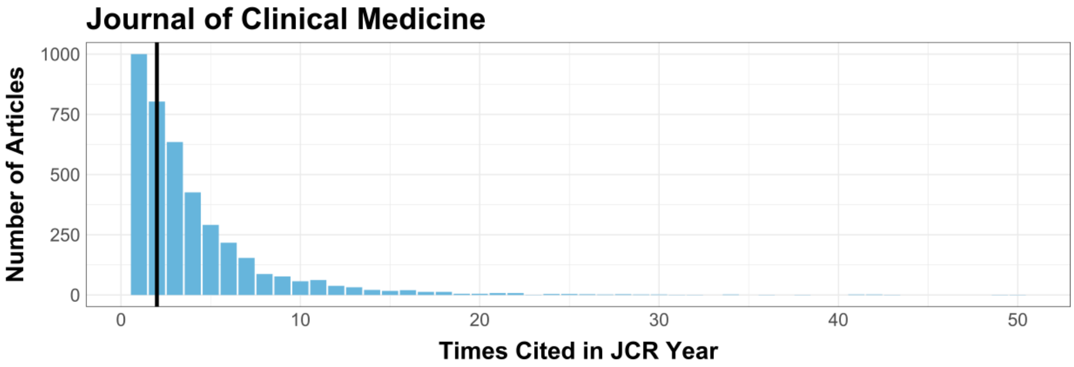


*Figure F.14:* Citation Distribution graph describing the number of articles that are cited in a year in the MDPI journal "Journal of Clinical Medicine". The median article citation is marked by the vertical black line ($\tilde{x} = 2$)

### F.2.4 PLoS ONE

*PLoS ONE* is a multidisciplinary open-access journal that publishes research related with the natural sciences, medical research, engineering, as well as the related social sciences and humanities.

| | |
|---|---|
| Abbreviation | PLoS ONE |
| Editorial | Public Library Science |
| Address | 1160 Battery Street, STE100, San Francisco, California, 94111, USA |
| First Index Year | 2009 |
| Current Fields (Rank, Quartile) | General & Multidisciplinary Sciences (29 / 73, Q2) |
| Current JIF | 3.752 |
| Current JIF (no self-citations) | 3.608 |
| Current JCI | 0.88 |
| Nº PhD Publications | 1 |

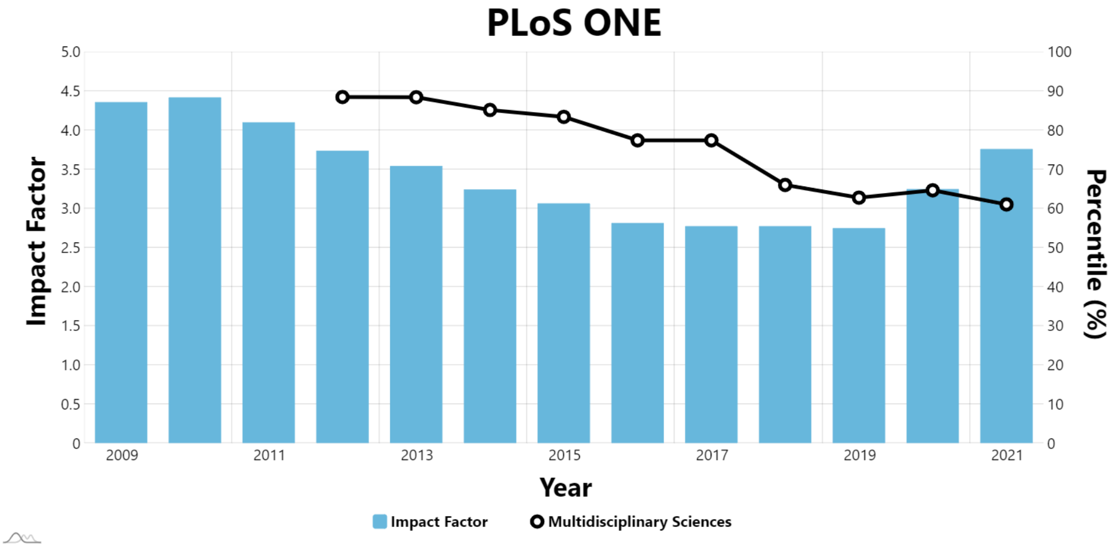


***Figure F.15:*** *Journal Impact Factor and percentile ranking of the Public Library Science journal "PLoS ONE".*

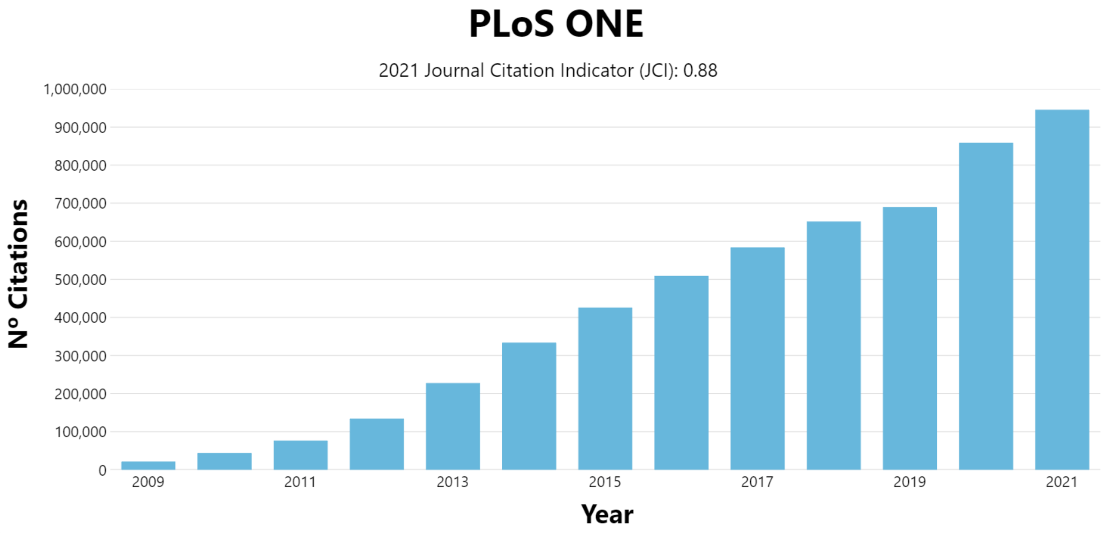


*Figure F.16:* *Number of yearly citations for the Public Library Science journal "PLoS ONE".*

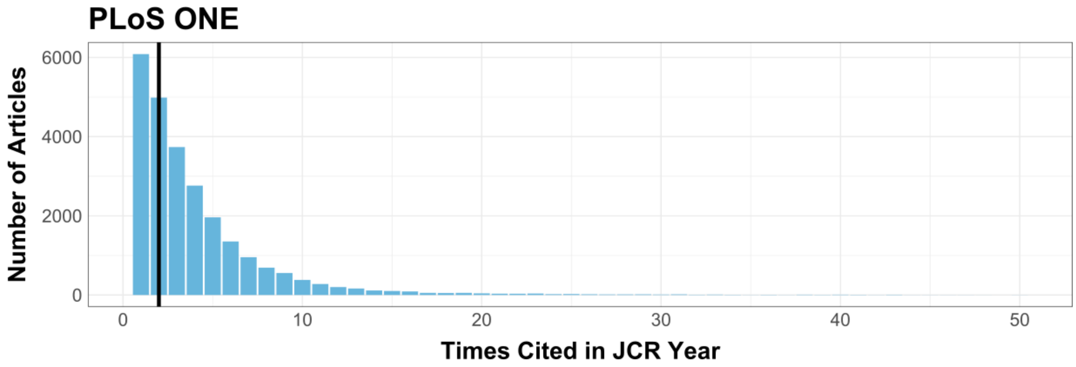


*Figure F.17:* *Citation Distribution graph describing the number of articles that are cited in a year in the Public Library Science journal "PLoS ONE". The median article citation is marked by the vertical black line* ($\tilde{x} = 2$)

## F.2.5 Quaternary Science Reviews

*Quaternary Science Reviews* is a journal that publishes research related with all aspects of Quaternary Science, including research in geology, geomorphology, geography, archaeology, soil science, palaeobotany, palaeontology and palaeoclimatology.

| | |
|---|---|
| Abbreviation | Quat. Sci. Rev. |
| Editorial | Elsevier |
| Address | The Boulevard, Langford Lane, Kidlington, Oxford OX51GB, England |
| First Index Year | 1997 |
| Current Fields (Rank, Quartile) | Multidisciplinary Geosciences (48 / 201, Q1) |
| | Physical Geography (12 / 48, Q1) |
| Current JIF | 4.456 |
| Current JIF (no self-citations) | 3.916 |
| Current JCI | 1.16 |
| Nº PhD Publications | 1 |

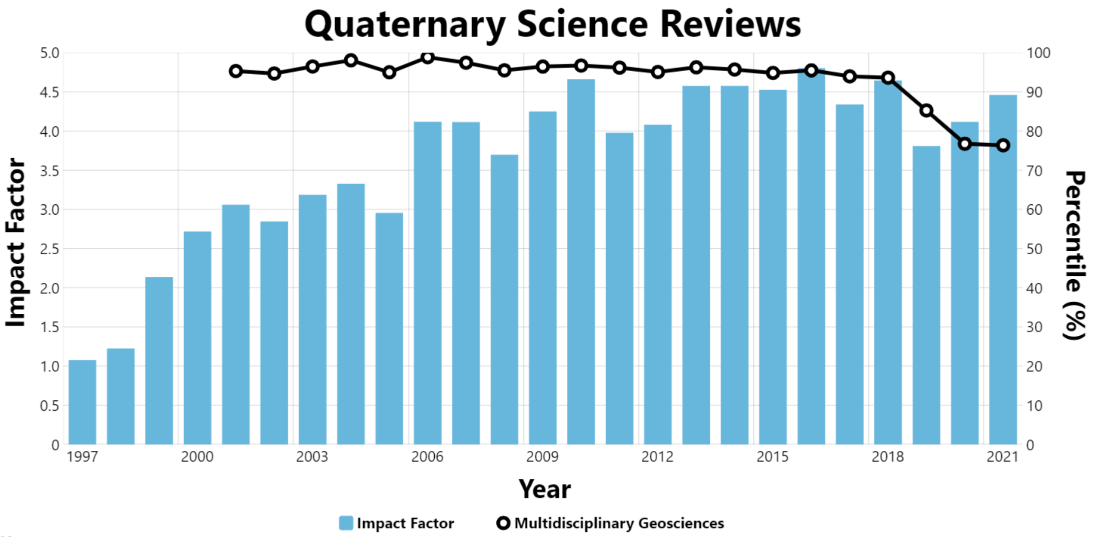


***Figure F.18:*** *Journal Impact Factor and percentile ranking of the Elsevier journal "Quaternary Science Reviews".*

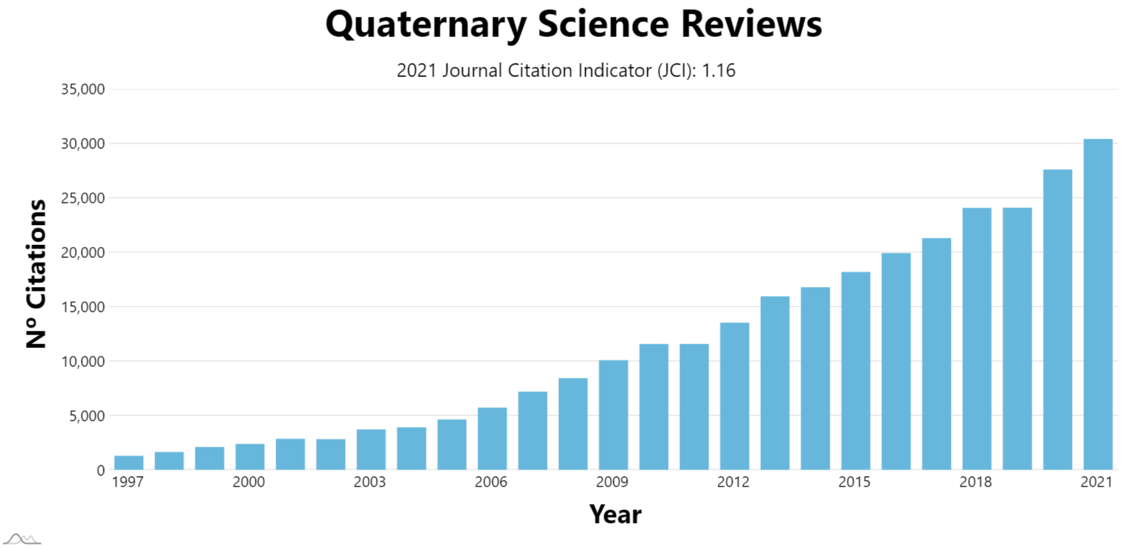


*Figure F.19:* *Number of yearly citations for the Elsevier journal "Quaternary Science Reviews".*

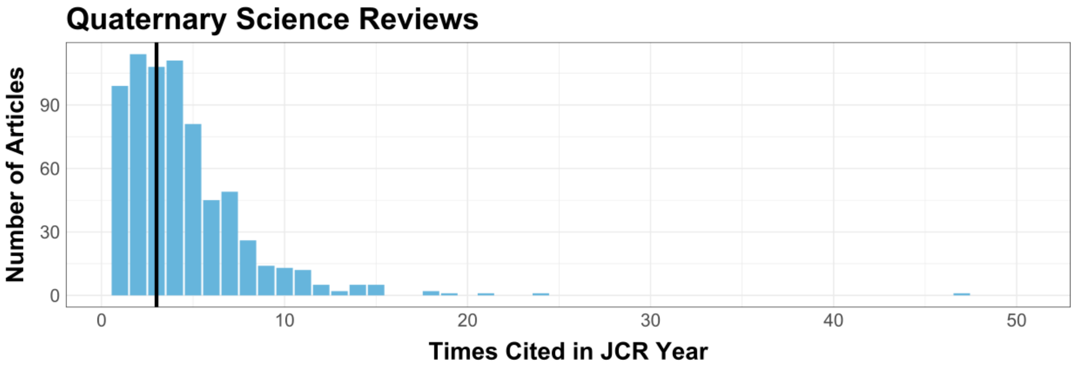


*Figure F.20:* *Citation Distribution graph describing the number of articles that are cited in a year in the Elsevier journal "Quaternary Science Reviews". The median article citation is marked by the vertical black line ($\tilde{x} = 3$)*

### F.2.6 Scientific Reports

*Scientific Reports* is a multidisciplinary open-access journal that publishes research related with all areas of the natural sciences, psychology, medicine and engineering.

| | |
|---|---|
| Abbreviation | Sci. Rep. |
| Editorial | Nature Portfolio |
| Address | Heidelberg Platz 3, Berlin 14197, Germany |
| First Index Year | 2011 |
| Current Fields (Rank, Quartile) | Multidisciplinary Sciences (19 / 73, Q2) |
| Current JIF | 4.996 |
| Current JIF (no self-citations) | 4.784 |
| Current JCI | 1.05 |
| Nº PhD Publications | 1 |

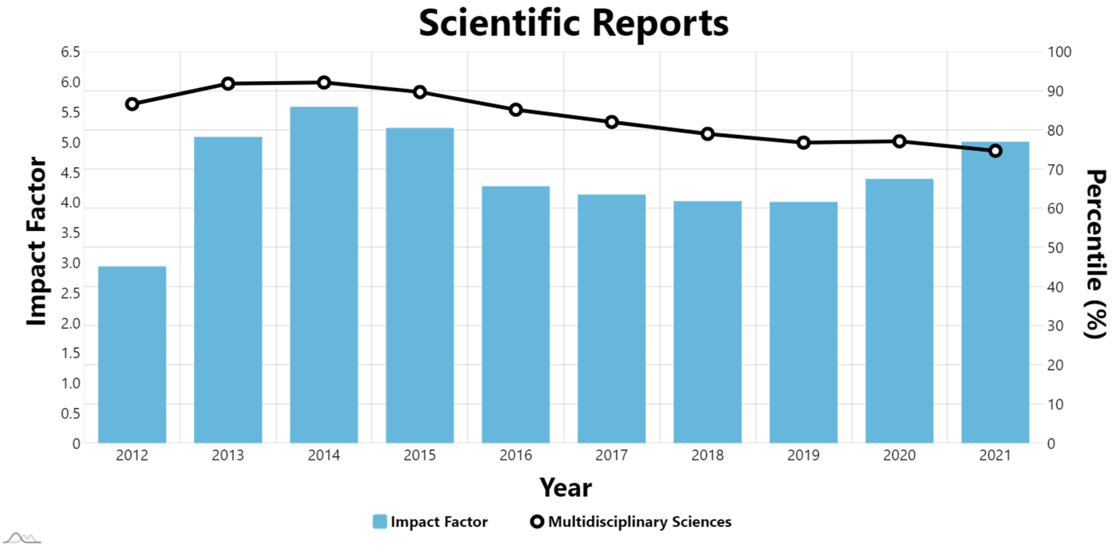


***Figure F.21:*** *Journal Impact Factor and percentile ranking of the Nature Portfolio journal "Scientific Reports".*

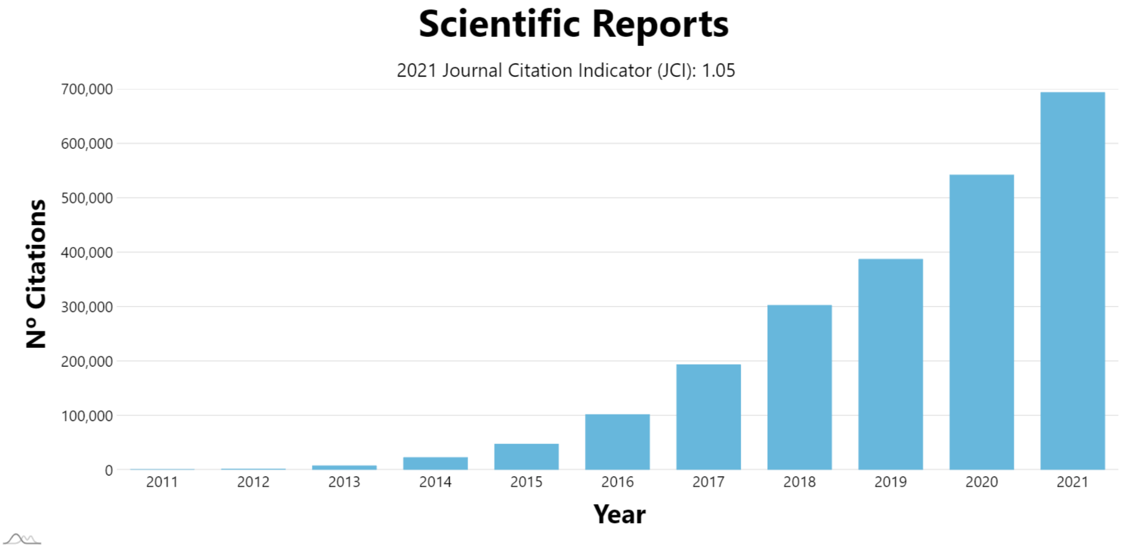


*Figure F.22: Number of yearly citations for the Nature Portfolio journal "Scientific Reports".*

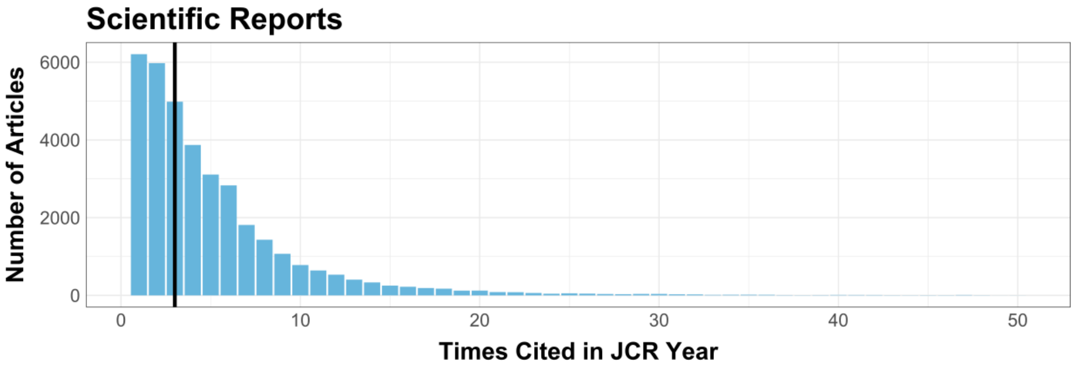


*Figure F.23: Citation Distribution graph describing the number of articles that are cited in a year in the Nature Portfolio journal "Scientific Reports". The median article citation is marked by the vertical black line ($\tilde{x} = 3$)*

## F.3 Overall Academic Career of the Doctoral Candidate

Beyond the work of his PhD, the Doctoral Candidate is author and co-author of an additional 33 scientific publications in multiple fields of research; 10 of these publications were carried out prior to beginning his PhD, while he is now a corresponding and primary author to a total of 16 articles. As of December, 2022, Lloyd Austin Courtenay's Scopus profile calculates a h-index (a measure of productivity against citation impact) of 12, with a total of 402 registered citations of his work. Google Scholar on the other hand calculates a h-index of 14, with 547 citations. In continuation, a series of figures details the precise nature of the Doctoral Candidate's research, including some additional information on the nature of his collaborations.

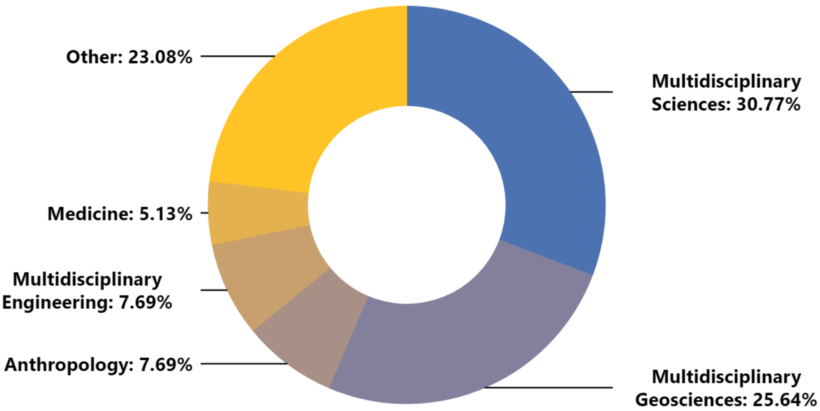


***Figure F.24:*** *Doughnut diagram presenting the different fields of science where the Doctoral Candidate has published, or collaborated in the publication of research, between the years 2017 and 2022*

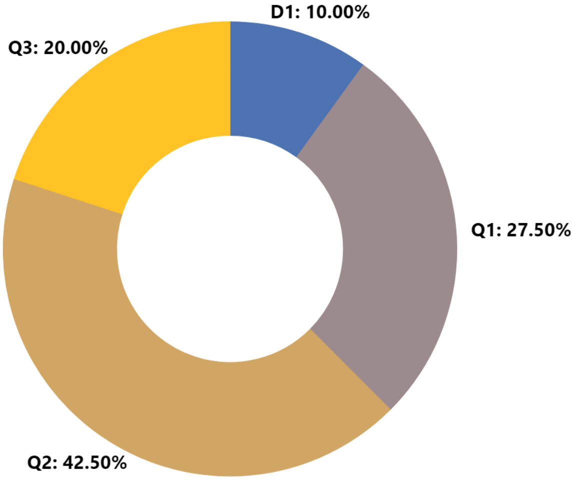


***Figure F.25:*** *Doughnut diagram presenting the quartile ranking of each journal where the Doctoral Candidate has published, or collaborated in the publication of research, between the years 2017 and 2022*

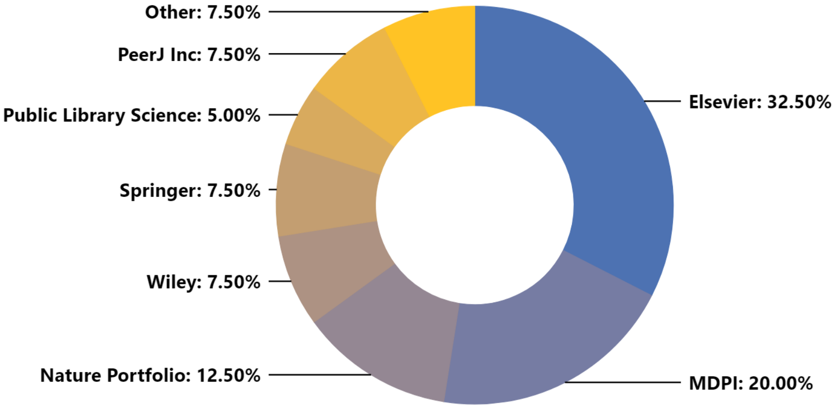


***Figure F.26:*** *Doughnut diagram presenting the editorials where the Doctoral Candidate has published, or collaborated in the publication of research, between the years 2017 and 2022*

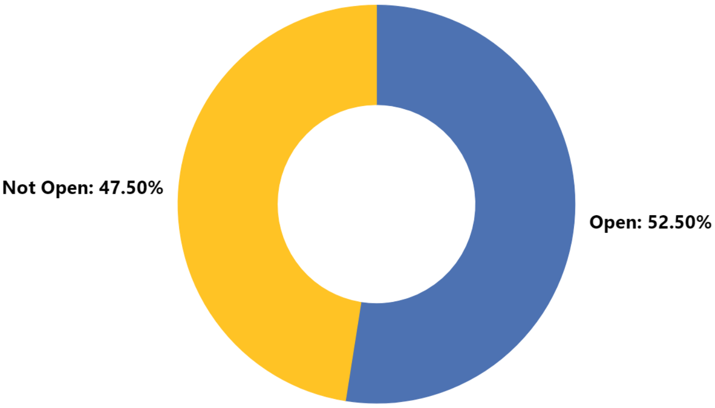


***Figure F.27:*** *Doughnut diagram presenting the proportion of scientific articles published as Open-Access by, or in collaboration with, the Doctoral Candidate between the years 2017 and 2022*

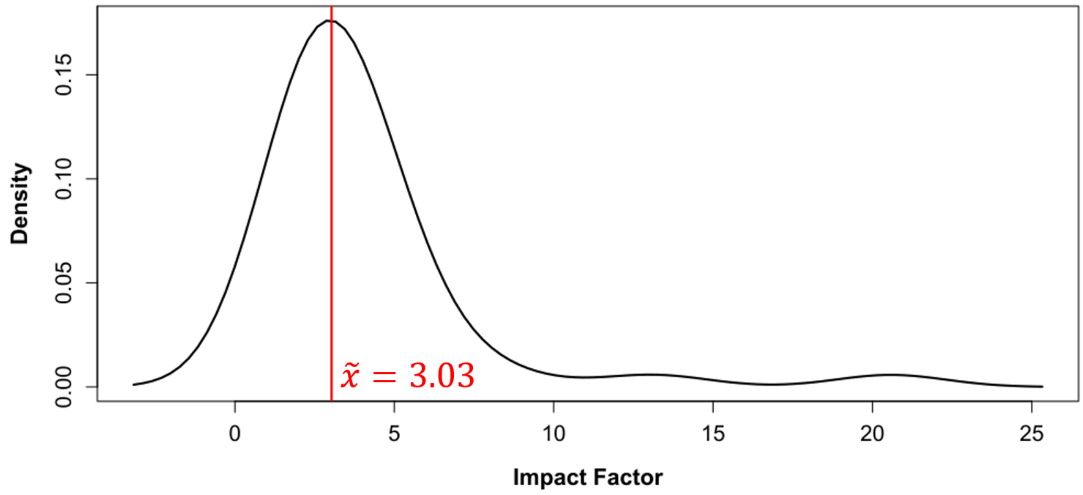


***Figure F.28:*** *Density plot of the JIF metrics associated to each of the articles published by, or in collaboration with, the Doctoral Candidate between the years 2017 and 2022. The median Impact Factor (2.84) is indicated by the red line.*

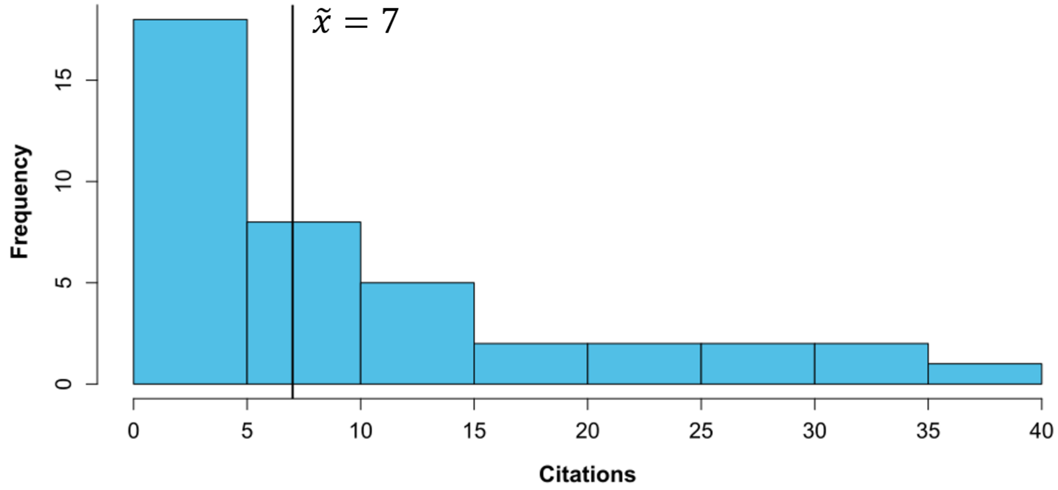


***Figure F.29:*** *Histogram of the number of times a document published by, or in collaboration with, the Doctoral Candidate has been cited between the years 2017 and 2022, according to the Scopus database (`https://www.scopus.com/`, as of the 13th of December, 2022). The median number of citations (7) is indicated by the black line.*

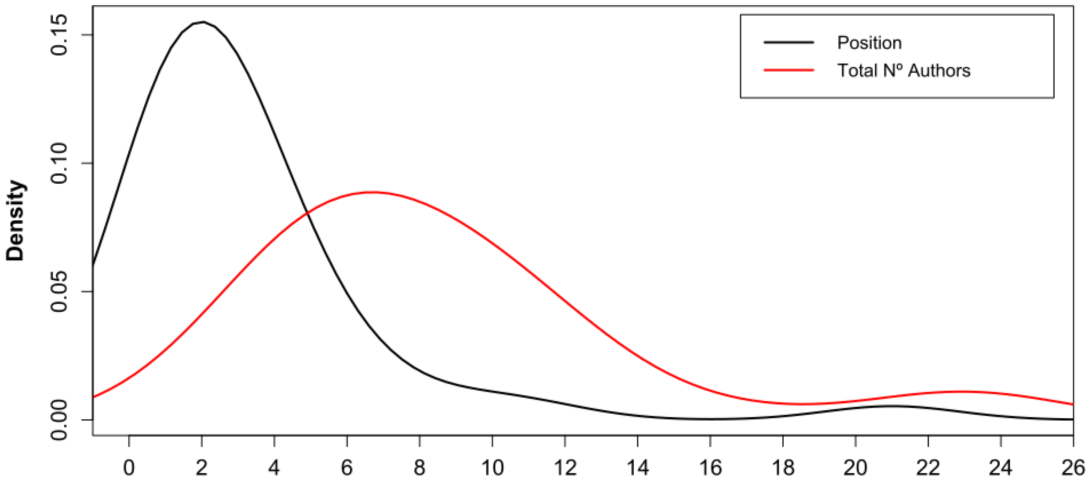


***Figure F.30:*** *Density plots of (black) the general position that the Doctoral Candidate has signed each of the publications between the years 2017 and 2022 with respect to (red) the total number of authors per article. The mode position has been calculated at 1, while the mode total number of authors per articles is 7.*

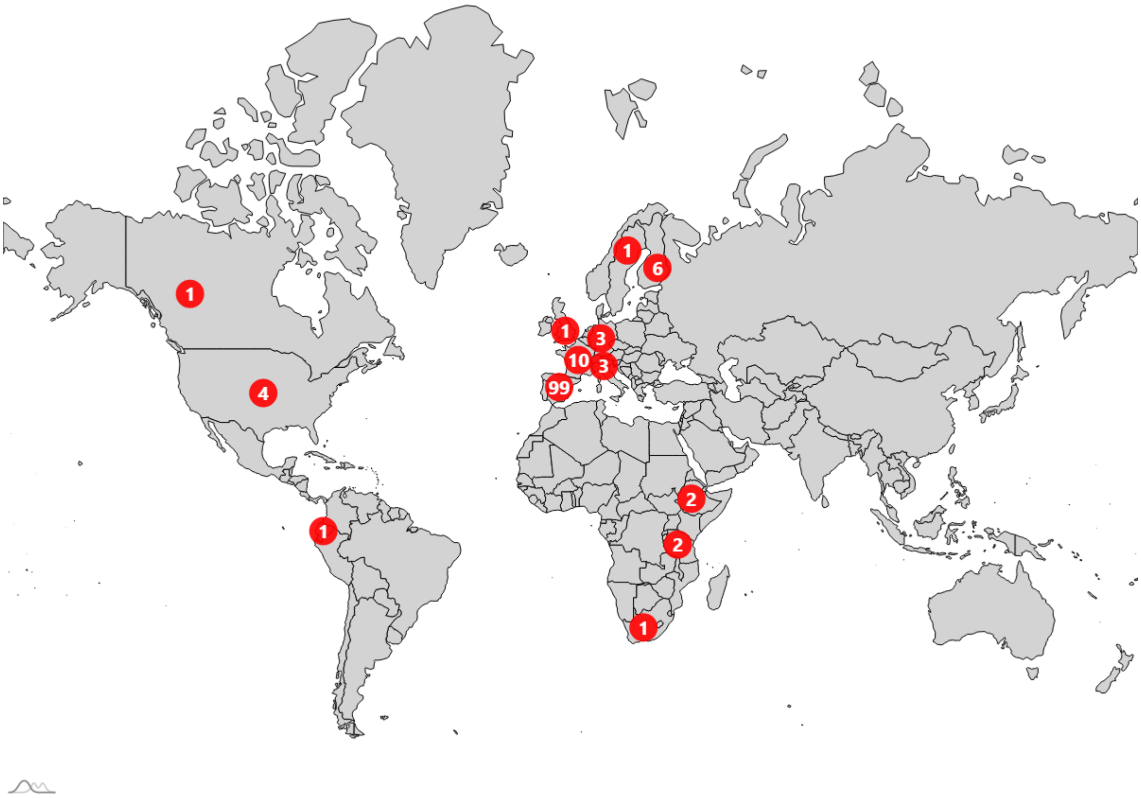


***Figure F.31:*** *A map detailing the number of authors from different countries the Doctoral Candidate has collaborated with in scientific publications between the years 2017 and 2022.*