

Research



Cite this article: Vanhoof MJM, Croquet B, De Grootte I, Vereecke EE. 2023 Principal component and linear discriminant analyses for the classification of hominoid primate specimens based on bone shape data. *R. Soc. Open Sci.* **10**: 230950.

<https://doi.org/10.1098/rsos.230950>

Received: 3 July 2023

Accepted: 1 September 2023

Subject Category:

Organismal and evolutionary biology

Subject Areas:

evolution/palaeontology/biomedical engineering

Keywords:

carpal, logistic regression, PCA, LDA, morphology

Author for correspondence:

Marie J. M. Vanhoof

e-mail: marie.vanhoof@kuleuven.be

†Joint first author.

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.6837169>.

Principal component and linear discriminant analyses for the classification of hominoid primate specimens based on bone shape data

Marie J. M. Vanhoof^{1,†}, Balder Croquet^{2,3,†},
Isabelle De Grootte^{4,5} and Evie E. Vereecke¹

¹Department of Development & Regeneration, Biomedical Sciences Group, KU Leuven Campus Kulak, Kortrijk, Belgium

²Medical Imaging Research Center, UZ Leuven, Leuven, Belgium

³Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium

⁴Department of Archaeology, Ghent University, Ghent, Belgium

⁵Research Centre in Evolutionary Anthropology and Paleoecology, Liverpool John Moores University, Liverpool L3 3AF, UK

MJMV, 0000-0002-4160-5220; IDG, 0000-0002-9860-0180

In this study, we tested the hypothesis that machine learning methods can accurately classify extant primates based on triquetrum shape data. We then used this classification tool to observe the affinities between extant primates and fossil hominoids. We assessed the discrimination accuracy for an unsupervised and supervised learning pipeline, i.e. with principal component analysis (PCA) and linear discriminant analysis (LDA) feature extraction, when tasked with the classification of extant primates. The trained algorithm is used to classify a sample of known fossil hominoids. For the visualization, PCA and uniform manifold approximation and projection (UMAP) are used. The results show that the discriminant function correctly classified the extant specimens with an F1-score of 0.90 for both PCA and LDA. In addition, the classification of fossil hominoids reflects taxonomy and locomotor behaviour reported in literature. This classification based on shape data using PCA and LDA is a powerful tool that can discriminate between the triquetrum shape of extant primates with high accuracy and quantitatively compare fossil and extant morphology. It can be used to support taxonomic differentiation and aid the further interpretation of fossil remains. Further testing is necessary by including other bones and more species and specimens per species extinct primates.

1. Introduction

The discovery of new fossils and the intensive study of this fossil evidence during the past decade has provided valuable insights into the evolution of primates, including their locomotion, diet, social behaviour and cognition (e.g. [1–10]). Fossil evidence of primates can be traced back to the early Eocene, at least 56.8 million years ago [11–13]. Some key features that are used to identify primate fossils include dental characteristics, cranial morphology and postcranial elements. The fossil record of early primates is largely comprised of dentitions. However, although teeth can indicate phylogenetic relationships and dietary preferences or feeding behaviour, they do not provide much information on positional behaviour or substrate preference [14]. The shape and structure of the skull, such as the size of the braincase or the position of the eye sockets, can provide important information about the primate's evolutionary relationships (e.g. [15,16]), while postcranial skeletal elements, such as limb bones, can provide important information on the locomotor behaviour and adaptations to different environments (e.g. [17–20]).

The preservation of complete bones offers the opportunity to examine a range of primitive and derived skeletal traits preserved in these fossils. Unfortunately, the identification of primate fossils can be challenging as the fossil record is mainly represented by isolated bones or bone fragments. The lack of complete skeletons means that researchers have to rely on a limited number of bones to identify the species and its overall morphology. Moreover, hand bones are underrepresented in the fossil record due to taphonomic processes and burial practice [21]. Especially in the case of secondary burial, the small hand bones are more likely to be left behind compared to other skeletal elements [22,23]. However, fossil long bones can become fragmented or eroded, making it difficult to determine their shape or size accurately. Despite these challenges, a variety of techniques is available to identify primate fossils, such as molecular techniques (e.g. DNA sequencing [24,25]), medical imaging techniques (e.g. CT-scanning), three-dimensional geometric morphometrics (e.g. [26–29]) and machine learning (e.g. [30]). By combining these techniques, researchers can gain a more complete understanding of the evolutionary history of primates [31].

Over the last several decades, machine learning has become an increasingly fine-tuned approach for classification purposes [32–37]. Unlike automated classification techniques, machine learning depends on the 'learning' capacity of the model, improving classification and generalization via quantitative repetition and adjustment through a training process. In a previous study, we showed that morphological characteristics of the primate triquetrum can be used to distinguish among different extant primate taxa [29]. The results revealed that the triquetrum shape of quadrupedal primates (e.g. chimpanzees and gorillas), which mainly use their wrist under compressive conditions, differs from that of suspensory primates (e.g. orangutans and gibbons) which have a wrist that is potentially exposed to tensile and torsional forces (see electronic supplementary material, for more details). In the present study, we want to use a classification algorithm for categorization of the triquetrum of known primate fossils to investigate if the results of the classification match information on taxonomy and locomotor behaviour that is available in literature. The large dataset of our previous study on primate triquetra [29] will be used in the training process, and the triquetrum of four extinct fossil primate species (*Ekembo heseloni*, *Australopithecus sediba*, *Homo naledi*, *Homo neanderthalensis*) is included to test the performance of the classification analysis.

Ekembo heseloni (25–30 mya) is one of the earliest hominoids [38,39]. *Ekembo heseloni* was formerly placed in *Proconsul* but later attributed to its own genus, together with *E. nyanzae*, to account for the substantial morphological variation between *Ekembo* and *Proconsul* [40–42]. Based on fossil evidence, it is suggested that *E. heseloni* was mainly an arboreal pronograde quadrupedal primate [2,43]. *Australopithecus sediba* (2 mya) is an extinct hominin species with a hand, foot, pelvis and spine that combined primitive *Australopithecus*-like and derived *Homo*-like character states [44–47]. Moreover, the forelimb was apparently adapted to competence in climbing and suspensory locomotor behaviours [48]. To date, there is still some debate about the exact phylogenetic position of *Au. sediba* [49]. *Homo naledi* (335 000–236 000 years ago) is an extinct hominin species that was bipedal and stood upright [50]. They share a derived wrist morphology with Neanderthals and modern humans, which is considered as an adaptation for manipulation such as tool use [7]. However, the more curved digits of *H. naledi* indicate frequent use of the hand for grasping during climbing and suspension behaviour [7,51]. *Homo neanderthalensis* (approximately 40 000 years ago), also known as Neanderthals, were a close evolutionary relative of modern humans. Neanderthals were adapted to a cold, harsh environment and had adaptations such as a robust ribcage, wide pelvis, and short limbs that helped

to conserve heat [52,53]. They were capable of bipedal walking and are known for their sophisticated tool-making abilities and cultural practices [54,55]. Their robust hands suggest that they were primarily adapted for power and force transmission through the wrist during manipulation [56,57], although recent research has shown that Neanderthals used systematic forceful precision grasping, during which the thumb forcefully secures a tool against the fingers and/or the palm [58].

In this study, we use a step-wise machine learning approach to test the following hypotheses: H1) we expect that the outcome of the classification analyses will confirm previous results of a 3DGM analysis of the primate triquetrum [29]; H2) we expect that the extant primates of the test dataset will be classified under the correct taxon on species level; H3) we expect that the classification of known hominoid fossils will support information on the locomotor behaviour that is available in literature.

2. Methods

2.1. Data acquisition

In this study, we analyse the classification of extant anthropoid primate and fossil hominoid triquetra, where triquetrum shape is discretized as a collection of fixed homologous landmarks.

2.1.1. Sample details

The extant sample used in this study includes three-dimensional surface meshes of the triquetrum of 304 anthropoid primate specimens representing 15 different species of four taxonomic clades, including plathyrrhines (*Ateles geoffroyi*), cercopithecoids (*Macaca mulatta*, *Macaca fascicularis*, *Mandrillus sphinx*, *Papio anubis*), hylobatids (*Hylobates lar*, *Hoolock hoolock*, *Symphalangus syndactylus*) and hominids (*Gorilla gorilla*, *Gorilla beringei*, *Pongo abelii*, *Pongo pygmaeus*, *Pan troglodytes*, *Pan paniscus*, *Homo sapiens*). The fossil sample includes three-dimensional surface meshes of the triquetrum of six hominoid specimens representing four extinct species (*Ekembo heseloni*, *Australopithecus sediba*, *Homo naledi*, *Homo neanderthalensis*). Details of the sample are provided in table 1 and electronic supplementary material, table S1. The extant sample was used to develop the classification model and was split into a training and test dataset (253/51) using stratification on the 19 species labels (electronic supplementary material, figure S1). The fossil sample is used as a test case and projected in the feature space. For each specimen, three-dimensional surface meshes of the left triquetrum were used and, when not available, the right triquetrum was mirrored. Only adult healthy specimens were included in the sample.

2.1.2. Landmarks

To capture the overall shape of the triquetrum, we used fixed landmarks. We positioned 18 landmarks on the surface mesh of the triquetrum, based on definitions of previous publications [29]. Full details of the landmark definitions and positioning are provided in table 2 and electronic supplementary material, figure S2. All landmark positioning was done in *Landmark Editor* software (version 3.0) [59].

2.2. Feature extraction

Feature extraction is the process of retrieving relevant information from the data, removing noise, and reducing the dimensionality [33]. From the user perspective, this can improve the interpretability and facilitate subsequent pattern recognition. In this work, we start with 18 manually placed fixed three-dimensional landmarks [29] (electronic supplementary material, figure S2), i.e. 54 dimensions. A generalized Procrustes analysis (GPA) [60,61] was carried out on all fixed landmark coordinates to remove the effects of variation in location, orientation, and scale from the coordinates, and superimpose the objects into a common coordinate system. These aligned shape coordinates are used in two linear feature extraction techniques to convert the data into a lower dimensional representation to improve classification and interpretability: principal component analysis (PCA) and linear discriminant analysis (LDA). These dimensionality reduced landmarks are further referred to as features.

2.2.1. Principal component analysis (PCA)

Principal component analysis (PCA) is a widely used unsupervised machine learning technique that can be used for feature extraction. Unsupervised learning is a branch of machine learning algorithms in

Table 1. Total triquetrum sample analysed in this study by species and sex.

genus	species/subspecies	female	male	unknown	total
EXTANT PRIMATES					
<i>Homo</i>	<i>sapiens</i>	24	3	2	29
<i>Pan</i>	<i>paniscus</i>	10	11	0	21
	<i>trogodytes</i>	24	34	4	62
<i>Gorilla</i>	<i>gorilla</i>	23	21	2	36
	<i>beringei</i>	5	12	1	18
<i>Pongo</i>	<i>pygmaeus</i>	14	9	1	24
	<i>abelii</i>	11	5	0	15
<i>Symphalangus</i>	<i>syndactylus</i>	2	3	1	6
<i>Hoolock</i>	<i>hoolock</i>	4	4	1	9
<i>Hylobates</i>	<i>lar</i>	7	10	2	19
<i>Papio</i>	<i>anubis</i>	8	10	0	18
<i>Macaca</i>	<i>fascicularis</i>	7	11	1	19
	<i>mulatta</i>	1	2	5	8
<i>Mandrillus</i>	<i>sphinx</i>	2	6	1	9
<i>Ateles</i>	<i>geoffroyi</i>	9	2	0	11
FOSSIL PRIMATES					
<i>Ekembo</i>	<i>heseloni</i>	0	0	1	1
<i>Australopithecus</i>	<i>sediba</i>	0	0	2	2
<i>Homo</i>	<i>naledi</i>	0	0	1	1
<i>Homo</i>	<i>neanderthalensis</i>	0	0	2	2
Total sample					310

which patterns can be extracted from unlabelled data. Formally, PCA is defined as ‘the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized.’ [62]. This implies that PCA constructs a new feature representation where the original data are represented as a linear combination of the previous features and for which the components are organized by variance. Therefore, the first components describe more variance in the data while the latter are assumed to be of less importance in the description of the data.

2.2.2. Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) is a supervised machine-learning technique that also can be used for feature extraction. Supervised learning is a branch of machine learning algorithms that makes use of labelled data, which often results in a model that is more driven towards a certain outcome (e.g. classification). LDA constructs a new feature representation in which the separation between the means of the projected classes is maximized and the within-class variance is minimized. In other words, it projects the data to a subspace in which the classes can be optimally separated.

2.3. Classification

The classification model is designed to assign a pre-defined species label to the feature representation of a specimen, therefore it is able to perform taxonomic classifications of the triquetrum samples. We used logistic regression as a classification algorithm. Logistic regression first linearly projects the input data and then applies a SoftMax function [33]. The result is a vector of which the size is equal to the amount of classes and of which the rows contain values between 0 and 1, which can be interpreted as the probability that the specimen is of the corresponding class. In this work, we used a multinomial

Table 2. Definitions of the fixed landmarks to capture external triquetrum shape.

#	type ^a	description
1	II	most proximopalmar point on the lunate surface
2	III	most convex point on the dorsal border of the lunate surface, between 1 and 3
3	II	most proximodorsal point on the lunate surface
4	II	most dorsodistal point between the lunate and hamate surfaces
5	III	most concave point along the distal ridge of the lunate surface connecting 4 and 6
6	II	most palmodistal point between the lunate and hamate surfaces
7	II	most concave point around the surface center of the lunate surface
8	II	most dorsal point on the hamate surface, ridge between 7 and 9
9	II	most ulnar point on the hamate surface
10	II	most palmar point on the hamate surface, ridge between 6 and 9
11	II	most concave point around the center of the hamate surface
12	II	most ulnar point of the pisiform surface
13	II	most dorso-ulnar point of the pisiform surface
14	II	most radial point of the pisiform surface
15	II	most palmaradial point of the pisiform surface
16	II	most concave point around the center of the pisiform surface
17	II	tubercle, most ulnarly projecting point
18	II	most proximally projecting point of the ulnar/meniscus surface

^aLandmark type after [76].

classification scheme with 12 regularizations. To evaluate the classification performance, mean accuracy and weighted F1-score are used on the test dataset. Mean accuracy is the number of correct predictions and the score ranges between 0% and 100%. The F1 score ranges from 0 to 1 and is the harmonic mean of precision and recall and gives a better measure of the incorrectly classified cases. The F1 score is often preferred over accuracy when data are unbalanced [63], such as when the quantity of specimens belonging to one class significantly outnumbers those found in other classes.

2.4. Visualization

To visualize the feature representations, we need to compress these to a two-dimensional vector, for which we used PCA and Uniform Manifold Approximation and Projection (UMAP).

2.4.1. Principal component analysis (PCA)

PCA can be used for feature extraction (outlined in §1.2.1) as well as for data visualization. In data visualization, PCA projects the input on a two-dimensional principal subspace, where these two dimensions explain the most variance. As such, a visualization created with PCA highlights the global structure of the data in which the spatial relations can be studied. A limitation of using PCA is that there can be a lot of overlap between the datapoints, since the data are being linearly projected in a two-dimensional space, and it does not always show which data are grouped together in a higher dimensional space.

2.4.2. Uniform manifold approximation and projection (UMAP)

Uniform manifold approximation and projection (UMAP) is an unsupervised manifold learning technique that can be used for data visualization. It was developed as an alternative to existing dimensionality reduction methods, particularly t-SNE (t-Distributed Stochastic Neighbour Embedding) [64], which was widely used for visualizing high-dimensional data but had some limitations. UMAP aims to address some of these limitations and provides a more flexible and efficient approach to

capturing the structure of complex data in lower-dimensional spaces. Intuitively, from a data visualisation perspective, UMAP first constructs a representation of the structure of the data and then reconstructs this structure in a two-dimensional space. While constructing the representation, UMAP will primarily focus on the datapoints that are close together in the high-dimensional space, referred to as neighbouring nodes [65]. A visualization created with UMAP is therefore good at conveying the local structure of the data, i.e. the datapoints that are close together in the high dimensional space will end up close together in the visualization. In our implementation we used the following parametrization: $n_neighbours = 15$, $min_dist = 0.9$ and $spread = 0.9$, which allows the technique to capture the local structure of the data while still maintaining readability.

2.5. Final pipeline

All models are built as a pipeline of three components: (1) *feature extractor*, which reduces the dimensionality of the input data and constructs a feature space in which data points can be analysed and compared; (2) *standardizer*, the feature representations are standardized to zero mean and unit variance; (3) *classifier*, a classifier is added to the final layer of the pipeline which assigns a class to the datapoints. The number of components of the feature extractor are determined using 5-fold cross-validation with stratification after which the best parameter is selected based on the model performance (see electronic supplementary material, figure S3) [66,67]. Each pipeline can be combined with a visualizer. Therefore, we refer to these combinations as PCA-PCA, PCA-UMAP, LDA-PCA and LDA-UMAP in the results and discussion sections below.

The model pipelines were developed using Scikit-learn in Python [67].

3. Results

3.1. Feature extraction

Using the 5-fold cross-validation on the training dataset, we recorded and aggregated the average accuracy for the folds that were left out (i.e. the test datasets) for both PCA and LDA. The retained number of components is 22 for PCA and 12 for LDA (electronic supplementary material, figure S3).

Figure 1a shows the PCA pipeline using two visualizers, (A) PCA and (B) UMAP. For PCA-PCA, three major clusters can be identified: (1) platyrrhines, (2) cercopithecoids, and (3) hylobatids and hominids. When looking at the third cluster, we see that all hominid genera partially overlap and that *Homo* is situated in between *Pongo*—which show the highest overlap with the hylobatids—and the African apes. The PCA-UMAP subdivides the three main clusters into subclusters (purely based on shape, non-supervised). Here, we can clearly distinguish *Pan* from *Gorilla*, and *Homo* from the hylobatids and *Pongo*.

Figure 1b shows the LDA pipeline, using A) PCA and B) UMAP as visualizers. The LDA-PCA is very similar to the PCA-PCA model, and the same three clusters can be identified. In contrast to the PCA-UMAP, the LDA-UMAP shows a more fine-grained separation of the classes as *Pongo* is clearly differentiated from the hylobatids, and *Gorilla* is distinct from *Pan*.

3.2. Classification analysis

To develop the classification model, the extant data sample was split into a training dataset and test dataset. The average accuracy values and F1 scores of the classification can be found in table 3. For the test set in the PCA-pipeline, the mean accuracy is 0.84 and the weighted F1 score is 0.82, while for the LDA-pipeline this is 0.90 for both performance scores.

The classes of our dataset are imbalanced, which means that the mean accuracy can produce results which do not accurately reflect the performance of the model. However, the F1 score, which optimizes both precision and recall, is highly similar to the mean accuracy for both the PCA- and LDA-pipeline. This shows that the models are able to distinguish, for example, a gibbon from a gorilla (precision), and each specimen from every class (recall), meaning that both models are able to classify the extant primates of the test dataset under the correct species (electronic supplementary material, figure S4 and electronic supplementary material, figure S5). However, for both pipelines, species with a similar triquetrum morphology can be confused (e.g. *G. beringei* and *G. gorilla*; *H. hoolock* and *H. lar*; *P. paniscus* and *P. troglodytes*) (figure 2).

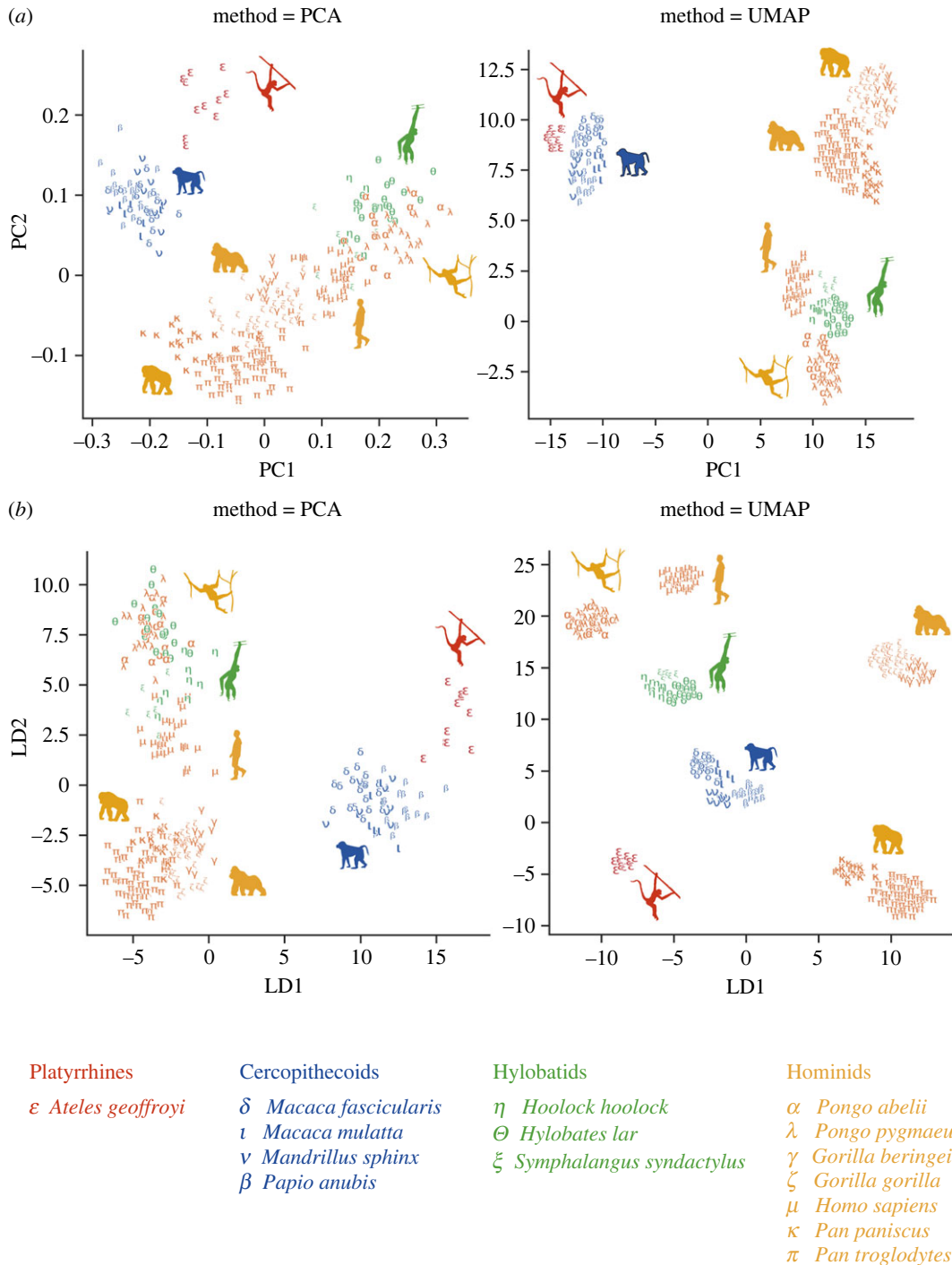


Figure 1. Two-dimensional representation of the feature space extracted from the extant specimen sample using PCA (a) and LDA (b). The left panel demonstrates the embedding performed using PCA visualizer, and shows a more continuous distribution, revealing the global structure within the dataset. The right panel illustrates the embedding performed using UMAP visualizer, which highlights the clusters found in the high dimensional feature space, revealing the local structure within the dataset.

Table 3. Average accuracy values for classification performance of the training and test datasets for PCA and LDA pipelines.

	mean accuracy		weighted F1-score	
	training	test	training	test
PCA + LR	0.992	0.843	0.992	0.824
LDA + LR	0.976	0.902	0.976	0.897

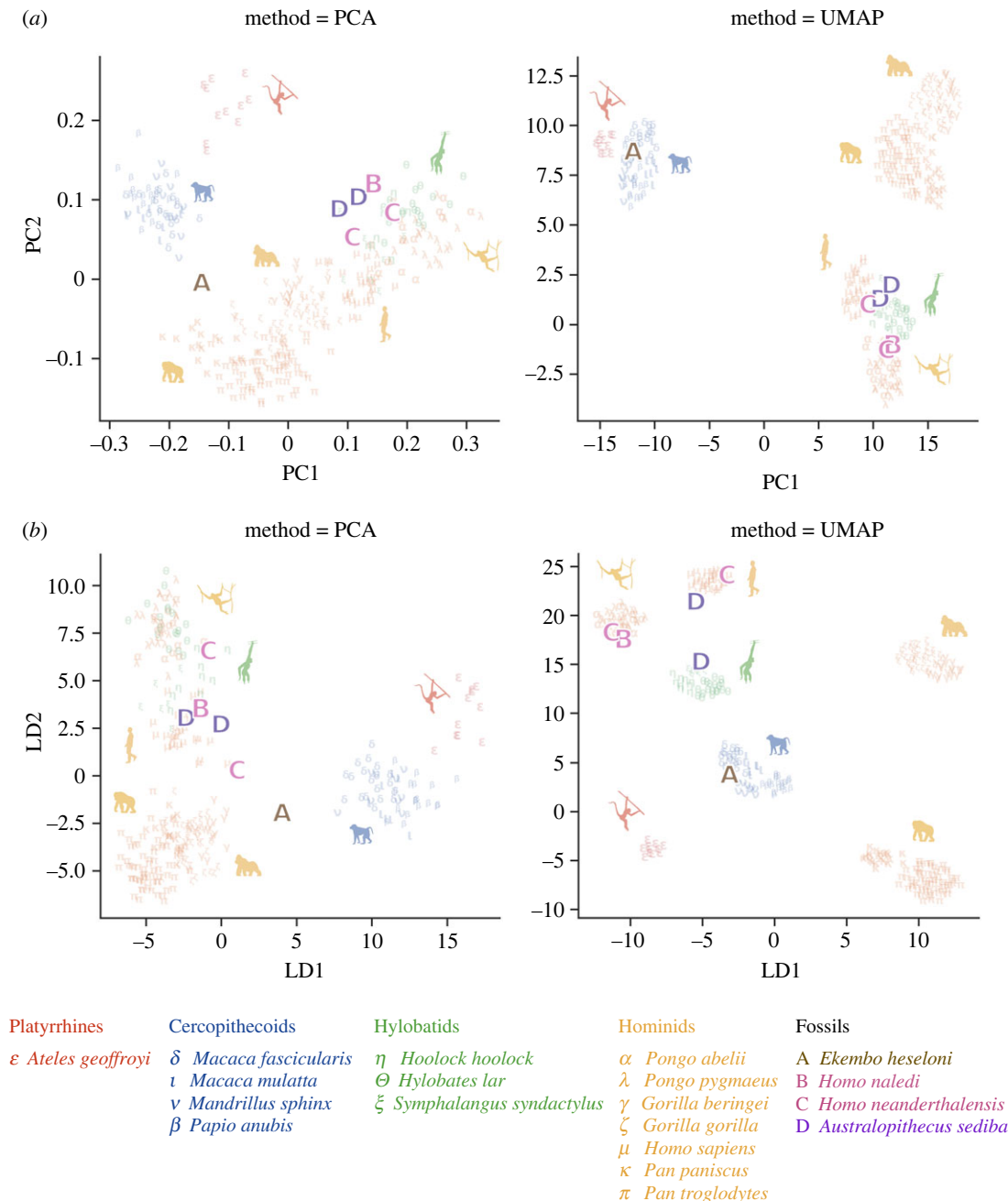


Figure 3. Two-dimensional visualization of the feature space extracted from the fossil specimen and extant specimen datasets using PCA (a) and LDA (b). The left panel demonstrates the embedding performed using PCA, the right panel illustrates the embedding performed using UMAP. The extant specimens are depicted in a lower opacity to highlight the projection of the fossil specimens.

4. Discussion

4.1. Feature extraction

We expected that the outcome of the feature extraction analyses would confirm previous results obtained using 3DGM of the triquetrum [29]. In that study, a bivariate scatterplot of PC1 against PC2 separated the platyrrhines and cercopithecoids from the other clades while the hylobatids showed overlap with the hominids in the morphospace. In addition, the different hominid genera partially overlapped, more specifically *Pongo/Homo* and *Pan/Gorilla*.

These results are confirmed by the classification analysis of this study. The PCA-PCA and LDA-PCA show the same results, with the platyrrhines and cercopithecoids being separated from the hominoids. PCA-UMAP shows further separation of the hominoids, with a clear distinction between *Pongo/Homo/*

hylobatids and the African apes. This supports our results on triquetrum shape, as the triquetrum of *Pongo* is similar to that of hylobatids, and that of *Gorilla* similar to *Pan*. In addition, we did find specific morphological traits that can be linked to a specific genus. For example, the differences in triquetrum shape between *Pan* and *Gorilla* might be related to differences in hand positioning during knuckle-walking [68,69]. For the highly arboreal hylobatids and *Pongo*, the differences in triquetrum shape might be linked to weight transfer through the ulnar side of the wrist in *Pongo* [70–72] and the frequent use of (ricochetal) brachiation of hylobatids [73–75]. This is supported by the LDA-UMAP, as *Pan* and *Gorilla* are separated into different clusters as well as *Pongo* and the hylobatids.

4.2. Classification analysis

For the classification analysis, we expected that the extant primates of the test dataset would be classified under the correct taxon on species level. This hypothesis is supported as we find that for both the PCA and LDA classification models, the test dataset is classified with high accuracy. The PCA-pipeline is slightly overfitted to the training dataset which results in a lower score on the test dataset compared to the LDA-pipeline, but the higher performance of the LDA-pipeline can be explained by the better separation in the feature space.

For both pipelines, species with a similar morphology can be confused in the classification of the test dataset (e.g. both species of *Gorilla* and both species of *Pan*). In our previous study, we did find significant differences for the triquetrum shape between both *Gorilla* species and between both *Pan* species [29] even though they showed high overlap within the morphospace. This means that although the classification models can discriminate between the triquetrum shape of extant primates with high accuracy, results need to be interpreted with caution when looking at species of the same genus.

4.3. Classification of fossil specimens

The triquetrum of the fossil *E. heseloni* lies between *Pan* and the cercopithecoids in the feature space, and using the UMAP visualization it is clearly classified in the cercopithecoid group. The cercopithecoids are mainly terrestrial quadrupedal primates which confirms the quadrupedal locomotion of *E. heseloni* that has been suggested in literature [2,43]. Although *E. heseloni* is more distinct from the other hominoid fossils, its close position relative to the cercopithecoids does not fully support its taxonomic position as one of the earliest hominoids.

Australopithecus sediba, *H. naledi* and *H. neanderthalensis* are clustered closely together in the feature space, which supports their close phylogenetic relationship. *H. naledi* is classified in the *Pongo* cluster using the UMAP visualization. This fits with the hypothesis that climbing remained a significant component of *H. naledi*'s locomotor repertoire, which is put forward as explanation of their 'primitive' shoulder morphology and curved manual phalanges [7,18]. *H. naledi* share a derived wrist morphology with the other *Homo* species (*H. neanderthalensis* and *H. sapiens*), which is supported by our analysis as these species are clustered closely together in the feature space.

Australopithecus sediba and *H. neanderthalensis* are clustered together with the hylobatids/*Pongo*/*Homo*. The clustering of *Au. sediba* close to the hylobatids and *Pongo* might be explained by their frequent use of climbing behaviour, while for Neanderthals there is no clear explanation. Both species show some of the derived features of *H. sapiens*, which might explain their classification close to the *Homo* cluster. In the LDA-PCA model, one Neanderthal specimen is clustered closer to the African ape cluster. This might indicate that this triquetrum specimen is more 'block-shaped', while the other specimen shows a more cylindrical shape, similar to the hylobatid/*Pongo* cluster (see also [29]). However, in the LDA-UMAP model, the specimens of *Au. sediba* are classified in different clusters. One of these fossils probably does not lie in a well-defined region in the feature space and is therefore pushed to the other cluster. The same accounts for the Neanderthals. This shows the danger of constructing a feature space on specific subclasses that do not directly align with the fossil data. To improve the feature space when investigating an unknown specimen, a dataset as complete as possible should be used and fossil specimens should continuously be added to the training/test dataset.

5. Conclusion

With this paper, we can demonstrate that machine learning methods have the potential for taxon identification and aid the interpretation of primate fossil remains.

The PCA model gives us a more appropriate feature space for projection of the existing data and analysis of new data. This model is more nuanced compared to LDA as it is an unsupervised technique that projects the data in a principal subspace in which the most important patterns of the data are preserved. The LDA model, on the other hand, gives us a feature space that is better suited for the separation of the different classes. For visualization, PCA can be used to find the global structure in the dataset, as you can use the distance between the datapoints to interpret the results, while UMAP is better suited to look at local structures in the data. This means that UMAP can be used to find specific groups in the feature space, even though these groups show overlap using PCA.

With this classification analysis, we want to encourage the use of traditional morphometric methodologies in combination with machine learning in order to provide additional support for identifying isolated primate fossil remains based on morphological features. This will help to solve contradicting taxonomic issues, to suggest phylogenetic relationships among fossil and living taxa, or to infer locomotory patterns (depending on the understanding of the origin of variation in the bone under study). Although this needs to be tested further on other (carpal) bones and with more specimens per species, the results of this study seem promising for future work.

Ethics. For this paper we only used three-dimensional surface meshes, therefore no ethical approval was needed.

Data accessibility. Electronic supplementary material, is published at Figshare: 10.6084/m9.figshare.24057558 [77].

The data are provided in electronic supplementary material [78].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. M.J.V.: conceptualization, data curation, investigation, methodology, resources, visualization, writing—original draft, writing—review and editing; B.C.: methodology, software, visualization, writing—original draft; I.D.G.: supervision, writing—review and editing; E.V.: project administration, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare no conflicts of interest.

Funding. We received no funding for this study.

Acknowledgements. The authors thank the different zoos and institutes which provided the extant primate specimens: Pieter Cornillie (Ghent University, campus Merelbeke), Koen Nelissen (KU Leuven, campus Gasthuisberg), François Druelle (Zoological and Botanical Park of Mulhouse, France), Robby Van der Velden (Pakawi Park, Belgium), Sergio Almécija (American Museum of Natural History, Division of Anthropology, New York), Craig Wuthrich (East Carolina University, Brody School of Medicine, Greenville, South Carolina), Emmanuel Gilissen (Royal Museum for Central Africa, Belgium), Pierre de Wit (Adventure Zoo Emmen, the Netherlands). Furthermore, we thank dr. Olivier Vanovermeire and Henk Lacaeyse from the Medical Imaging Department, AZ Groeninge (Kortrijk, Belgium) for CT-scanning of the specimens. Finally, we would like to thank the students who assisted during segmentation of the CT scan images.

The fossil primate specimens were provided by: Bernard Zipfel (University of Witwatersrand, South-Africa), Sergio Almécija (American Museum of Natural History, Division of Anthropology, New York), Antonio Rosas (National Museum of Natural Sciences, CSIC, Spain). Access to the digital data was facilitated by Tracy Kivell and Mykolas Imbrasas.

References

- Berger LR, Hawks J, Dirks PH, Elliott M, Roberts EM. 2017 Homo naledi and Pleistocene hominin evolution in subequatorial Africa. *ELife* **6**, e24234. (doi:10.7554/eLife.24234)
- Böhme M, Spassov N, Fuss J, Tröscher A, Deane AS, Prieto J, Kirscher U, Lechner T, Begun DR. 2019 A new Miocene ape and locomotion in the ancestor of great apes and humans. *Nature* **575**, 489–493. (doi:10.1038/s41586-019-1731-0)
- Butler PM. 1963 Tooth morphology and primate evolution. In *Dental anthropology*, pp. 1–13. Elsevier.
- Byrne RW. 2000 Evolution of primate cognition. *Cogn. Sci.* **24**, 543–570. (doi:10.1207/s15516709cog2403_8)
- Gebo DL, Dagosto M, Beard KC, Qi T, Wang J. 2000 The oldest known anthropoid postcranial fossils and the early evolution of higher primates. *Nature* **404**, 276–278. (doi:10.1038/35005066)
- Grabowski M, Jungers WL. 2017 Evidence of a chimpanzee-sized ancestor of humans but a gibbon-sized ancestor of apes. *Nat. Commun.* **8**, 880. (doi:10.1038/s41467-017-00997-4)
- Kivell TL, Deane AS, Tocheri MW, Orr CM, Schmid P, Hawks J, Berger LR, Churchill SE. 2015 The hand of Homo naledi. *Nat. Commun.* **6**, 8431. (doi:10.1038/ncomms9431)
- Rose KD. 2005 The earliest primates. *Evol. Anthropol.: Issues News Rev.* **3**, 159–173. (doi:10.1002/evan.1360030505)
- Rosenberger AL. 2010 Adaptive Profile Versus Adaptive Specialization: Fossils and Gummivory in Early Primate Evolution. In *The evolution of exudativory in primates* (eds AM Burrows, LT Nash), pp. 273–295. New York: Springer.
- Wuthrich C, MacLathry LM, Nengo IO. 2019 Wrist morphology reveals substantial locomotor diversity among early catarrhines: An analysis of capitates from the early Miocene of Tindret (Kenya). *Sci. Rep.* **9**, 3728. (doi:10.1038/s41598-019-39800-3)
- Ni X, Wang Y, Hu Y, Li C. 2004 A euprimate skull from the early Eocene of China. *Nature* **427**, 65–68. (doi:10.1038/nature02126)
- Rose KD, Chester SGB, Dunn RH, Boyer DM, Bloch JJ. 2011 New fossils of the oldest North American euprimate *Teilhardina brandti* (Omomyidae) from the paleocene-eocene thermal maximum. *Am. J. Phys. Anthropol.* **146**, 281–305. (doi:10.1002/ajpa.21579)
- Smith T, Rose KD, Gingerich PD. 2006 Rapid Asia–Europe–North America geographic dispersal of earliest Eocene primate *Teilhardina*

- during the Paleocene–Eocene Thermal Maximum. *Proc. Natl Acad. Sci. USA* **103**, 11 223–11 227. (doi:10.1073/pnas.0511296103)
14. Boyer DM, Toussaint S, Godinot M. 2017 Postcrania of the most primitive euprimate and implications for primate origins. *J. Hum. Evol.* **111**, 202–215. (doi:10.1016/j.jhevol.2017.07.005)
 15. Bastir M, Rosas A, Stringer C, Manuel Cuétara J, Kruszynski R, Weber GW, Ross CF, Ravosa MJ. 2010 Effects of brain and facial size on basicranial form in human and primate evolution. *J. Hum. Evol.* **58**, 424–431. (doi:10.1016/j.jhevol.2010.03.001)
 16. Schoenemann PT. 2013 Hominin Brain Evolution. In *A companion to paleoanthropology* (ed. DR Begun), pp. 136–164. Blackwell Publishing Ltd.
 17. Ankel-Simons F. 1972 Vertebral morphology of fossil and extant primates. In *The functional and evolutionary biology of primates*, pp. 18, 1st edn.
 18. Feuerriegel EM, Green DJ, Walker CS, Schmid P, Hawks J, Berger LR, Churchill SE. 2017 The upper limb of *Homo naledi*. *J. Hum. Evol.* **104**, 155–173. (doi:10.1016/j.jhevol.2016.09.013)
 19. Fleagle JG, Lieberman DE. 2015 Major transformations in the evolution of primate locomotion. In *Great transformations in vertebrate evolution*, pp. 257–279. University of Chicago Press.
 20. Young NM, Wagner GP, Hallgrímsson B. 2010 Development and the evolvability of human limbs. *Proc. Natl Acad. Sci. USA* **107**, 3400–3405. (doi:10.1073/pnas.0911856107)
 21. Brothwell DR. 1981 *Digging up bones: The excavation, treatment, and study of human skeletal remains*, 3rd edn. Ithaca, NY: Cornell University Press.
 22. Bello S, Thomann A, Lally L, Signoli M, Rabino-Massa E, Dutour O. 2003 *Calcul du 'Profil théorique de survie osseuse la plus probable' et son utilisation dans l'interprétation des processus taphonomiques pouvant déterminer la formation d'un échantillon ostéologique humain*.
 23. Bello S, Andrews P. 2006 The intrinsic pattern of preservation of human skeletons and its influence on the interpretation of funerary behaviours. In *The Social Archaeology of Funerary Remains* (eds R Gowland, C Knüsel). Barnsley, UK: Oxbow Books.
 24. Glazko GV. 2003 Estimation of Divergence Times for Major Lineages of Primate Species. *Mol. Biol. Evol.* **20**, 424–434. (doi:10.1093/molbev/msg050)
 25. Raam RL, Sterner KN, Noviello CM, Stewart C-B, Disotell TR. 2005 Catarrhine primate divergence dates estimated from complete mitochondrial genomes: Concordance with fossil and nuclear DNA evidence. *J. Hum. Evol.* **48**, 237–257. (doi:10.1016/j.jhevol.2004.11.007)
 26. Couette S, White J. 2010 3D geometric morphometrics and missing-data. Can extant taxa give clues for the analysis of fossil primates? *C.R. Palevol* **9**, 423–433. (doi:10.1016/j.crpv.2010.07.002)
 27. Daver G, Déroit F, Berillon G, Prat S, Grimaud-Hervé D. 2014 Homininés fossiles, primates quadrupèdes et l'origine de la bipédie: Une analyse morphométrique géométrique 3D de l'hamatum chez les primates. *Bulletins et Memoires de La Societe d'Anthropologie de Paris* **26**, 121–128. (doi:10.1007/s13219-014-0111-4)
 28. Terhune CE, Kimbel WH, Lockwood CA. 2007 Variation and diversity in *Homo erectus*: A 3D geometric morphometric analysis of the temporal bone. *J. Hum. Evol.* **53**, 41–60. (doi:10.1016/j.jhevol.2007.01.006)
 29. Vanhoof MJM, Galletta L, De Groot I, Vereecke EE. 2021 Functional signals and covariation in triquetrum and hamate shape of extant primates using 3D geometric morphometrics. *J. Morphol.* **282**, 1382–1401. (doi:10.1002/jmor.21393)
 30. Monson TA, Armitage DW, Hlusko LJ. 2018 Using machine learning to classify extant apes and interpret the dental morphology of the chimpanzee-human last common ancestor. *PaleoBios* **35**, 1–20. (doi:10.5070/P9351040776)
 31. Püschel TA, Marcé-Nogué J, Gladman JT, Bobe R, Sellers WL. 2018 Inferring locomotor behaviours in Miocene New World monkeys using finite element analysis, geometric morphometrics and machine-learning classification techniques applied to talar morphology. *J. R. Soc. Interface* **15**, 20180520. (doi:10.1098/rsif.2018.0520)
 32. Alpaydin E. 2020 *Introduction to machine learning*, 4th edn. Cambridge, MA: The MIT Press.
 33. Bishop CM. 2016 *Pattern recognition and machine learning*. New York: Springer.
 34. Kotsiantis BS. 2006 Supervised machine learning: a review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **39**, 223–243.
 35. Miele V, Dussert G, Cucchi T, Renaud S. 2020 Deep learning for species identification of modern and fossil rodent molars [Preprint]. *Zoology*, 1–27. (doi:10.1101/2020.08.20.259176)
 36. Moclán A, Domínguez-García AC, Stoetzel E, Cucchi T, Sevilla P, Laplana C. 2023 Machine Learning interspecific identification of mouse first lower molars (genus *Mus* Linnaeus, 1758) and application to fossil remains from the Estrecho Cave (Spain). *Quat. Sci. Rev.* **299**, 107877. (doi:10.1016/j.quascirev.2022.107877)
 37. Torkzaban B, Kayvanjoo AH, Ardalan A, Mousavi S, Mariotti R, Baldoni L, Ebrahimie E, Ebrahimi M, Hosseini-Mazinani M. 2015 Machine Learning Based Classification of Microsatellite Variation: An Effective Approach for Phylogeographic Characterization of Olive Populations. *PLoS ONE* **10**, e0143465. (doi:10.1371/journal.pone.0143465)
 38. Begun DR. 2023 Catarrhine Origins and Evolution. In *A companion to biological anthropology* (ed. CS Larsen), pp. 381–399, 1st edn. Wiley.
 39. Stevens NJ et al. 2013 Palaeontological evidence for an Oligocene divergence between Old World monkeys and apes. *Nature* **497**, 611–614. (doi:10.1038/nature12161)
 40. McNulty KP, Begun DR, Kelley J, Manthi FK, Mbua EN. 2015 A systematic revision of *Proconsul* with the description of a new genus of early Miocene hominoid. *J. Hum. Evol.* **84**, 42–61. (doi:10.1016/j.jhevol.2015.03.009)
 41. Walker A. 1997 *Proconsul: Function and Phylogeny*. In *Function, phylogeny, and fossils* (eds DR Begun, CV Ward, MD Rose), pp. 209–224. US: Springer.
 42. Ward CV. 1998 *Afropithecus, Proconsul, and the Primitive Hominoid Skeleton*. In *Primate locomotion* (eds E Strasser, JG Fleagle, AL Rosenberger, HM McHenry), pp. 337–352. US: Springer.
 43. Nakatsukasa M. 2019 Miocene Ape Spinal Morphology: The Evolution of Orthogrady. In *Spinal evolution* (eds E Been, A Gómez-Olivencia, P Ann Kramer), pp. 73–96. Berlin, Germany: Springer International Publishing.
 44. Kibii JM, Churchill SE, Schmid P, Carlson KJ, Reed ND, de Ruiter DJ, Berger LR. 2011 A Partial Pelvis of *Australopithecus sediba*. *Science* **333**, 1407–1411. (doi:10.1126/science.1202521)
 45. Kivell TL, Kibii JM, Churchill SE, Schmid P, Berger LR. 2011 *Australopithecus sediba* Hand Demonstrates Mosaic Evolution of Locomotor and Manipulative Abilities. *Science* **333**, 1411–1417. (doi:10.1126/science.1202625)
 46. Williams SA, Ostrofsky KR, Frater N, Churchill SE, Schmid P, Berger LR. 2013 The Vertebral Column of *Australopithecus sediba*. *Science* **340**, 1232996. (doi:10.1126/science.1232996)
 47. Zipfel B, DeSilva JM, Kidd RS, Carlson KJ, Churchill SE, Berger LR. 2011 The Foot and Ankle of *Australopithecus sediba*. *Science* **333**, 1417–1420. (doi:10.1126/science.1202703)
 48. Churchill SE et al. 2013 The Upper Limb of *Australopithecus sediba*. *Science* **340**, 1233477. (doi:10.1126/science.1233477)
 49. Mongle CS, Strait DS, Grine FE. 2023 An updated analysis of hominin phylogeny with an emphasis on re-evaluating the phylogenetic relationships of *Australopithecus sediba*. *J. Hum. Evol.* **175**, 103311. (doi:10.1016/j.jhevol.2022.103311)
 50. Berger LR et al. 2015 *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. *ELife* **4**, e09560. (doi:10.7554/eLife.09560)
 51. Williams SA, García-Martínez D, Bastir M, Meyer MR, Nalla S, Hawks J, Schmid P, Churchill SE, Berger LR. 2017 The vertebrae and ribs of *Homo naledi*. *J. Hum. Evol.* **104**, 136–154. (doi:10.1016/j.jhevol.2016.11.003)
 52. Ocobock C, Lacy S, Niclou A. 2021 Between a rock and a cold place: Neanderthal biocultural cold adaptations. *Evol. Anthropol.: Issues News Rev.* **30**, 262–279. (doi:10.1002/evan.21894)
 53. Steegmann AT, Cerny FJ, Holliday TW. 2002 Neanderthal cold adaptation: Physiological and energetic factors. *Am. J. Hum. Biol.* **14**, 566–583. (doi:10.1002/ajhb.10070)
 54. Çep B, Schürch B, Münzel SC, Frick JA. 2021 Adaptive capacity and flexibility of the Neanderthals at Heidenschmiede (Swabian Jura) with regard to core reduction strategies. *PLoS ONE* **16**, e0257041. (doi:10.1371/journal.pone.0257041)
 55. Hofferker JF. 2018 The complexity of Neanderthal technology. *Proc. Natl Acad. Sci. USA* **115**, 1959–1961. (doi:10.1073/pnas.1800461115)
 56. Bardo A, Moncel MH, Dunmore CJ, Kivell TL, Poydebat E, Cornette R. 2020 The implications of thumb movements for Neanderthal and modern human manipulation. *Sci. Rep.* **10**, 19323. (doi:10.1038/s41598-020-75694-2)

57. Niewoehner WA. 2006 Neanderthal hands in their proper perspective. In *Neanderthals revisited: New approaches and perspectives* (eds JJ Hublin, K Harvati, T Harrison), pp. 157–190, Netherlands: Springer.
58. Karakostis FA, Hotz G, Tourloukis V, Harvati K. 2018 Evidence for precision grasping in Neandertal daily activities. *Sci. Adv.* **4**, eaat2369. (doi:10.1126/sciadv.aat2369)
59. Wiley D *et al.* 2005 Evolutionary Morphing. In IEEE Visualization Conference, Minneapolis, MN, 23–28 October 2005, pp. 431–438. New York, NY: IEEE.
60. Dryden I, Mardia K. 1998 *Statistical shape analysis*. London: Wiley.
61. Gower JC. 1975 Generalized procrustes analysis. *Psychometrika* **40**, 33–51. (doi:10.1007/BF02291478)
62. Hotelling H. 1933 Analysis of a complex of statistical variables into principal components. *J. Edu. Psychol.* **24**, 417–441. (doi:10.1037/h0071325)
63. Goodfellow I, Bengio Y, Courville A. 2016 *Deep learning*. Cambridge, MA: The MIT Press.
64. Van der Maaten L, Hinton G. 2008 Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.
65. McInnes L, Healy J, Melville J. 2018 *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. doi:10.48550/ARXIV.1802.03426
66. Buitinck L *et al.* 2013 *API design for machine learning software: Experiences from the scikit-learn project*. doi:10.48550/ARXIV.1309.0238
67. Pedregosa F *et al.* 2011 Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
68. Matarazzo S. 2013 Manual pressure distribution patterns of knuckle-walking apes. *Am. J. Phys. Anthropol.* **152**, 44–50. (doi:10.1002/ajpa.22325)
69. Wunderlich RE, Jungers WL. 2009 Manual digital pressures during knuckle-walking in chimpanzees (*Pan troglodytes*). *Am. J. Phys. Anthropol.* **139**, 394–403. (doi:10.1002/ajpa.20994)
70. Mackinnon J. 1974 The behaviour and ecology of wild orang-utans (*Pongo pygmaeus*). *Anim. Behav.* **22**, 3–74. (doi:10.1016/S0003-3472(74)80054-0)
71. Thorpe SKS, Crompton RH. 2006 Orangutan positional behavior and the nature of arboreal locomotion in Hominoidea. *Am. J. Phys. Anthropol.* **131**, 384–401. (doi:10.1002/ajpa.20422)
72. Tuttle R. 1988 Positional behaviour, adaptive complexes, and evolution. *Orangutan Biology*.
73. Reichard UH, Barelli C, Hirai H, Nowak MG. 2016 The Evolution of Gibbons and Siamang. In *Evolution of gibbons and siamang* (eds UH Reichard, C Barelli, H Hirai, MG Nowak), pp. 3–41. Berlin, Germany: Springer Science & Business Media.
74. Tuttle RH. 1969 Quantitative and Functional Studies on the Hands of the Anthropoidea. I. The Hominoidea. *J. Morphol.* **128**, 309–363. (doi:10.1002/jmor.1051280304)
75. Usherwood JR, Larson SG, Bertram JEA. 2003 Mechanisms of force and power production in unsteady ricochet brachiation. *Am. J. Phys. Anthropol.* **120**, 364–372. (doi:10.1002/ajpa.10133)
76. Bookstein FL. 1992 *Morphometric tools for landmark data*. Cambridge: Cambridge University Press.
77. Vanhoof MJM, Croquet B, De Groot I, Vereecke EE. 2023 Principal component and linear discriminant analyses for the classification of hominoid primate specimens based on bone shape data. Figshare. (doi:10.6084/m9.figshare.24057558)
78. Vanhoof MJM, Croquet B, De Groot I, Vereecke EE. 2023 Principal component and linear discriminant analyses for the classification of hominoid primate specimens based on bone shape data. Figshare. (doi:10.6084/m9.figshare.c6837169)