MORPHIX: Resolving sample identification bias in morphometrics analysis with a supervised machine learning package.

Authors: Nima Mohseni[1]*, Eran Elhaik[1]

Affiliations:

[1]Lund University, Department of Biology, Lund, Sweden.

*Correspondence to: nima.mohseni@biol.lu.se

Keywords: Modern geometric morphometrics, Principal Component Analysis (PCA), Generalized Procrustes Analysis (GPA), Morphometric landmarks, Papionin, MORPHIX.

Abstract

Evolutionary biologists, primarily anatomists and ontogenists, employ modern geometric morphometrics to quantitatively analyse physical forms (e.g., skull morphology) and explore relationships, variations, and differences between samples and taxa using landmark coordinates. The standard approach comprises two steps, Generalised Procrustes Analysis (GPA) followed by Principal Component Analysis (PCA). PCA projects the superimposed data produced by GPA onto a set of uncorrelated variables, which can be visualised on scatterplots and used to draw phenetic, evolutionary, and ontogenetic conclusions. Recently, the use of PCA in genetic studies has been challenged. Due to PCA's central role in morphometrics, we sought to evaluate the standard approach and claims based on PCA outcomes. To test PCA's accuracy, robustness, and reproducibility using benchmark data of the crania of five papionin genera, we developed MORPHIX, a Python package containing the necessary tools for processing superimposed landmark data with classifier and outlier detection methods, which can be further visualised using various plots. We discuss the case of *Homo Nesher Ramla*, an archaic human with a questionable taxonomy. We found that PCA outcomes are artefacts of the input data and are neither reliable, robust, nor reproducible as field members may assume and that supervised machine learning classifiers are more accurate both for classification and detecting new taxa. Our findings raise concerns about PCA-based findings in 18,000 to 32,900 studies. Our work can be used to evaluate prior and novel claims concerning the origins and relatedness of inter- and intra-species and improve phylogenetic and taxonomic reconstructions.

Introduction

Morphometrics is one of the oldest approaches of investigation in biology used to address a variety of questions ranging from how shapes vary over time (evolution, growth, and development) to how they vary inter- and intra-species. It has been pursued by graphical and computational means, based on millennia-old geometric principles, since the European Renaissance (36, 73). Even though the quantitative analysis of morphological data has been long pursued (29, 107), it was not until the mid-20th century that the mathematical foundations of forms matured and in the mid-1990s (17, 35, 45, 102) that relevant graphic and statistical tools were made available that modern geometric morphometrics became an essential biological tool dominating shape studies. A major contribution to the field was made by Sokal and Sneath's Principles of Numerical Taxonomy (105) book, which developed the field of phenetics and introduced quantitative methods to address questions of classification (see also review by Sneath (104)).

Morphology is a critical source of information concerning an object's shape and size and is used throughout the natural sciences, materials sciences, and engineering. Concerning organisms, morphometrics is the main technique for answering questions about form- (shape and size) relatedness in physical anatomy. It has many applications in evolutionary biology, anthropology, palaeontology/palaeoanthropology, and even forensics (115). Morphometrics is the leading technique in studying variations, differences, and evolutionary dynamics and relationships in forms, which is suggestive of ancestry and ontogeny. Members of the field assign morphometrics ancestral and phynetic interpretations and apply it to gain ontogenetic insight into growth and development (8, 28, 67, 88, 114).

Modern geometric morphometrics is typically applied to the coordinate data of two- or three-dimensional landmarks comprising anatomical landmarks, mostly sutures or processes identified as biological homologues across samples, or geometric landmarks, used to capture positional and proportional information within a shape without relying on biological correspondence such as most distal (farthest) and the most proximal (closest) points to a specific reference point or structure (115). In addition to landmarks, semi-landmarks are also used to capture and analyse shape variation in biological structures (9, 36). While landmarks provide crucial information about the structure's overall shape, semi-landmarks are points placed along curves or outlines that require additional information to define their precise locations. Semi-landmarks are often used to capture fine-scale shape variation that cannot be adequately represented by landmarks alone, such as curves or surfaces. Semi-landmarks are heavily relied upon as the source of shape information to break the continuity of regions in the specimen without clearly identifiable landmarks (53). Semi-landmarks are typically aligned based on their relative positions to landmarks, allowing for the comprehensive analysis of shape changes and deformations within complex structures (36). Determining the sufficient number of measurements and data points for a valid morphometric analysis is older than modern geometric morphometrics (87). Unsurprisingly, the use of semi-landmarks is controversial. For instance, Bardua et al. (9) claim that high-density sliding semi-landmark approaches are beneficial over landmark-only studies, while

Cardini (22) advises caution about potential biases and subsequent inaccuracies in high-density morphometric analyses.

Geometric morphometrics analysis uses different approaches that vary greatly in detail, while most maintain similar conventional core elements (8, 36, 41, 67, 77). The standard approach to processing landmark data comprises two major steps: Generalised Procrustes Analysis (GPA) followed by Principal Component Analysis (PCA) (Figure 1). GPA superimposes the landmark coordinates by reducing the shape-independent variations in samples. PCA is a multivariate statistical analysis that projects the data onto a set of uncorrelated variables known as principal components (PCs), each explaining a reduced proportion of the variance in the data (115). After GPA, all landmark data are in the shape space manifold, meaning that each landmark configuration is a point in a high-dimensional space, where the dimensions represent the coordinates of the landmarks after GPA. However, this high-dimensional space is often too complex to visualise and analyse directly. Thereby, a tangent space projection with PCA is applied to move the points from the shape space manifold to a linearised space. In practice, only the variations along the first two or three PCs are discussed, and the rest are ignored (30, 48, 49, 53, 57, 111, 114). These PCs are then used to visualise the samples in colourful two-dimensional scatterplots to evaluate shape relationships. Typically, researchers expect biologically related shapes to cluster more closely than unrelated ones. Originally, these plots were simply developed as a tool to visually explore the structure of multivariate data in a low-dimensional subspace that summarizes the bulk of shape variation along a few components (15, 36), but the interpretation of the scatterplot patterns spread in terms of origins, relatedness, evolution, gene flow, speciation, and phenotypic/genotypic variation of the samples and taxa (8, 30, 41, 42, 48, 49, 53, 57, 59, 67, 84, 111, 114). It is also commonly assumed that the proximity of the samples is evidence of relatedness and shared evolutionary history (48, 79, 114), whereas their absence indicates the opposite. Moreover, the variation/covariation along some PCs is considered to reflect specific morphological traits (16, 53, 115). Needless to say all these interpretations are subjective, and while some authors advised caution in interpreting the results (79), biological conclusions are commonly drawn nonetheless. For example, claims that one PC reflects ‘population effects’ and another ‘sex effects’ are common (48, 84), although the PCs are statistical manifestations agnostic to the data (86).

One popular application of modern geometric morphometrics is the classification of newly discovered partial pieces of human skulls with respect to the human phylogenetic tree by visually observing the *clustering* and *proximity* of the specimens with each other in PC plots (48, 53, 114) in conjunction with other analyses. Both *clusters* and *distances* between samples and groups in these plots are determined subjectively. Similarly, plots produced by different PCs might yield inconsistent results, and authors choose plots based on subjective and inconsistent criteria. For instance, in the case of *homo Nesher Ramla* (NR) bones found in Israel, Hershkovitz et al. (53) compared the osteological remains with those of other hominins using PC1, PC2, and PC3 scatterplots to investigate their origins (Figures 2, 3, S1, and S17 in (53)). The plots produced conflicting results. Excluded from that study was the lower molar PC2-PC3 plot shown in our Figure 2C. In this plot, *homo* NR clusters with Neanderthals, which was inconsistent with the PC1-PC2 (Figures 2A, Figure S17 A in (53))

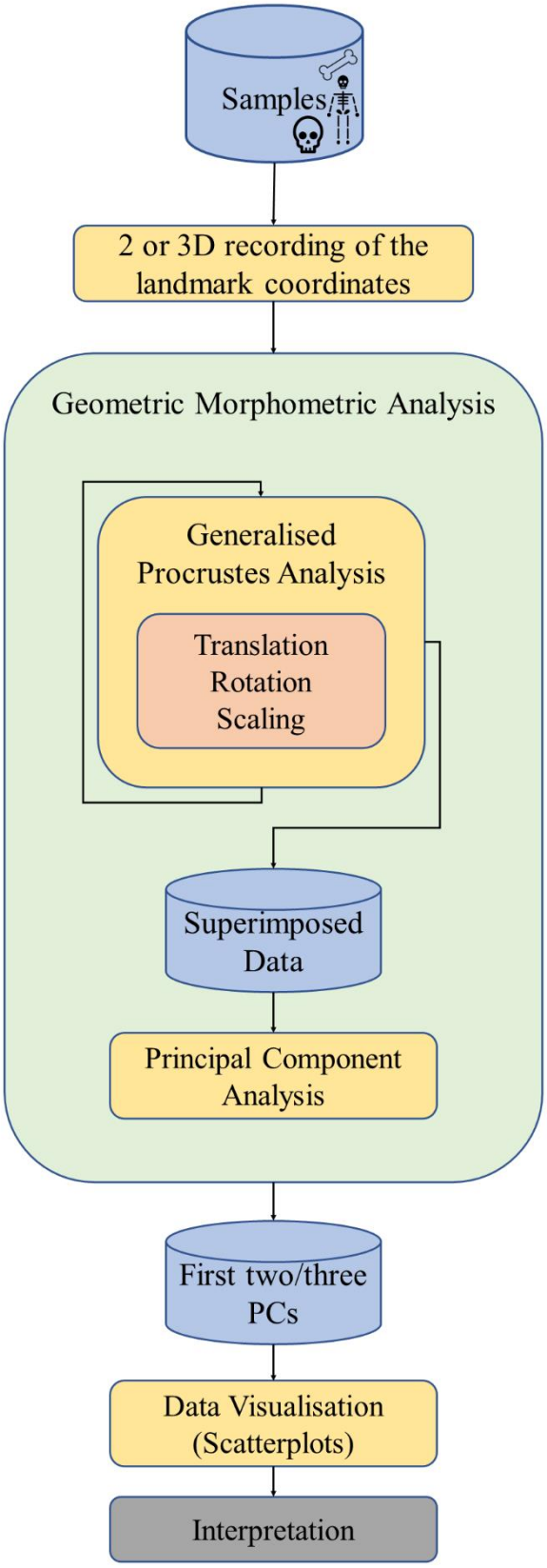
and PC1-PC3 (Figures 2B, Figure S17 B in (53)) plots, even though shape variation along both PC2 and PC3 was relied upon and interpreted in the text (53). Selecting a few PC dimensions and excluding the conflicting results, the authors concluded that *homo* NR was one of the last survivors of an as-of-yet-unknown population contributing to the *Neanderthals* and East Asian *Homo*. A similar methodology was used by Wu et al. (2023) for the Hualongdong sample 6 (HLD 6), comprising a mandible and partial crania (114). The authors analysed the clustering, overlap, and the *position* of HLD 6 relative to the samples in PC scatterplots (114). They reported that the overall mandibular shape with size of HLD 6 has an intermediary position among five other overlapping groups in PC1-3 scatterplot. Moreover, in PC1-2 scatterplot of the shape of the symphysis, HLD 6 was located in the middle of the overlap of Neanderthals and the Late Pleistocene modern humans in PC2. Along the PC1 axis HLD 6 was located at the edge of the range of variation of the Middle Pleistocene hominins and recent modern humans (114). Based on these observations the authors claimed that HLD 6 could belong to a new branch in human family tree.

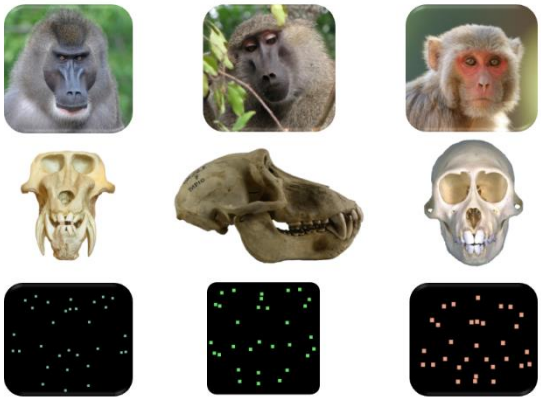
Criticism of PCA in morphometrics is not new. Rohlf's (98) already observed that distances between close neighbours are not well represented by PCA ordination. However, due to the lack of satisfactory methods for deriving clusters from similarity matrices (Sokal and Sneath 1963 (105), Pg. 252), and the premise of using a quantitative method, PCA usage continued and spread despite its stated limitations. No sooner had PCA been established as the main element of geometric morphometric analysis than even the criticism from its pioneers was dismissed (17). More recent concerns about the adequacy of PCA (5) and its extensions (17, 23) in morphometrics and phylogenetics were also published. It has been shown that between-group PCA can inflate and even create differences and patterns where they do not exist (23) and that common PCA (multi-group PCA) (39) introduces systematic statistical biases in morphological trait analysis, even when the data are in *very good condition* (13). Furthermore, through evaluating numerous cases using population genomics data, it has been shown that projections from high to low dimensional space shown in colourful PC scatterplots are artefacts of the input data and can produce a multitude of outcomes based on variation in the input data (38). These artefacts are not limited to the nearest neighbours, as argued by Rohlf's (98), but also include distances between the major clusters in subsets of PCs which are usually limited to the first two. PCA's prominent role in morphometrics analyses and, more generally, physical anthropology is inconsistent with the recent criticisms, which raises concerns regarding its validity and, consequently, the value of the results reported in the literature. If PCA of morphometric landmark data produces biased results, a large corpus of the literature should be reevaluated. Conservatively, we estimate that, as of August 2023, between 18,000 to 32,900 landmark-based geometric morphometric studies employed PCA (see Methods).

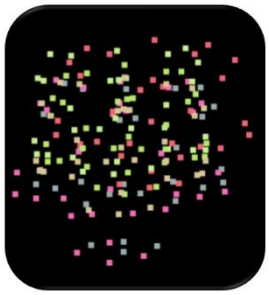
Here, we verified that this indeed is the case. We also evaluated the argument that a high Explained Variance (EV) could be counted as a measure of reliability (36). We found no association between high EV amounts and the subjectiveness of the results.

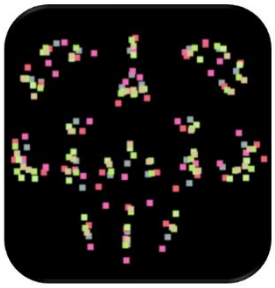
As an alternative to the PCA step in geometric morphometrics analysis, we evaluated leading supervised learning classifiers and outlier detection methods using the original and

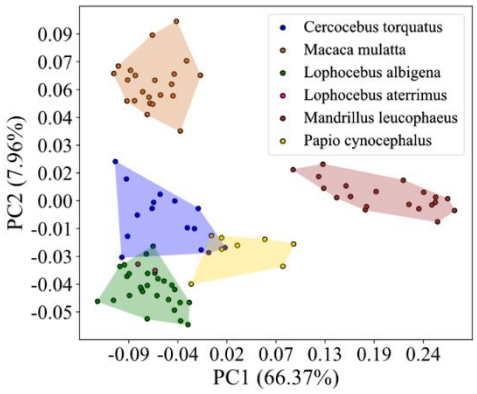
altered three-dimensional landmark datasets. We found that the nearest neighbour classifier and local outlier factor outperformed all other methods in classification and outlier detection. Concerningly, we also observed that the accuracy of linear discriminant analysis (LDA), a classifier already utilised in geometric morphometrics, was strongly influenced by the selection of the number of principal components (PCs) used as input. We developed MORPHIX, a Python package that contains the necessary tools for processing superimposed landmark data with any classifier and outlier detection method of choice, which can be further visualised using various plots. MORPHIX is freely and publicly available at https://github.com/NimaMohseni/Morphometric-Analysis and includes numerous detailed examples.


Samples
2 or 3D recording of the landmark coordinates
Geometric Morphometric Analysis
Generalised Procrustes Analysis
Translation
Rotation
Scaling
Superimposed Data
Principal Component Analysis
First two/three PCs
Data Visualisation (Scatterplots)
Interpretation








Cercocebus torquatus
Macaca mulatta
Lophocebus albigena
Lophocebus aterrimus
Mandrillus leucophaeus
Papio cynocephalus
PC2 (7.96%)
PC1 (66.37%)
0.09
0.07
0.06
0.04
0.03
0.02
0.00
-0.01
-0.03
-0.04
-0.05
-0.09 -0.04 0.02 0.07 0.13 0.19 0.24

182

183

Figure 1: A flow chart of conventional modern geometric morphometric analysis. The standard pipeline comprises several steps (left). First, a set of landmarks are chosen based on distinct anatomical features believed to be equivalent in some sense (e.g., biologically homologous) among all the specimens (papionins), and their cartesian coordinates are scanned and recorded. Next, the landmark data are superimposed using GPA through translation, rotation, and scaling. Next, after performing PCA, the scatterplots of the first two or three PCs are used to assess patterns of shape variation based on the distribution of the observations. The outcome of the pipeline (right) are depicted for the benchmark data. For brevity, only three taxa are shown on the top right.

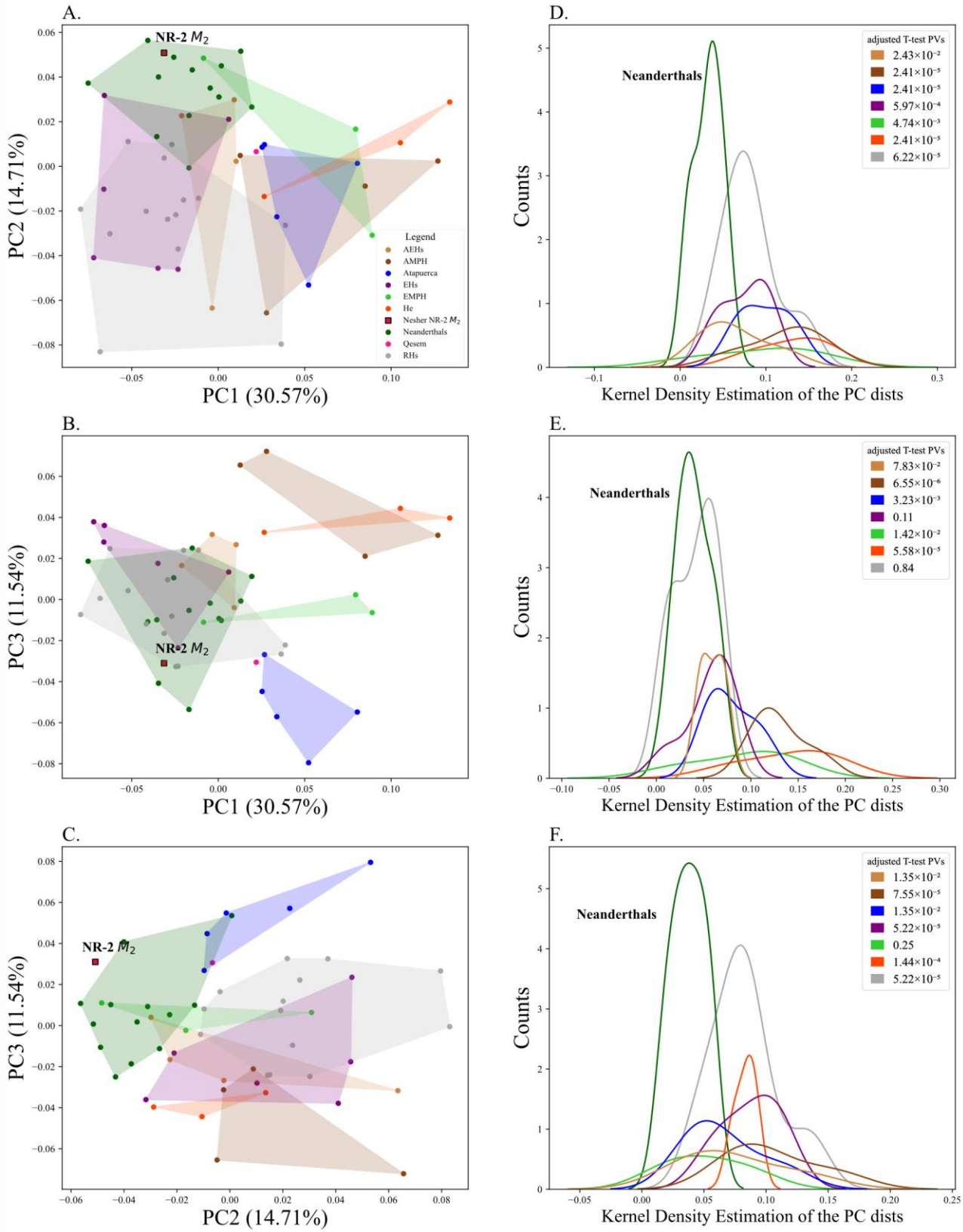

A.
NR-2 M_2
PC2 (14.71%)
PC1 (30.57%)
Legend
AEHs
AMPH
Atapuerca
EHs
EMPH
He
Nesher NR-2 M_2
Neanderthals
Qesem
RHs
B.
NR-2 M_2
PC3 (11.54%)
PC1 (30.57%)
C.
NR-2 M_2
PC3 (11.54%)
PC2 (14.71%)
D.
Neanderthals
adjusted T-test PVs
2.43×10^{-2}
2.41×10^{-5}
2.41×10^{-5}
5.97×10^{-4}
4.74×10^{-3}
2.41×10^{-5}
6.22×10^{-5}
Counts
Kernel Density Estimation of the PC dists
E.
Neanderthals
adjusted T-test PVs
7.83×10^{-2}
6.55×10^{-6}
3.23×10^{-3}
0.11
1.42×10^{-2}
5.58×10^{-5}
0.84
Counts
Kernel Density Estimation of the PC dists
F.
Neanderthals
adjusted T-test PVs
1.35×10^{-2}
7.55×10^{-5}
1.35×10^{-2}
5.22×10^{-5}
0.25
1.44×10^{-4}
5.22×10^{-5}
Counts
Kernel Density Estimation of the PC dists

Figure 2: PC scatterplots of the shape space of the lower molar of *Homo* NR2 and other *Hominins* (left). The analysis combines the enamel–dentine junction (EDJ) and the cemento-enamel junction (CEJ). The Kernel Density Estimation (KDE) plots the distribution of the distances of each *Homo* species from *Homo NR* in the space of the corresponding PCs (right). The legends of the right sub-figures present the adjusted (Benjamini/Hochberg) p-values of the T-test for the means of the distances of each *Hominin* population and *Neanderthals* from *Homo NR*. Legend guide: African Early *Homo sapiens* (AEHs), African *Middle Pleistocene Homo* (MPH), Early *Homo sapiens* (EHs), European *Middle Pleistocene Homo* (EMPH), *Homo erectus* (He), Recent *Homo sapiens* (RHs). For consistency, the colours of the species are similar to those of Hershkovitz et al. (53) PC plots.

Results

We compared the classification accuracy of PC scatterplots and eight ML tools for the complete dataset of the crania of 94 papionin adults belonging to six species (*Cercocebus torquatus, Papio cynocephalus, Lophocebus albigena, Lophocebus aterrimus, Macaca mulatta,* and *Mandrillus leucophaeus*) with 31 three-dimensional landmarks (the benchmark). Each of the four cases of altered datasets includes two examples of species removal, sample removal, landmark removal, and their combination. If PC-based shape inferences are valid (30, 41, 48, 53, 67) and if PC scatterplots were valid tools for classifying samples and investigating phylogeny (8, 20, 33, 114), as many members of the field assume, different species should not overlap in PC space, whereas related species who share recent evolutionary origins are expected to cluster, overlap, or exhibit short distances from each other. It is also widely assumed that the distribution of samples and groups along PCs accurately reflects relatedness in certain morphological traits (48, 53).

The benchmark data

First, we carried out traditional geometric morphometric analyses using GPA and PCA. In the second part of our study, we reanalysed the data using alternative classifiers. Even without any alterations, the PC-based clusters are inconsistent with our taxonomical knowledge of the papionin taxa. It is difficult to see how these incongruent PC clusters can be used to facilitate the inference of an accurate phylogeny, as commonly practiced (e.g., (92)). In all PC plots (Figures 3A-C), at least two groups (not necessarily the same ones) overlap. For example, in Figure 3A *Cercocebus torquatus (blue)* overlaps with both *Papio cynocephalus (yellow)* and *Lophocebus albigena (green)*, in Figure 3B *Macaca mulatta (orange), Cercocebus torquatus (blue)* and *Lophocebus albigena (green)*, and in Figure 3C *Mandrillus leucophaeus (red)* and *Cercocebus torquatus (blue)* overlap; consequently, and since the results can be expected to continue to change as higher dimensions are explored, it is unclear whether these overlapping samples should be treated as misclassification as there is room for other evolutionary interpretations. Based on the proximity of the samples to their neighbours, misinterpretations can range from phenotypic variation (67, 93), admixture and hybridization (41), to speciation (8) and gene flow (53) and phenetic evolutionary trends (114). Observations can be further explained in different manners. For example, interpreting the results as gene flow might also require explaining its direction; this was the case of *homo NR*, whom as Hershkovits et al. claimed (53), could have belonged to a population ancestral to the *Neanderthals* and East Asian *homo*, while Rightmire argued (59) that *homo NR* could have belonged to a group of European *Neanderthals* who had migrated to the area based on the same findings. Considering the *Cercocebus torquatus (blue)* samples that fall near or within *Lophocebus albigena (green)* in Figure 3A; if they were newfound samples of unknown origins, they could be misinterpreted as being *Lophocebus albigena (green)*. Additionally, several different gene flow scenarios could also be proposed, such as the

Lophocebus albigena (green)-proximate *Cercocebus torquatus* samples being from an as-of-yet unknown origin contributing to the *Cercocebus torquatus* (blue) and *Lophocebus albigena* (green) or being a hybrid of them.

It is customary to interpret the sum of the explained variance of the PCs as a measure to evaluate accuracy (30, 111). Despite occasional advised caution (26, 78), higher percentages of explained variance are still interpreted as reflecting higher accuracy (64, 71). We first show that this is not always the case. In Figure 3A, PCs can have the highest EV (Figure 3A, PC1+PC2=74%), yet the overlap of taxa *Cercocebus torquatus (blue)* and *Papio cynocephalus (yellow)* would necessarily lead to the aforementioned misinterpretations. By contrast, in Figure 3C with the lowest EV (EV=12%), the PCs well separate *Cercocebus torquatus (blue)* and *Papio cynocephalus (yellow)*. In other words, by pure chance, here, the PC plot with the lesser variance depicts morphometrical relatedness and distances more adequately and accurately than the plot explaining more than six times shape variation.

Next, we investigated the classification *accuracy* of the alternative methods in classifying superimposed landmark data. We calculated the accuracy as the portion of the correctly classified samples to the total number of samples and *balanced accuracy* as the arithmetic mean of sensitivity and specificity, which can be generalised to multiclass classification under a one versus rest transformation (19, 61). We implemented eight ML methods that could effectively classify the samples in a five-fold Cross-Validation (CV) scheme, including Linear Discriminant Analysis (LDA), using the first two and 10 PCs of the superimposed landmarks. To assess the ability of PC scatterplots to classify samples (8, 20, 33), we defined accuracy as the ratio of the non-overlapping samples to all samples in each case. Since, in case of overlap, false negatives cannot be defined, no balanced accuracy could be calculated for PC plots. It should be emphasised that LDA was solely used as a classifier and not a dimensionality reduction method.

Three ML classifiers, including the Logistic regression (LR), Gaussian Process (GP), and Support Vector Classifier (SVC), as well as the Nearest Neighbours (NN) (Figure 3F) had balanced accuracies higher than 96%. The three classifiers from the decision tree family had lower performances. LDA had an accuracy of 93% and 98% for the first two and ten PCs, respectively. We visualised the samples and the decision boundaries of the most effective classifier using a t-distributed Stochastic Neighbour Embedding (t-SNE) (110). However, we note that the confusion matrices (supplementary Figures S1-9) and Procrustes distances dendrograms created using agglomerative clustering (82), should be the main infographic source. Table 1 presents the accuracy and balanced accuracy of each classifier.

Table 1: The accuracy and balanced accuracy of the eight classifiers. The accuracy was evaluated in a five-fold stratified CV on the benchmark dataset and the six alteration cases. The numbers in parentheses are the balanced accuracies.

	NN	**LR**	**GP**	**SVC**	**RF**	**ET**	**XGB**	**LDA(2)**	**LDA(10)**	**PC1-2**	**PC1-3**	**PC2-3**
Benchmark	0.97	0.97	0.98	0.97	0.92	0.90	0.92	0.93	0.98	0.93	0.42	0.79

data	(0.97)	(0.96)	(0.98)	(0.97)	(0.88)	(0.86)	(0.91)	(0.90)	0.98	-	-	-
Removing *Green*	1 (1)	0.98 (0.96)	1 (1)	1 (1)	0.96 (0.95)	0.96 (0.95)	0.95 (0.92)	0.92 (0.92)	1 (1)	0.97 -	0.98 -	0.92 -
Removing *Orange*	1 (1)	1 (1)	1 (1)	1 (1)	0.92 (0.90)	0.94 (0.93)	0.9 (0.88)	0.8 (0.79)	1 (1)	0.62 -	0.75 -	0.37 -
***Green* sample removal**	0.98 (0.98)	0.97 (0.96)	0.98 (0.98)	0.98 (0.98)	0.97 (0.97)	0.98 (0.98)	0.95 (0.93)	0.91 (1)	0.88 (1)	0.91	0.5	0.96
***Yellow* and *Blue* sample removal**	0.97 (0.94)	0.93 (0.86)	0.97 (0.94)	0.98 (0.97)	0.95 (0.91)	0.95 (0.91)	0.90 (0.85)	0.91 (0.88)	1 (1)	0.93	0.54	1
landmark removal (23, 24, 25, 26 and 27)	0.98 (0.98)	0.96 (0.94)	0.98 (0.98)	0.97 (0.97)	0.95 (0.93)	0.93 (0.90)	0.90 (0.84)	0.91 (0.90)	1 (1)	0.90 -	0.38 -	0.78 -
landmark removal (11, 12, 16, 17, 18, 30 and 31)	0.95 (0.95)	0.93 (0.91)	0.95 (0.95)	0.94 (0.93)	0.90 (0.88)	0.91 (0.89)	0.85 (0.81)	0.93 (0.90)	0.96 (0.96)	0.87 -	0.48 -	0.76 -
Extreme case 1	0.71 (0.57)	0.71 (0.54)	0.64 (0.46)	0.67 (0.52)	0.71 (0.54)	0.75 (0.59)	0.64 (0.52)	0.71 (0.59)	0.71 (0.63)	- -	- -	- -
Extreme case 2	61 (0.51)	0.5 (0.38)	0.58 (0.42)	0.52 (0.41)	0.6 (0.54)	0.55 (0.47)	0.58 (0.43)	0.41 (0.34)	0.47 (0.42)	- -	- -	- -

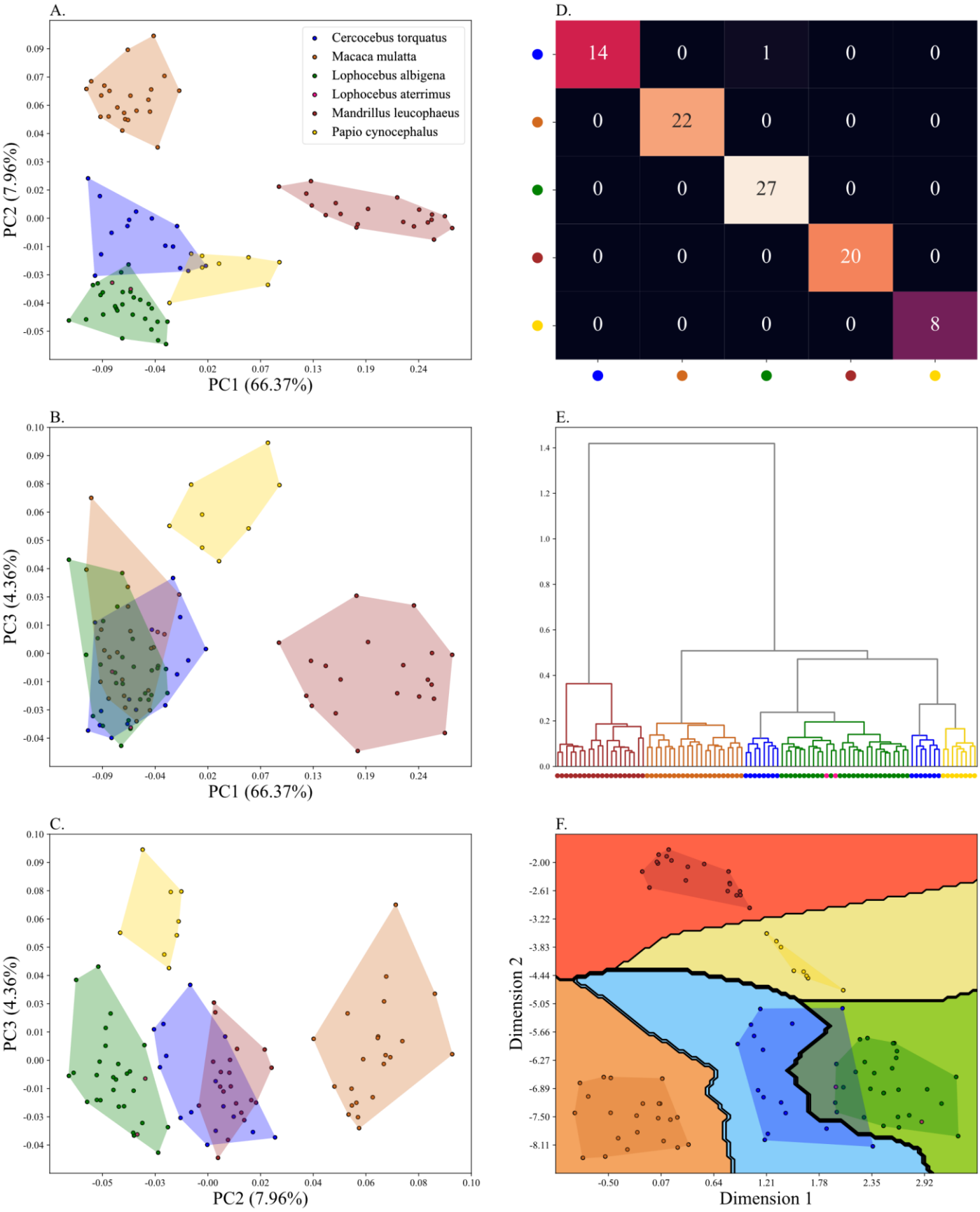


Figure 3: Plots depicting the papionin benchmark data. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) Confusion matrix of a 2NN classifier with 5-fold CV, E) Dendrogram of Procrustes distances of the samples using agglomerative clustering, F) Decision boundary of the 2NN classifier after t-SNE embedding with 5-fold CV.

The case of missing taxa

Paleontological findings often differ in quality and completeness, which poses profound challenges to morphometric analyses. Rarely all related taxa are represented; thereby, only partial datasets are inevitably used (48, 53, 114). Consequently, the extent of the bias created by this incomplete sampling is unclear. To study how PCA and alternative classifiers are affected by non-representative partial sampling, we excluded one taxon at a time from the dataset, applied the pipeline (Figure 1) and classifiers, and compared the outcome to the benchmark (Figure 3).

As before, we present two instances where the removal of *Lophocebus albigena (green)* (Figure 4) and *Macaca mulatta (orange)* (Figure 5) resulted in profound deviations compared to the benchmark. In both cases, the distances within and among groups in the PC plots changed dramatically. Removing the *Lophocebus albigena (green)* results in three instances of taxa overlap (separated in the benchmark). *Macaca mulatta (orange)* overlaps with *Cercocebus torquatus (blue)* in Figure 4A and Figure 4B and with *Mandrillus leucophaeus (red)* in Figure 4C. This would promote the misclassification of several samples. For example, the single *Cercocebus torquatus (blue)* sample that falls inside the convex hull of the *Macaca mulatta (orange)* (Figure 4A) previously clustered with other samples from its own taxon (Figure 3A); now it clusters with *Macaca mulatta (orange)*. Misinterpretations of evolutionary processes, such as gene flow and phenotypic variations, can also be made regarding this sample. This *Cercocebus torquatus (blue)* sample could be claimed to belong to a taxon related to the *Cercocebus torquatus (blue)* and *Macaca mulatta (orange)* experiencing gene flow. It could also be narrowed down to being their hybrid. Another notable example is the supposed relationships between the *Cercocebus torquatus (blue)*, *Papio cynocephalus (yellow)*, and *Macaca mulatta (orange)* perceived from different plots. Figure 4A implies an old event of admixture between the *Papio cynocephalus (yellow)* and *Macaca mulatta (orange)*, which resulted in the *Cercocebus torquatus (blue)* with the appearance of breeding barriers separating it from both. Such a trend cannot be observed in the benchmark; By contrast, an opposite trend is obvious in Figure 3A, which implies a more recent event of speciation between the *Cercocebus torquatus (blue)* and *Papio cynocephalus (yellow)*, both being distanced from *Macaca mulatta (orange)*.

The changes in PC plots are even more profound in the case of the *Macaca mulatta (orange)* removal. Here the *Cercocebus torquatus (blue)* and *Lophocebus albigena (green)* clusters almost overlap entirely across all the PC plots (Figures 5A -5C), whereas their overlap was minimal in the benchmark plot (Figure 3A). While the benchmark implies that the evolutionary distance between them is considerably high due to the accumulation of genetic differences to the point of speciation, Figure 5A is indicative of much higher genetic relatedness. The presence of only five *Lophocebus albigena (green)* samples not overlapping *Cercocebus torquatus (blue)* (Figure 5A) may be perceived as uniqueness within the genera; at the same time, the possibility of a speciation event cannot be ruled out. Considering the distribution of *Cercocebus torquatus (blue)* and *Lophocebus albigena (green)* over PC3, Figures 5B and 5C may imply an opposite trend, that the genetic divergence between the two

is high, which can be indicative of penultimate states of speciation, especially if based on the location of the excavations, geographic isolation, and breeding barriers could be proposed. We note that the EV of these plots (Figures 4A and 5A) is higher than that of the benchmark (Figure 3A) since there are fewer specimens, implying a lower dimensional shape space, supposedly lending them more credibility by the widely assumed interpretation.

The overlap and separateness of taxa over PCs are frequently associated with morphological traits and trends. Methods useful for the study of morphometrics are expected to maintain an accurate representation of the relationships of the studied taxa and remain, at least largely, unaffected by the addition or removal of a taxon. In the case of PCA, we observed the exact opposite; while in the benchmark, the *Macaca mulatta (orange)* and *Mandrillus leucophaeus (red)* are completely separated by PC2 (Figures 3A and 2C), after the removal of *Lophocebus albigena (green)*, *Macaca mulatta (orange)* overlaps considerably with *Mandrillus leucophaeus (red)* (Figures 5A and 5C). Thereby, it could either be claimed that PC2 *separates well* between the *Macaca mulatta (orange)* and *Mandrillus leucophaeus (red)* (Figures 3A to C) or that it is *taxonomically noninformative* about them (Figures 5B and 5C). These results are unsurprising because adding or removing specimens to PCA may result in a rotation of the PCs and reflect completely different portions and aspects of shape variation with different interpretations of the new space. The extent of this bias, shown here for the first time, demonstrates that PCA is a hazardous and unpredictable tool for studying morphological trends.

Concerning the use of EV as a measure of accuracy or reliability, it is notable that Figures 4A and 5A each explain more than 75% of the variance in shape (very close to that of the benchmark). Yet, the distances between specimens/species and, consequently, interpretations based on them greatly diverge. Moreover, classifying the samples using PC plots would lead to the misclassification of many in the case of *Macaca mulatta (*orange) removal. For instance, in Figure 5A (EV = 78%), at least 20 samples would potentially be misclassified.

Of the eight classifiers applied to the data, three (NN, GP, and SVC) and four (NN, LR, GP, and SVC) reached perfect accuracy (100%) in classifying the samples in a five-fold stratified CV in the first and second test, respectively. Similar to the benchmark, NN, LR, GP, and SVC show superior performance to the tree-based methods in the two cases of taxon removal. The LDA of the first two PCs had a considerably lower performance than other classifiers (accuracy = 80% in the second case), while the LDA of the first 10 PCs had perfect accuracy (100%) in both tests. The t-SNE plot creates well-separated clusters (Figures 4D and 4D), and the simple 2NN classifier has no difficulty differentiating groups (Figures 4E and 4F and 5E and 4F).

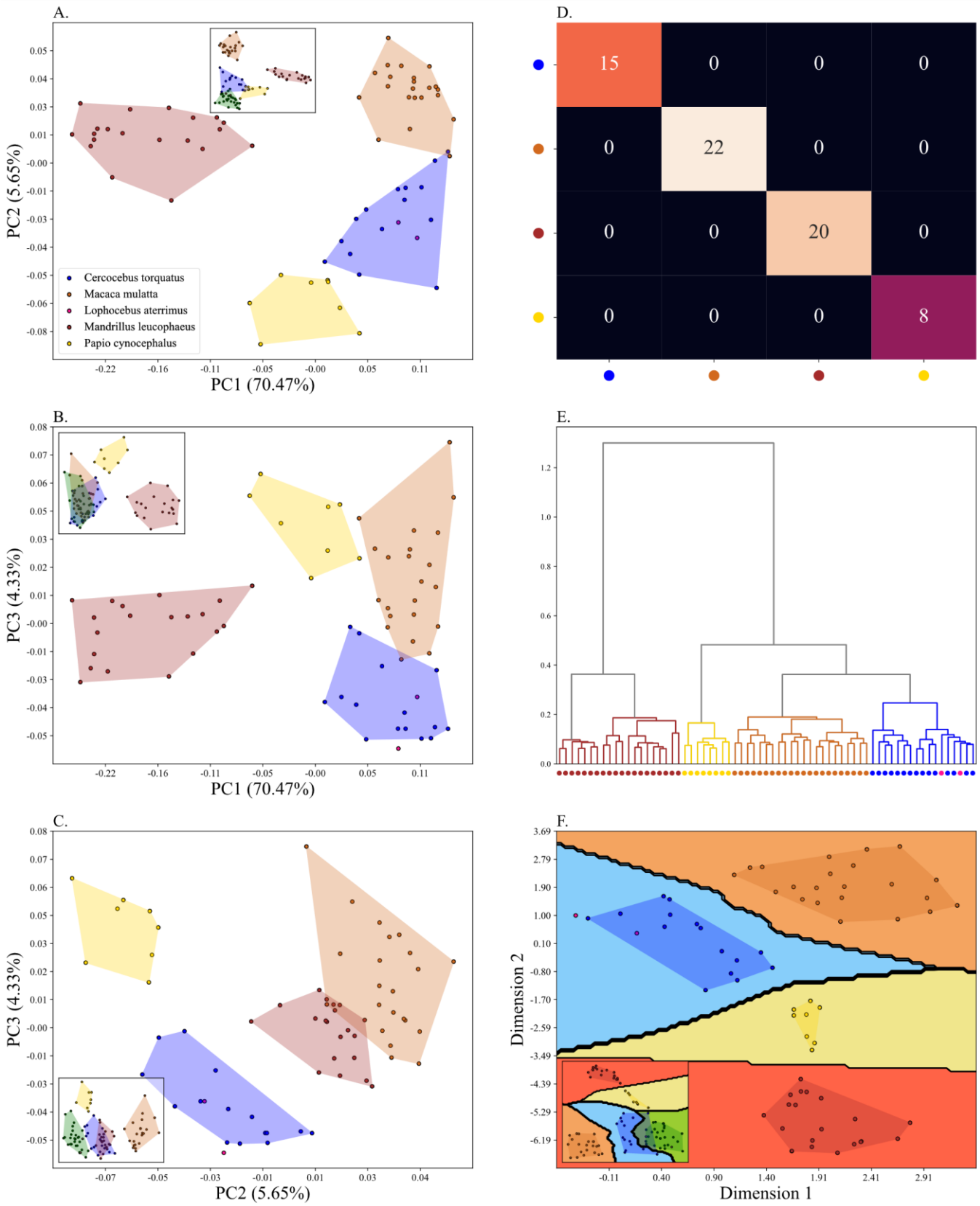


Figure 4: Plots depicting the papionin data after the removal of the Lophocebus albigena (green) taxon. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot, E) Dendrogram of Procrustes distances of the samples using agglomerative clustering, F) Decision boundary of

the 2NN classifier after t-SNE embedding with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).

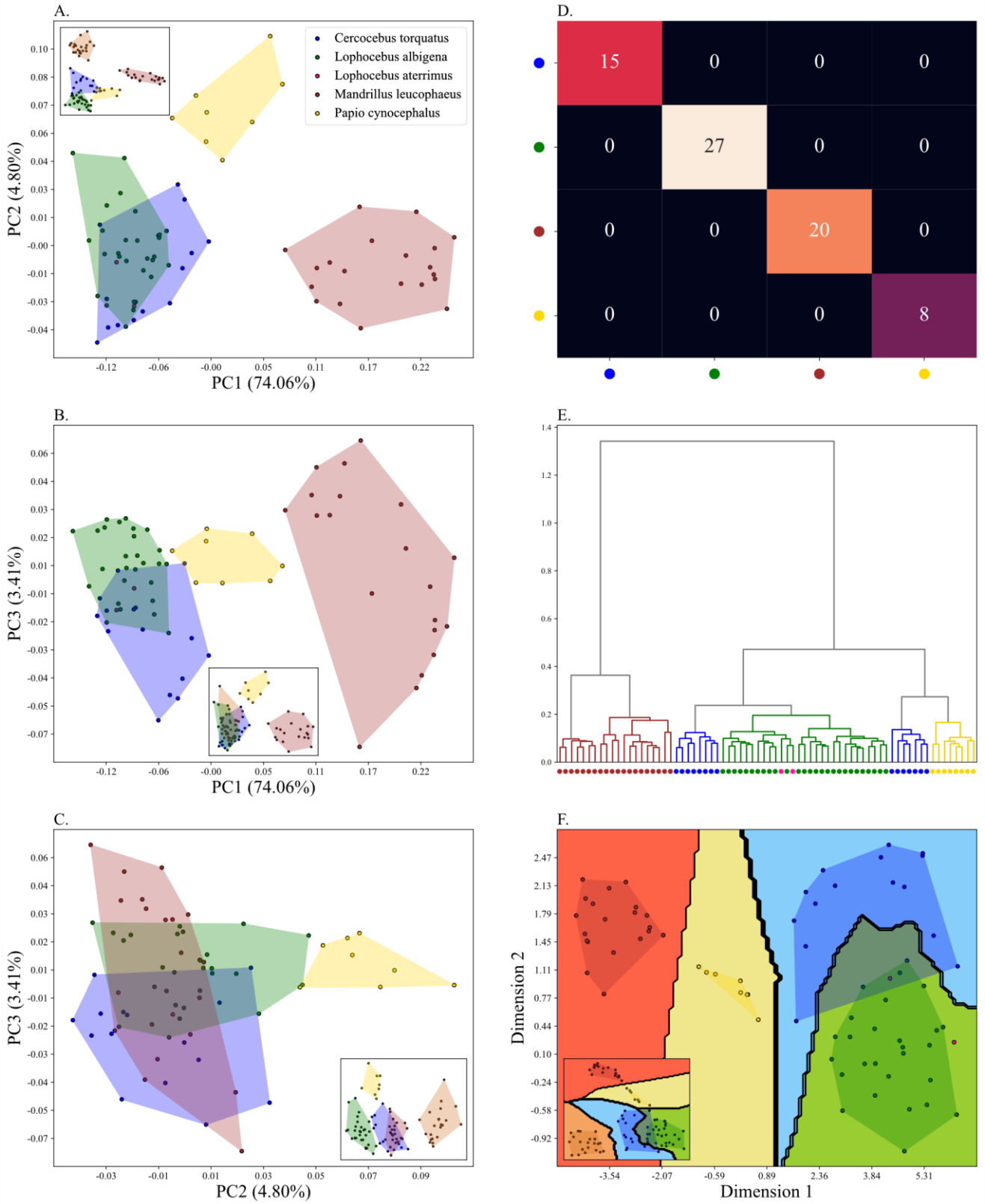


Figure 5: Plots depicting the papionin data after removing the Macaca mulatta (orange) taxon. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot, E) Dendrogram of Procrustes distances of the samples using agglomerative clustering, F) Decision boundary of the 2NN classifier after t-SNE embedding with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).

The case of missing samples

To examine the effect of missing specimens on PCA and the alternative methods, we removed 11 and 19 samples from one (first test) and two (second test) taxa, respectively. As expected, we found that the shape and the area of the convex hull change with respect to the benchmark, specifically, that of the taxon(a) from which the samples were removed. This reduction is evident in all the PC scatterplots (Figures 6A-C and 7A-C). More concerningly, we observed changes in the clusters that may lead to misclassification of the unaltered taxa. For example, we note one of the two *Lophocebus aterrimus (pink)* samples in Figure 6A and 6C and 7A and 6C, which clustered with *Lophocebus albigena (green)* close to the *Cercocebus torquatus (blue)* in the benchmark (Figure 3A). In Figure 6A, it is positioned close to the *Cercocebus torquatus (blue)*, while in Figure 7A, it is outside the *Lophocebus albigena (green)*, far from other taxa. In both Figures 6C and 7C, it exactly falls in the space between the *Cercocebus torquatus (blue)* and *Lophocebus albigena (green)*. Based on its position in Figures 6A, 6C, and 7C, a gene flow interpretation may be proposed. The direction of the gene flow may be interpreted to be either from the pink to the *Cercocebus torquatus (blue)* and *Lophocebus albigena (green)* or contrariwise; therefore, it could either be an ancestor or a hybrid. Based on its positioning, *Lophocebus albigena (green)* could have, arguably, migrated to the area occupied by *Cercocebus torquatus (blue)* (Figure 6A) and exchanged genes with the inhabitants.

Other inconsistent outcomes based on the PC plots were also evident; for example, the *Cercocebus torquatus (blue)* and *Mandrillus leucophaeus (red)* proximity varied in PC2 (Figures 3C, 6C, and 7C). The interpretation of these relationships depends on the specific dataset. In the benchmark (Figure 3C), the two groups largely overlap in PC2 to the extent that PC2 may be considered *noninformative*; this overlap could also be inferred as a strong similarity in the genetic content implying a short evolutionary distance, i.e., the difference between the two taxa suggests *taxonomical identity*. By contrast, in Figure 6C, these taxa are largely distinct, which can be attributed to multiple possible factors, including the potential occurrence of divergence subsequent to the accumulation of sufficient genetic differences. This may also suggest the presence of breeding boundaries. Finally, Figure 7C implies that the taxa are completely separated. Here, PC2 appears to *separate all taxa*, and *speciation* could be concluded. Interestingly, this figure has the lowest EV (12.5%).

Investigating the classifiers proved them considerably robust against this type of missingness. For the first test, all the classifiers apart from XGBoost reached accuracies equal to 96% and higher. For the second test, which is more challenging, three classifiers (NN, GP, and SVC) have accuracies greater than 94% and the remaining range between 85% to 91% in the mentioned CV scheme (Table 2). In comparison to other classifiers, the LDA of the first two PCs exhibited poorer performance, whereas the LDA of the first 10 PCs achieved perfect accuracy (100%) in both tests. It can be noted that the 2NN classifier had difficulties in classifying two *Cercocebus torquatus (blue)* samples since, due to the removal of samples from their own taxon, their nearest neighbours are now *Lophocebus albigena (green)* (Figure 7F).

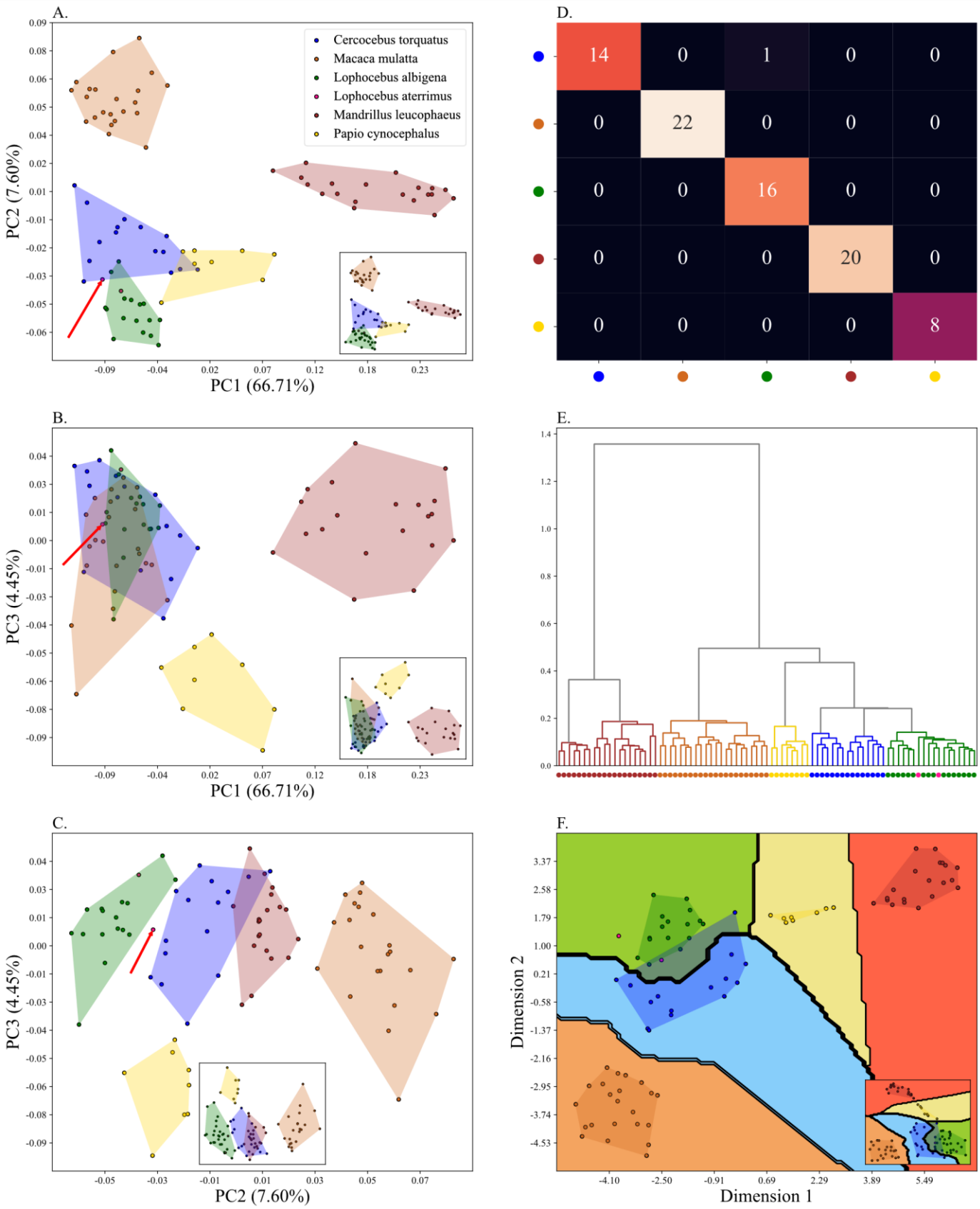
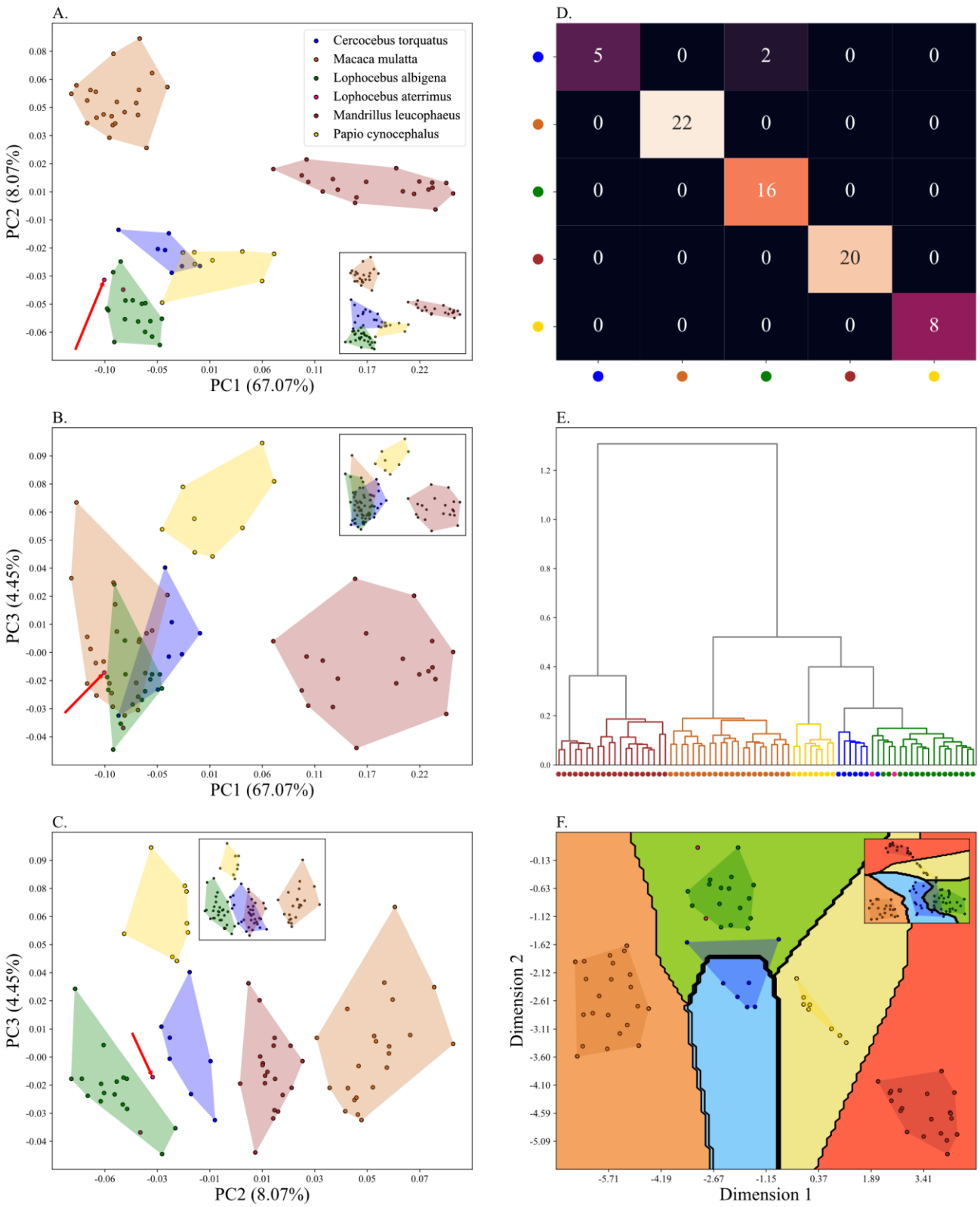


Figure 6: Plots depicting the papionin data after the first case of sample removal. 11 *Lophocebus albigena (green)* samples (index in benchmark dataset: 52, 56, 60, 63, 64, 66, 67, 73, 75, 76 and 77). A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot, E) Decision

boundary of 2NN classifier after t-SNE embedding, F) Decision boundary with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).



The *Lophocebus albigena (pink)* sample is marked by a red arrow.

Figure 7: Plots depicting the papionin data after the second case of sample removal. 19 samples from both *Lophocebus albigena (green)* (index in benchmark dataset: 52, 56, 60, 63, 64, 66, 67, 73, 75, 76 and 77) and *Cercocebus torquatus (blue)* (index in benchmark dataset: 80, 81, 82, 83, 84, 88, 91 and 94) A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot, E) Decision boundary of 2NN classifier after t-SNE embedding, F) Decision boundary with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).

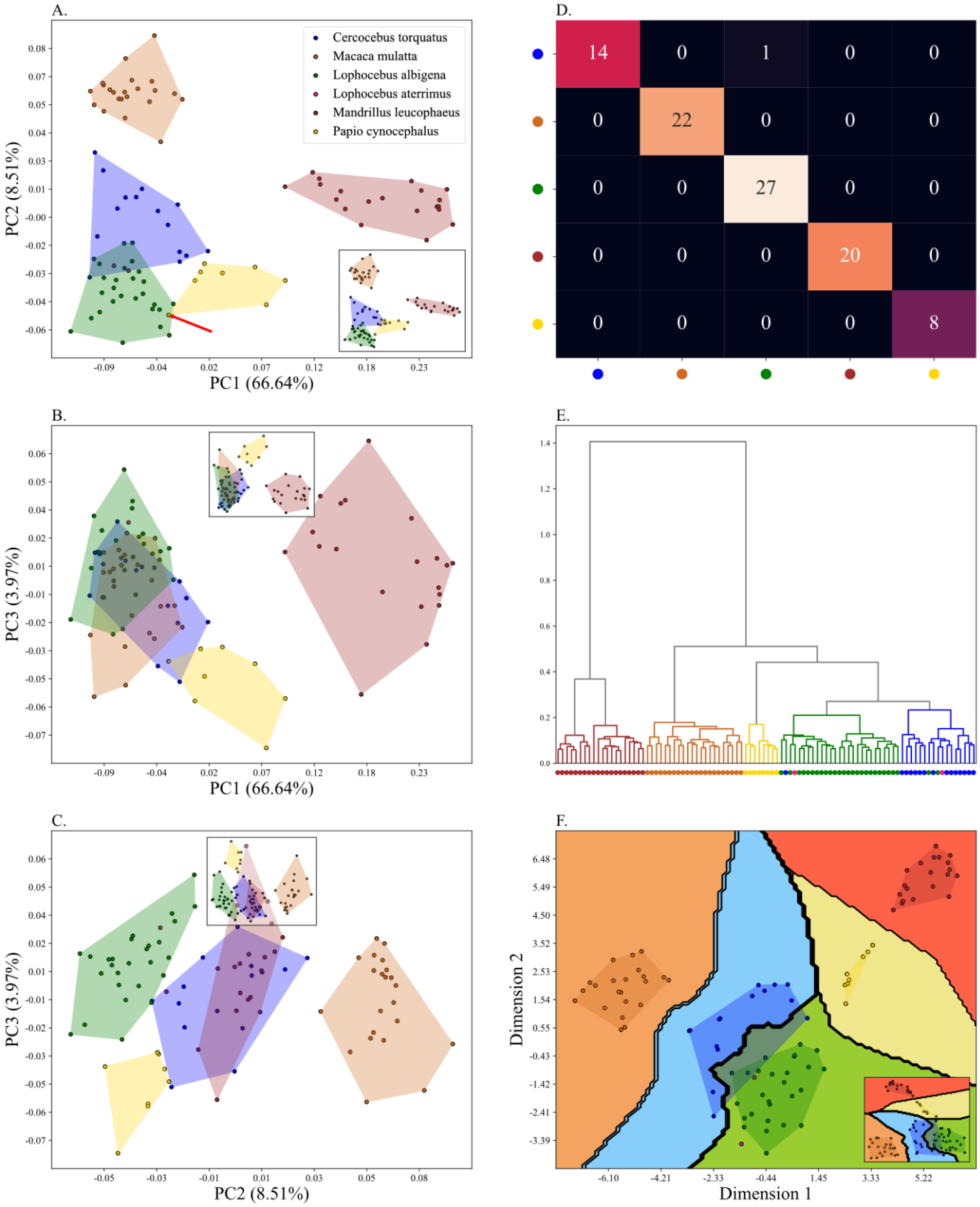
The case of missing landmark

Newfound samples often consist of incomplete bones or fossils (53, 114) and only present some form information. Choosing landmarks is an important and controversial subject (9, 22, 36). However, the ability of a landmark set to predict the shape of the incomplete part is questionable. To evaluate this effect in downstream analyses, we studied two cases of landmark removal, each time removing one set (comprising all three cartesian coordinates of each landmark). First, we removed one set (Figure 8) with the symmetric pairs of the omitted landmarks remaining (Table S3). Next, we removed another set, of which more than half had no symmetric pairs (Figure 9).

This type of alteration results in deviations from the benchmark in the shape of the convex hulls and the distances between the samples and taxa clusters across all the PCs (Figures 8A to 8C and 9A to 9C). For brevity, we only point out the most distinct outcomes. First, we investigate the perceived evolutionary relationships between *Papio cynocephalus (yellow)*, *Lophocebus albigena (green)*, and *Cercocebus torquatus (blue)* (Figures 3, 8, and 9). In the benchmark (Figure 3A) *Cercocebus torquatus (blue)* slightly overlaps with *Lophocebus albigena (green).* This overlap increases considerably in both Figures 8A and 9A. Thereby, while the benchmark suggests that the accumulation of genetic variation has led to speciation, Figures 8A and 9A suggest that the difference can be due to genetic variation within the specie or reflect the penultimate steps of speciation. Moreover, by contrast to the benchmark where the *Papio cynocephalus (yellow)* and *Lophocebus albigena (green)* are completely separated, in Figure 8A, one single *Papio cynocephalus (yellow)* sample (index in the benchmark: 6) now clusters with *Lophocebus albigena (green)*, raising challenges in identifying its origins and relatedness to other taxa. Hence, a range of gene flow hypothesis and other possible causes of shape difference between this sample and both the *Papio cynocephalus (yellow)* and *Lophocebus albigena (green)* can be proposed. The position of the same *Papio cynocephalus (yellow)* sample in Figure 9A is also notable. After removing the second set of landmarks, it appears as an outlier. Thereby, while the complete skull would cluster with its own taxon, with only a small portion of its zygomatico-maxillary region missing, it could either be misclassified as *Lophocebus albigena (green)* (Figure 8A) or considered an outlier to it, possibly belonging to a new taxon (Figure 8A).

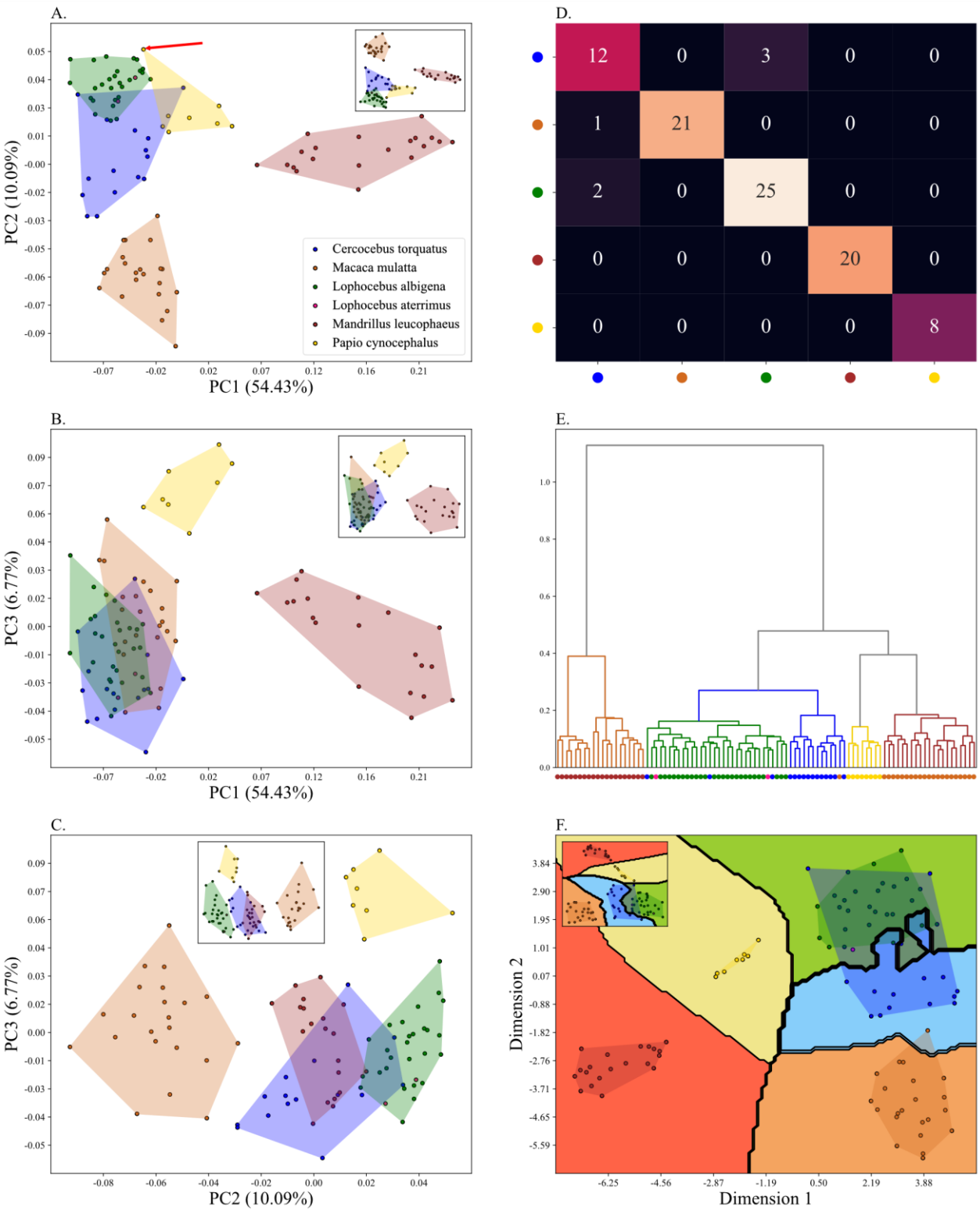
Another notable inconsistency is the relationship between *Cercocebus torquatus (blue)*, *Lophocebus albigena (green)*, and *Mandrillus leucophaeus (red)* over PC2, best seen in PC2-PC3 scatterplots (Figures 3C, 8C, and 9C). In the benchmark and the first case of landmark removal, they are *well separated* by that PC (Figures 3C and 8C, respectively); however, in Figures 9C *Lophocebus albigena (green)* and *Mandrillus leucophaeus (red)* are tangent over PC2. Accordingly, Figures 3C and 8C suggest considerably more evolutionary distance compared to Figures 9C. Different interpretations can be proposed for these plots. Additionally, in Figures 3C and 8C, *Cercocebus torquatus (blue)* appears only to exchange genetic material with *Mandrillus leucophaeus (red)*. However, in Figure 9C, the exchange was between *Lophocebus albigena (green)* and *Mandrillus leucophaeus (red)*. This, in turn, can be interpreted as a loss of biodiversity due to hybridization (14, 85, 94, 112). Similar scenarios can be observed when feral and domesticated species mate (95). This may lead to different population dynamics depending on factors such as the frequency of such events, the size of the populations involved, and the selection pressures. Consequently, this could even lead to homogenization (103).

Considering the supervised classifiers, this type of alteration resulted in a negligible drop in accuracy (Tables 1 and 2). However, their performances were acceptable in both cases. Three (NN, GP, SVC) and two (NN, GP) had accuracies above 97% and equal to 95% in the first and second cases, respectively, higher than the LDA of the first two PCs. The LDA of the first 10 PCs performs very well in both tests (accuracy = 100% and 96% in the first and second tests, respectively). Due to distortions in the second case, some of the *Cercocebus torquatus (blue)* and *Lophocebus albigena (green)* samples have moved close to the other taxon and get misclassified by the 2NN classifier (Figures 9D-F). In the first case, the accuracies of RF and ET were above 90%, while in the second case, they dropped below 90%. With an accuracy below 85%, XGBoost had the lowest performance. The results indicate the robustness of the classifiers to partial missingness in landmarks.


A.
Cercocebus torquatus
Macaca mulatta
Lophocebus albigena
Lophocebus aterrimus
Mandrillus leucophaeus
Papio cynocephalus
PC2 (8.51%)
PC1 (66.64%)
B.
PC3 (3.97%)
PC1 (66.64%)
C.
PC3 (3.97%)
PC2 (8.51%)
D.
14 0 1 0 0
0 22 0 0 0
0 0 27 0 0
0 0 0 20 0
0 0 0 0 8
E.
F.
Dimension 2
Dimension 1

508

Figure 8: Plots depicting the papionin data after the first case of landmark removal. Landmarks 23, 24, 25, 26 and 27 were removed. Scatterplots of the papionin data after the



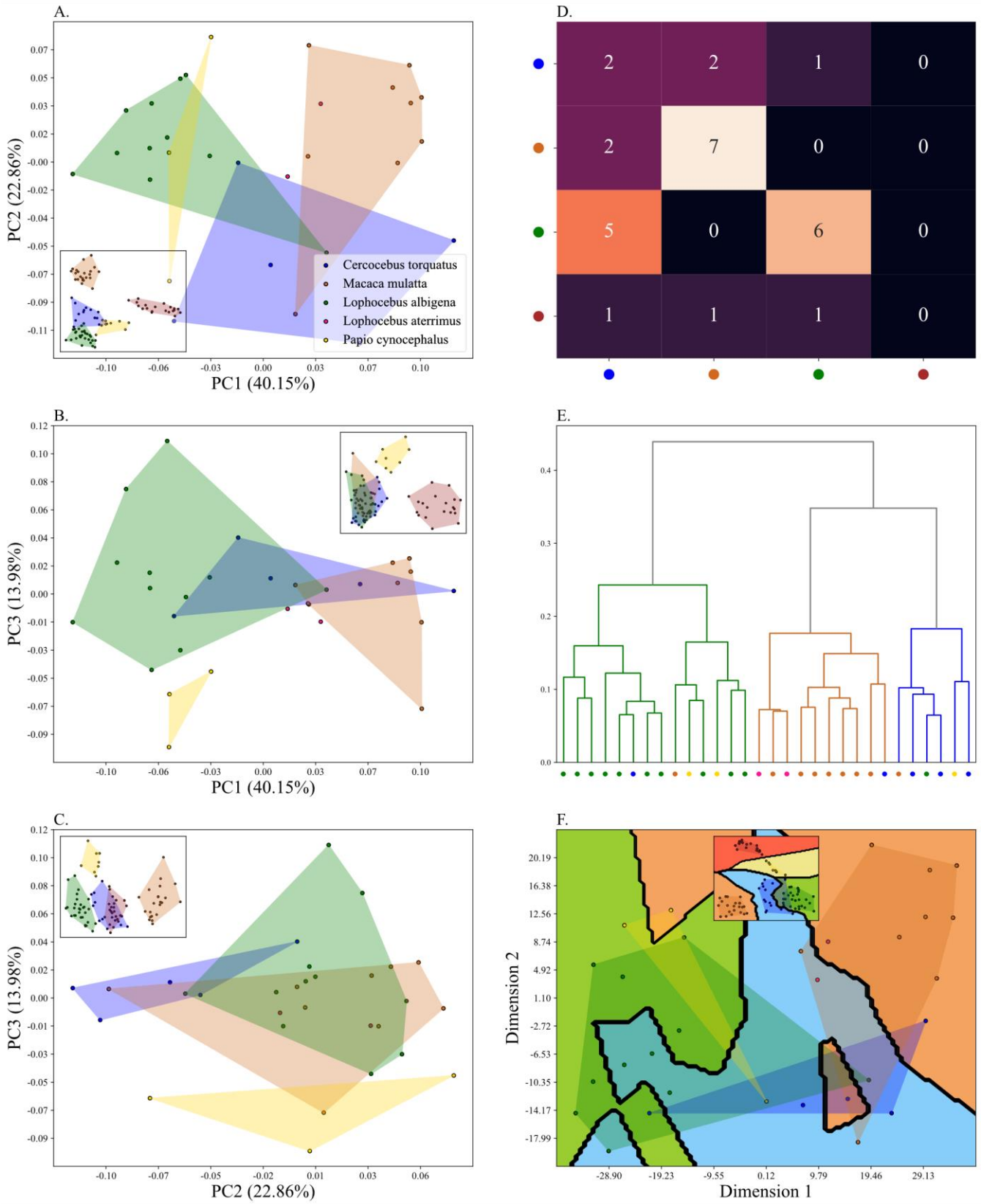
first case of sample removal. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot, E) Decision boundary of 2NN classifier after t-SNE embedding, F) Decision boundary with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).

Figure 9: Plots depicting the papionin data after the second case of landmark removal. Landmarks 11, 12, 16, 17, 18, 30 (symmetrical pair of 12) and 31 (pair of 13) were removed. Scatterplots of the papionin data after the first case of sample removal. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot E) Decision boundary of 2NN classifier after t-SNE embedding, F) Decision boundary with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).

The case of combined alterations

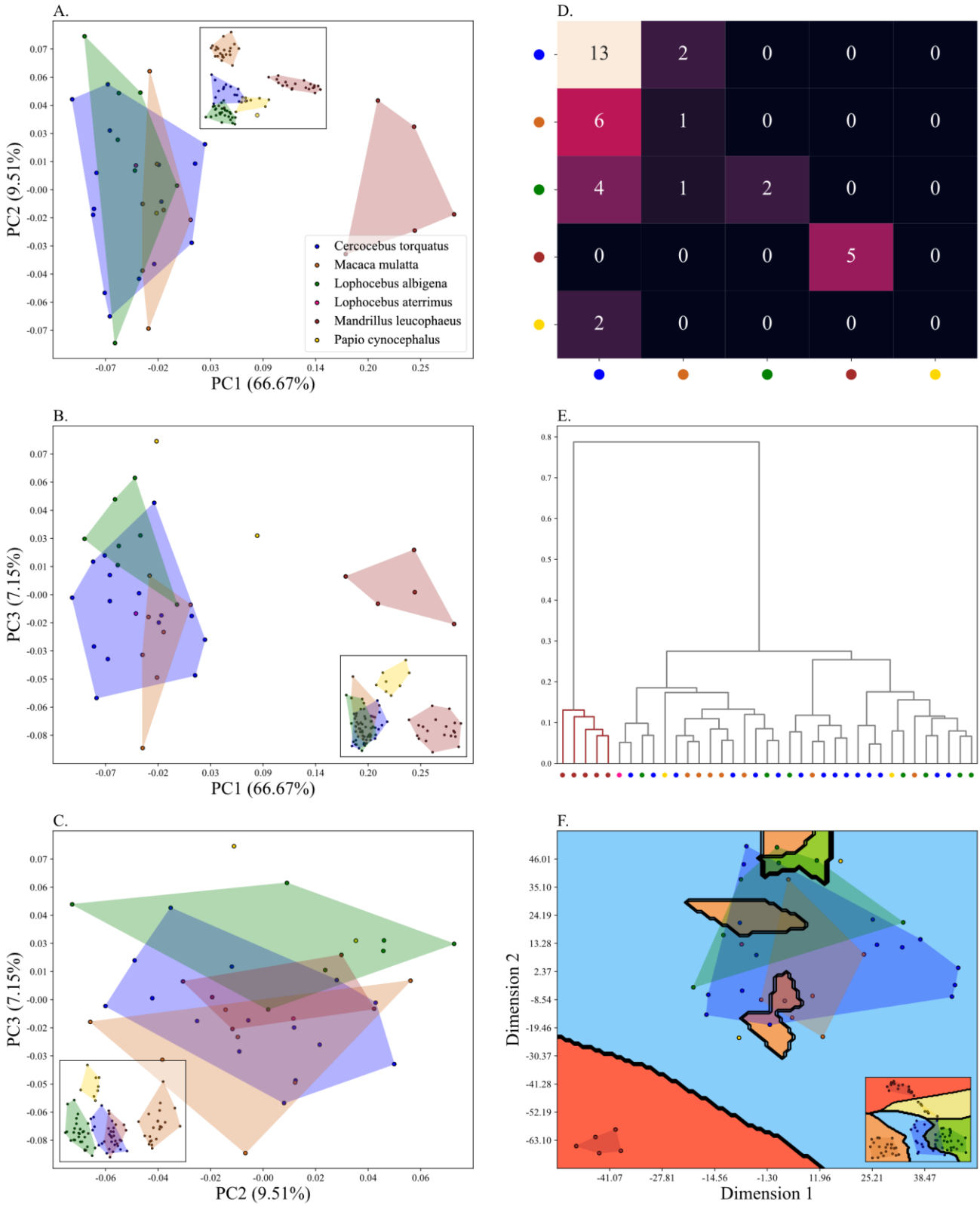
Thus far, for simplicity, we have introduced and studied the three types of alterations separately as if they represent separate challenges associated with ancient fossils; however, in reality (7, 53, 114), when a fossil is found, it might only be a fragment (fragments) of a bone (landmark removal) incomparable with some fossils (sample removal). It is also common to compare samples of interest with much larger cohorts of other taxa. For example, Harvati et al. (48) compared 12 *Neanderthal* specimens from Europe and the Near East; the early *Neanderthal* specimen from Reilingen, the Middle Pleistocene African specimen Kabwe, two early anatomically modern humans (AMH) from the Near East, and four Late Palaeolithic European AMHs and 270 modern humans from nine populations using fifteen temporal landmarks. In the case of *Homo NR*, Hershkovitz et al (53) Analysed four landmarks and 77 surface semi-landmarks in NR-1, 18 landmarks and 77 curved semi-landmarks in NR-2, and a set of six landmarks, 18 semi-landmarks (enamel–dentine junction (EDJ)) and two landmarks and 18 semi-landmarks (cemento-enamel junction (CEJ)) in the lower left second molar in a combined analysis of the junctions. In each analysis, the sample was compared with a different set of humans. In both studies (48, 53) the number of samples representing each group was utterly dissimilar; some groups were not presented, and some were under/over-presented. In the case of *homo NR (53)*, the use of only four landmarks (specifically on the cranium) was another limitation (53). Unfortunately, despite numerous appeals made to the authors and journals, neither group released their complete data publicly, thus hindering reproducibility.

To better understand how these combined limitations affect morphometric analyses, we investigated two cases. In the first case, one taxon (*Mandrillus leucophaeus, red*) and a greater portion of the collective samples were randomly removed (30/94 samples remained), and only a fracture of the cranium (landmarks #1, 2, 3, 4, 6, and 14) was investigated. In the second case, all the taxa were present, but only a few samples (37/94) and a small portion of the cranium (landmarks #1, 2, 3, 4, and 14) were kept. Figures 10 and 11 show the results of the PCA and the 2NN classifier for the two cases, respectively. At odds with the benchmark (Figure 3A), in the PC scatterplots, many groups overlap. Moreover, the samples are poorly separated. Unsurprisingly, the performance of the ML classifiers is considerably low for these tests (Tables 1 and 2), with balanced accuracies slightly above 50% and 40% for the first and the second case, respectively. In such cases, the interpretability of the data is questionable as a fundamental lack of information in the dataset will bias all methods. In other words, the



typical landmark datasets in palaeoanthropology analysed using the established pipeline (Figure 2) yield incorrect results that can only be partially improved with alternative methods.

Figure 10: Plots depicting the papionin data in the first extreme case. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot E) Decision boundary of 2NN classifier after t-SNE



embedding, F) Decision boundary with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C). the corresponding benchmark PC scatterplots (Figure 3A-C).

Figure 11: Plots depicting the papionin data in the second extreme case. A) PC 1-2, B) PC 1-3, C) PC 2-3, D) 2-D t-SNE scatterplot E) Decision boundary of 2NN classifier after t-SNE

embedding, F) Decision boundary with 5-fold CV. The insets in the subfigures are the corresponding benchmark PC scatterplots (Figure 3A-C).

The Case of Outlier Detection

Newfound samples may represent individuals (7, 53, 114) of unknown taxa (7, 114). Therefore, identifying them requires a different approach than classification because classifiers can only classify samples to already known classes and, thus, are not capable of detecting outliers (taxa). To address this issue, we investigated the ability of three outlier detection methods by measuring the outlier score for outliers using the Local Outlier Factor (LOF) (18), isolation forest (70), and once class SVM (101); since the isolation forest and one class SVM methods yielded significantly poor results compared to LOF (Figures S8 and S9) we did not investigate them further.

When exploring the origins of unidentified samples through morphometrics, it is crucial to exercise extreme caution. As early as 1978, Oxnard (87) emphasized the necessity for meticulous attention, particularly when examining a single specimen, more so in the case of fossils, where researchers may unconsciously seek to validate their intuitive speculations using analytical methods. To address this and investigate the performance of outlier detection methods, we carried out three trials by considering *Cercocebus torquatus (blue)* samples as novelties (outliers) and calculated their outlier score in all tests. First, to address the case of finding a single (53) specimen, we added each *Cercocebus torquatus (blue)* sample to the cohort separately. Second, for the case of several (7) individuals belonging to the same taxon, we performed two trials; we added six *Cercocebus torquatus (blue)* in each other's proximity together. In the last trial, we added 15 *Cercocebus torquatus (blue)* samples to the dataset. Figures 12-14 depict the results. Eq. 1 yields the radius of the circle around the *i_th* sample in the scatterplots. The LOF scores for each sample in all three tests are presented in Table 3. Samples receive considerably high scores when treated separately, which identifies them as outliers. As expected, adding samples together reduces their outlier scores (greater magnitude, smaller negative score) because LOF evaluates the samples based on the concept of local density, and samples from one group are expected to be in a neighbourhood in the shape space manifold (considering that the data is informative). If the LOF scores reduce after being treated together, then it could mean that they belong to an as-of-yet-unknown group. These results confirm the ability of LOF to identify novelties in landmark data shape space.

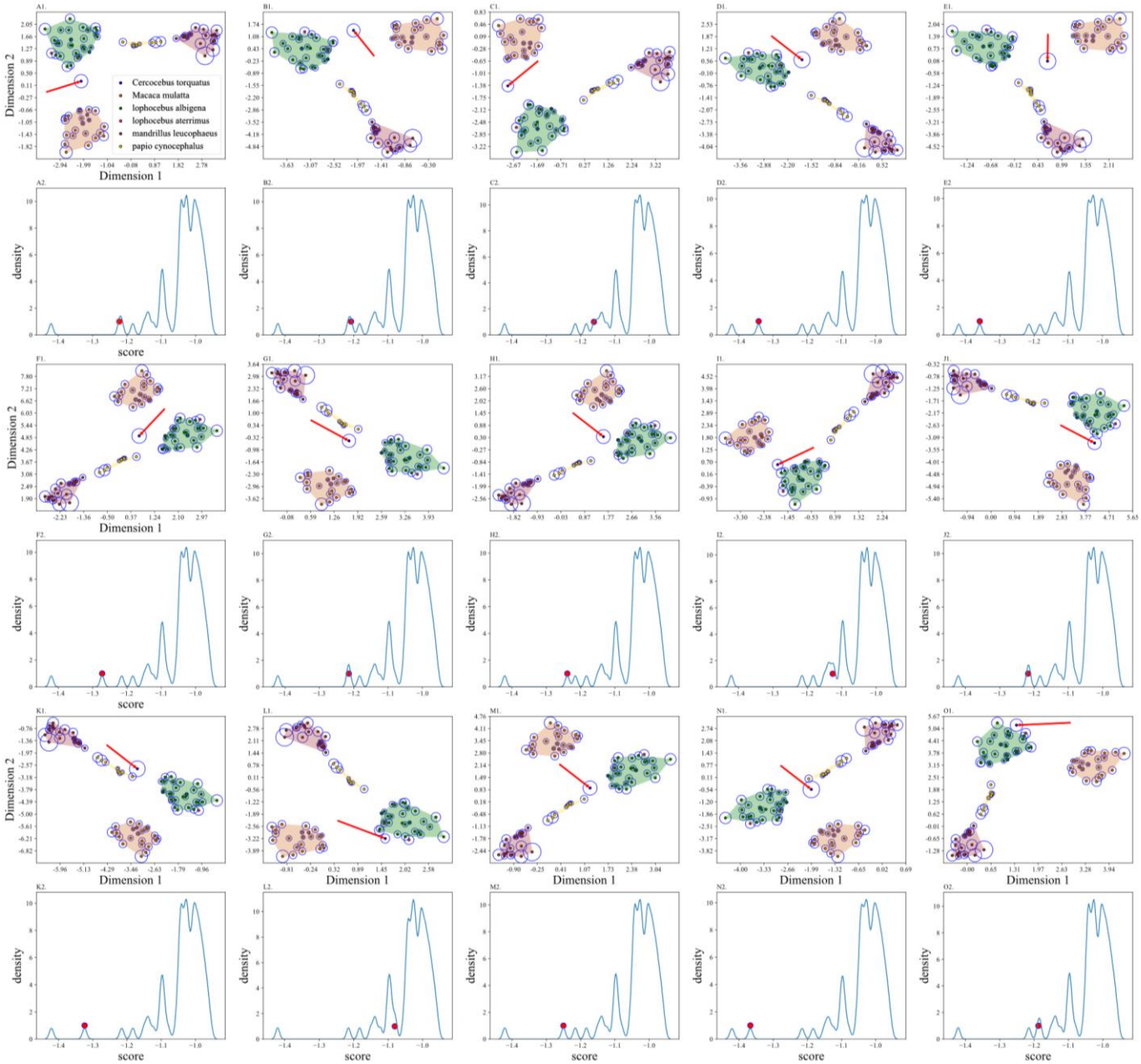


Figure 12: Detecting outliers using local outlier score and KDE plots. Each *Cercocebus torquatus (blue)* sample is treated as an outlier (red arrow, A1-O1 t-SNE plots) and the KDE of the scores are plotted for each case (A2-O2). The radius of the circle around each sample is proportional to the LOF score. The red dot in the KDE plots shows the LOF score of the outlier in each case.

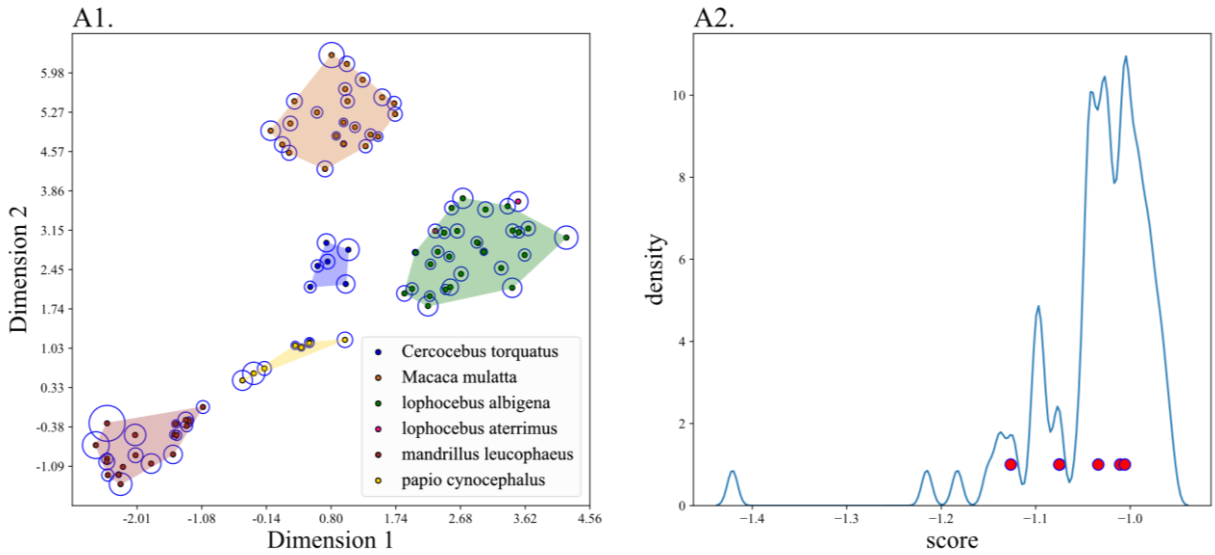


Figure 13: Detecting outliers using local outlier score and KDE plots. Six *Cercocebus torquatus (blue)* samples (indexes: 84,85,86,87,90 and 92) are treated as outliers (A1 t-SNE plots) and the KDE of the scores are plotted (A2). The radius of the circle around each sample is proportional to the LOF score. The red dots in the KDE plots show the LOF score of the outliers.

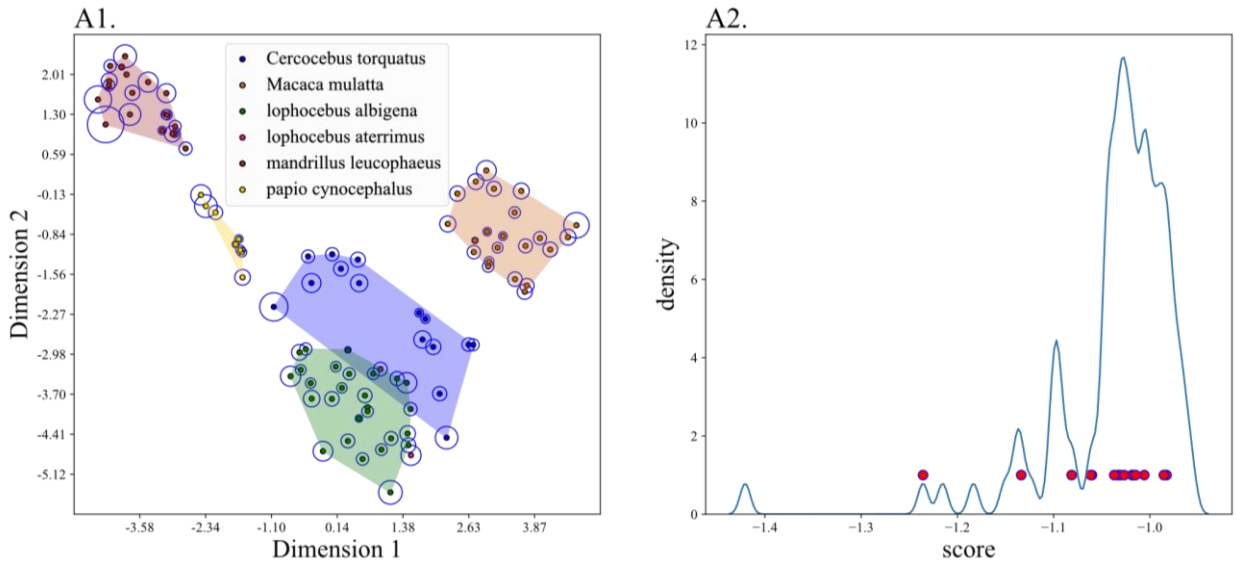


Figure 14: Detecting outliers using local outlier score and KDE plots. All *Cercocebus torquatus (blue)* samples are treated as outliers (A1 t-SNE plot) and the KDE of the scores are plotted (A2). The radius of the circle around each sample is proportional to the LOF score. The red dots in the KDE plots show the LOF score of the outliers.

Table 3: The LOF outlier scores. *Cercocebus torquatus (blue)* was treated as an outlier in three cases; A) each sample separately, B) 6 samples together, and C) the whole group together.

Sample	**A) LOF score (a)**	**B) LOF score (b)**	**C) LOF score (c)**
80	-1.22	-	-0.98
81	-1.21	-	-1.02
82	-1.16	-	-1.01
83	-1.34	-	-0.99
84	-1.36	-1.07	-1.03
85	-1.27	-1.03	-1.03
86	-1.21	-1.01	-1.02
87	-1.24	-1.12	-1.06
88	-1.13	-	-1.04
89	-1.22	-	-1.06
90	-1.32	-1.00	-1.01
91	-1.08	-	-1.03
92	-1.25	-1.07	-1.08
93	-1.37	-	-1.24
94	-1.19	-	-1.13

Evaluation of the informativeness of landmarks with feature selection

Using informative input is vital to the success of machine learning-based classification methods. It is well established that the required number of data points for an accurate outcome is (generally) exponentially proportional to the number of data dimensions (34), a phenomenon that Bellman termed the "*curse of dimensionality*" (12). In other words, an analysis feasible in low-dimensional spaces can be challenging or meaningless in higher dimensions as data becomes too sparse (47). To overcome this problem, it is necessary to reduce the number of dimensions through either feature selection methods that select informative features (54) or feature extraction/transformation methods that create new features from the original variables (109). To assess the informative or noninformative features, we carried out several iterations of feature selection, although the low number of landmarks in our benchmark dataset did not handicap the classifiers (Tables 1 and 2). For that, we applied a genetic algorithm (21, 51, 68, 72) as a wrapper method with a logistic regression classifier five times and calculated the accuracy of the classifier with the identified subsets in each case (Tables S2 and S3). Despite considerable differences in the identified landmark coordinate subsets, all subsets performed with similar accuracy, nearly equal to the original set, although half its size. In other words, half of the landmark coordinates are either redundant, noninformative, or both.

Discussion

The quantitative analysis of morphological data has always captivated researchers (29, 107). Employing graphical and computational methods, as done today, is rooted in ancient geometric principles developed since the European Renaissance (36, 73). However, it wasn't until the mid-1990s (15, 35, 45, 102) that the mathematical underpinnings of shape found solid ground, complemented by cutting-edge graphics and statistical tools and propelled modern morphometrics into the realm of essential tools spanning diverse fields, including evolutionary biology, anthropology, palaeontology/palaeoanthropology, and even forensics (8, 28, 67, 88, 105). In biological science, morphometrics is used to study temporal shape variation inter- and intra-species and is instrumental in exploring questions – from the intricate changes in shapes during evolution to questions concerning growth and development. There are various approaches in morphometrics, but among them, geometric morphometrics has left its indelible mark on biology, especially in physical anatomy and anthropology.

Geometric morphometrics encompasses a diverse range of approaches; most share a standard workflow involving two core steps: Generalised Procrustes Analysis (GPA) followed by Principal Component Analysis (PCA) (Figure 1) (8, 36, 41, 67, 77). GPA aligns landmark coordinates by minimising the shape-independent variations across samples. PCA projects the data onto uncorrelated variables called principal components (PCs), which capture different proportions of variation within the data (115). The PCs are then plotted on colourful scatterplots and examined visually. It is generally expected that biologically related shapes will exhibit closer clustering while unrelated shapes will be more distanced and dispersed. However, there are no rules on which plots to use. Researchers typically select favourable PCs, disregarding the remaining ones (30, 48, 53, 111), and claim that various PCs, i.e., statistical constructs agnostic to the underlying data (86), represent demographic and evolutionary processes (48, 84, 114). Interpretations may also encompass origins, relatedness, gene flow, speciation, phenetic evolutionary trends, and phenotypic/genotypic variation within samples and taxa (8, 30, 41, 42, 48, 53, 59, 67, 84, 111, 114). Additional assumptions include that variations or covariations along specific PCs are often linked to specific morphological traits (16, 53, 115) and that PCs with *sufficiently higher explained variances (EVs)* represent certain phenotypic (36, 48, 53, 84) or even sexual (48, 84) morphological traits.

Phylogenetic Principal Component Analysis (Phy-PCA) (92) and Phylogenetically Aligned Component Analysis (PACA) (27) are two methods used in geometric morphometrics to analyse shape variation while considering the supposed phylogenetic relationships among species. They differ in their approach to aligning landmark configurations and the role of PCA within them. Phy-PCA incorporates phylogenetic information by utilising a phylogenetic tree to model the evolutionary history of the species. This method aims to separate shape variation resulting from shared evolutionary history from other sources of variation. PCA plays a similar role in performing dimensionality reduction on the aligned landmark configurations in Phy-PCA (92). PACA takes a different approach to

alignment. It uses a Procrustes superimposition method based on a phylogenetic distance matrix, aligning the landmark configurations according to the evolutionary relationships among species. PCA is then applied to the aligned configurations to extract the principal components of shape variation (27). Both Phy-PCA and PACA are implemented to gain insights into the patterns and processes shaping biological form diversity while considering phylogenetic relationships. It is important to note that because these methods rely on PCA as part of the process, they are also subjected to PCA's limitations and biases.

Given the widespread popularity of PCA in geometric morphometrics (18,000-32,900 studies), researchers and practitioners naturally anticipate that it will yield accurate and valid results. The growing body of literature criticising the accuracy of various PCA applications (5, 11, 23, 31, 37, 38) raises concerns about its accuracy in geometric morphometrics. Surprisingly, despite its far and wide applications, PCA has never been shown to produce *correct* shape inferences. For that, we undertook the first thorough investigation of the possible shortcomings of PCA in geometric morphometrics, where we assessed its reliability, robustness, and reproducibility in shape analysis compared to eight alternative machine learning methods. Specifically, we evaluated PCA performance for four use cases using a benchmark dataset, to which we introduced eight alterations simulating real-life scenarios. In the fifth use case, we investigated the accuracy of machine learning methods to perform outlier detection in the case of novelties, i.e., the identification of new taxa.

A key strength of our study is using benchmark data as the "ground truth" to assess the accuracy of PCA using t-SNE, dendrograms and confusion matrices as alternative visualisation aids. We found that the five papionin taxa cluster separately in the dendrogram of the shape space and its t-SNE plot (Figures 3E and 3F). This allowed us to evaluate the performances of PCA in light of alterations, representing real-life scenarios, which reflect the restraints faced in morphometric analysis of ancient samples (e.g., *hominins*). PCA scatterplots of the altered datasets changed the taxa's positioning, the clusters' shape, and the distances within and between clusters with respect to the benchmark. For example, samples from one cluster would overlap with others; clusters tended to appear more or less heterogeneous, and separate clusters would overlap. We showed that the PC scatterplots produce inconsistent results for the same dataset based on the choice of the PC. We demonstrated how analysing different PCs gives different outcomes, which can lead to different conclusions (Figure 2). Using the *Homo Nesher Ramla* (*NR2*) as a test case (53), we demonstrated that the experimenter could influence the outcome by selecting preferable results out of all the conflicting options provided by PCA (Figure 2). Alternatively, researchers may see patterns emerging when analysing different PCs and uphold the most favourable ones. For example, Harvati (48), who analysed the *Skhul 5* (43), a 40,000-year-old human skull from *Mount Carmel* (Israel), marched between conflicting hypotheses based on cherry-picked PC outcome (based on PC8 separating it from *Neanderthals* and *modern humans* and associating it with the *Late Palaeolithic* specimen and based on PC12 associating it with *modern humans*). Such results would lead to contradictory evolutionary misinterpretations, promoting misidentification and misclassification. Our results (Figures 4-11) show that, compared to the benchmark (Figure 3), PCA produces results that are statistical artefacts of the input data, and although they can be interpreted in evolutionary

terms such as gene flow, admixture, or speciation – they would nonetheless be artefacts. Overall, we found that PCA results are not reliable, robust, or reproducible as many members of the field assume (Table 1).

The popularity of PCA stands in sharp contrast with its multiple limitations, some of which were also reported in other fields (e.g., 24, 38). To name a few, PCA lacks any measurable power, a test of significance, or a null model, making it a problematic tool, at the very least. PCA outcomes are open to manipulations and interpretations, and the ability to evaluate various dimensions allows for cherry-picking of the results where all the answers are equally acceptable. Without external validation methods, such as used by Marom and Rak (74) in their criticism of Hershkovitz et al.'s (53) work based on unique morphological characteristics of *Homo neanderthalensis* mandible (e.g., the medial pterygoid tubercle), PCA-based arguments are vacuously true based on the antecedent of their correctness. It may be argued that PC results with "*sufficiently high EVs"* should be favoured. Not only does no threshold for such a term exists, but we showed instances that higher EV plots were less accurate than lower EV plots (Figure S1). Overall, we found no justification for the use of PCA, by itself, in morphometrics.

Beyond evaluating PCA's accuracy, we comparatively investigated the accuracy of eight machine learning methods using the same use cases and benchmark in classifying superimposed landmarks, including LDA with the first two and 10 PCs. The ML classifiers showed superior performances across all the tests except the most extreme case of the combined alterations, where all the analyses produced inaccurate results (~50% accuracy). Not only were these methods more accurate in the first three use cases (80-100% accuracy), but they were also less susceptible to experimenter intervention. It can be observed that the classifiers are also less sensitive to and more stable in a decrease in the number of samples. Moreover, the low-dimensional embedding could form distinct clusters in the first three cases (Figures 4 to 9). Linear Discriminant Analysis (LDA) of the first two PCs exhibited relatively lower performance compared to the seven ML classifiers. Conversely, the LDA utilising the first 10 PCs demonstrated satisfactory performance. It is worth mentioning that while LDA is extensively utilised as the primary technique for classifying landmark data in the field, there is no established guideline regarding the optimal number of PCs to be incorporated in the Discriminant Analysis (DA). Moreover, a drawback of LDA applications is handling a large number of variable ((80)). In practice, this selection is often made by the experimenter to obtain the preferable performances in a cross-validation (CV) (10) scheme and may include the first 30 (81) or 50 (90) PCs.

Of all the classifiers tested here, the simpler classifier (NN) was more efficient than the more advanced machine learning methods. Unsurprisingly, the tree-based classifiers had relatively low accuracies since they have a higher variance error component as they are more sensitive to data distribution. Tree-based algorithms partition the feature space based on the available data. When the sample size is small, the algorithm becomes more sensitive to the specific distribution of the training data; thereby, small variations or outliers in the data can substantially impact the resulting tree structure, leading to higher variance in predictions (58).

The advantage of the NN classifiers as a substitute for PC scatterplots and to promote geometric morphometrics cannot be overstated.

The identification of novel taxa is always exciting – as long as it can be proven as such. Unsurprisingly, PCA has played a pivotal role in facilitating such "breakthroughs" (21, 54, 103, 117). To note one example, it has been assumed that craniofacial morphology can effectively serve as a population genetic proxy for investigating population structure, based on its broader spatial and temporal coverage (100, 101, 116). It has further been claimed that this evidence is stronger than paleogenetic evidence based on ancient DNA derived from certain populations such as inhabitants of the pre-Contact New World. Therefore it was employed for multiple dispersal events and ancestral origins inferences (100, 101, 116). However, it has been proven conclusively that this does not hold true generally in disputes surrounding the case of the "Kennewick Man" (96). Moreover, it is important to note that paleogenetics is based on several thousand of clear biological metrics with internal consistency which can be effectively compared (97) as opposed to limitations faced in morphometric comparisons of ancient cranial remains.

Unfortunately, the problems with physical anthropology are not limited to the misuse of statistical methods. It is common to encounter private and non-sharable datasets, which are maintained by an alliance between five key actors – authors, journals, funders, museums, and the media – to ensure the survivability of physical anthropology and, in turn – theirs. For example, Hershkovitz and May admitted that the co-authors of Hershkovitz et al. paper (15) had signed a contract to maintain the data private (personal communication), which nullifies the data sharing policy of Science (https://www.science.org/content/page/science-journals-editorial-policies). After falsely claiming that the data-sharing policy was not in place when the said paper was published, the Science editor upheld this violation of their own data-sharing policy and, once pointed out their problematic behaviour, ignored our further emails (personal communication). Excepting one funder, who claimed they did not fund that study, although listed in the paper, funders have either ignored our correspondence or saw no problem in this practice. The role of museums and the agreements they require scientists to sign when seeking access to morphometric information of samples should also be considered. On one end of the scale are museums willing to share data subjected to stipulated agreements, whereas on the other end are museums that adopt a more restrictive stance, particularly when dealing with valuable artefacts, and closely monitor who gains access to the data and how the information is utilised. Museums lie somewhere between these two ends. Finally, the media that publish uncritical hyped reports encourages researchers to use poor methods to satisfy this hype. For example, with respect to the *Homo Nesher Ramla*, The Times's headline was: "Meet Nesher Ramla Homo: New form of human found,"(1) The Guardian's headline was: "Fossilised bones found in Israel could belong to mystery extinct humans,"(4) and Science Alert's headline was "A Previously Unknown Type of Ancient Human Has Been Discovered in The Levant."(3) Remarkably, only a few sources, including Haarretz (2) and The Times (91) bothered to update their readers about the juicy controversy following the finding of Marom and Rak that the Homo Nesher Ramla is a Neanderthal (74). In the more recent case of HLD 6 (114), CNN's headline was '300,000-year-old skull found in China unlike any

early human seen before' (63), and The Daily Mail's headline was 'Have scientists discovered a new species of HUMAN? Ancient skull belonging to a child with no chin who lived 300,000 years ago suggests our family tree 'needs another branch'' (108). The hype was later echoed in a Nature News article titled "A new human species? Mystery surrounds 300,000-year-old fossil" (69). All in all, our findings, coupled with the poor scientific practices in the field, raise major concerns regarding a large corpus of morphometrics analyses and offers a way to revolutionise it.

In summary, the limitations of PCA render it an unserviceable tool for geometric morphometrics analyses. PCA is also unsuitable for detecting novel taxa. Our comparative analysis showed that three ML methods (namely, nearest neighbours classifier, gaussian process classifier, and logistic regression) have superior performances for classification and highlight the effectiveness of the Local Outlier Factor (LOF) tool in identifying outlier individuals and groups in shape space. As a remedy, we created MORPHIX, an open-source Python package encompassing a comprehensive set of tools for processing superimposed landmark data. MORPHIX enables users to apply their preferred classifiers and outlier detection methods and provides a range of visualisation options. MORPHIX is freely and publicly accessible via GitHub at https://github.com/NimaMohseni/Morphometric-Analysis. We also provided detailed examples to demonstrate the usage of the package. We are aware that PCA has additional usages in physical anthropology, for example, studying sex bias or phylogenetics. The applicability of PCA to these questions can be evaluated using the approach described here.

Conclusions

Modern geometric morphometrics is widely used to analyse physical forms, specifically in physical anthropology. Geometric morphometrics' heavy reliance on Principal Component Analysis (PCA) to address questions of classification as well as relationships, variations, and differences between samples and taxa, coupled with recent criticisms of this approach, merits an evaluation of its accuracy, robustness, and reproducibility. Through extensive investigations and alterations to landmark data, we showed that PCA introduces statistical artefacts specific to the data and fails to meet the required merits. We raise concerns about PCA usage in 18,000 to 32,900 studies and call for a revaluation of its usage. We also demonstrate that machine learning methods outperform PCA for classification and outlier detection, particularly when the data are informative. We made these tools freely and publicly available in a Python package.

Methods

The dataset

We used a dataset comprising the three-dimensional coordinates of 31 landmarks measured with an accuracy of 0.5 mm. The landmarks represent the locations of sutural junctions, maxima of curvature, and a number of other important anatomical features (Figure 15 and Table S1) of the crania of 94 adults of five papionin genera (84). Adult crania had been selected based on completed dentition (84).

The taxa were *Cercocebus* (15 specimens of *C. torquatus* including seven males and eight females), *Lophocebus* (27 specimens of *L. albigena* including 16 males and 11 females, plus two specimens of *L. aterrimus*, both female), Macaca (22 specimens *M. mulatta* including 10 males and 12 females), *Mandrillus* (20 specimens of *M. leucophaeus* including 13 males and seven females) and *Papio* (eight specimens of *P. cynocephalus* including three males and five females) (84).

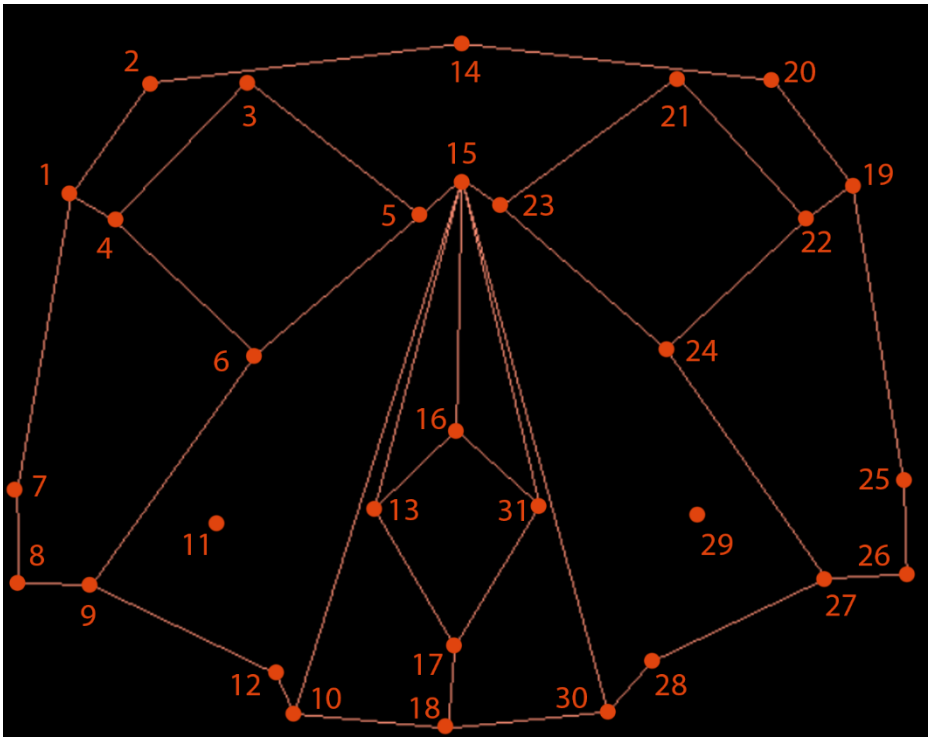


Figure 15: The 31 landmarks of Table S1. The wireframes connect the landmarks.

The three papionins and their skull pictures presented in Figure 1 were obtained from the following links under a Creative Commons Attribution-Share Alike 3.0 Unported license with the exception of the last link, which was used with permission:

https://tinyurl.com/mrh3ukz8, https://tinyurl.com/2p8777j4, https://tinyurl.com/7dxbt4zb, https://tinyurl.com/ar4uujdj, https://tinyurl.com/47n9z7ae, https://tinyurl.com/mxvr439c.

All the pictures were cropped. For two skulls, the background was removed.

The PC scores use for creating Figure 2 were obtained from Israel Hershkovitz.

Processing landmark data with General Procrustes Analysis (GPA) and Principal Component Analysis (PCA)

The process of making biological inferences from landmark is as follows: raw landmark data consist of three-dimensional coordinates of specific points on the surface of the object or organism that are biologically meaningful or correspond to homologous structures and can be reliably and accurately located across samples. Using specialized software and hardware, the coordinates of the landmarks are recorded. The location of each landmark is recorded as a set of X, Y, (and) Z coordinates in a digital file, which can then be used for further analysis. Raw landmark data cannot be directly used as the input of computational and learning algorithms. To make the samples effectively comparable, superimposing the samples of a dataset with General Procrustes Analysis (GPA) is necessary (36, 99). This method operates in non-Euclidean shape space (62) and, through translation, rotation and scaling, superimposes the landmark data by cantering all the samples in their centroid, standardising centroid size, and reducing variations related to rotation (5). Theoretically, these steps can underline possible differences independent of position, rotation, and isometry (36). PCA is often applied after GPA to identify the major patterns of shape variation among individuals or samples. Moreover, it is used to identify the major axes of shape variation and to quantify the amount of variation explained by each PC. The first few PCs usually capture the most variation in the data and are visualised for hypothesis testing, such as comparing the shape variation between groups or observing the association between shape and other variables. The GPA and PCA analyses in this study were carried out using the morphometrics package morphologika (83).

Employing classifiers to investigate the origin of the samples

We adopted seven machine learning classifiers to assess whether a new observation belongs to known taxa. Excepting for NN, this process involves a training phase in which the algorithm will try to learn the underlying patterns in the data to classify new observations. For each case, we evaluated and compared the performance of several machine learning algorithms, including Nearest Neighbours classifier (NN) (47), Logistic Regression (LR), Gaussian Process (GP) classifier (96), Support Vector Machine (SVM) classifier (52), Random Forest (RF) classifier (55, 56), Extra Trees (ET) classifier (44) and Extreme Gradient Boosting classifier (XGBoost) (25). We also adopted Linear Discriminant Analysis (36) for classifying the superimposed landmark data in the reduced space of the first two and

10 PCs. Tables 1 and 2 summarise their performance on the benchmark dataset and the alteration cases in terms of accuracy and balanced accuracy in a five-fold CV scheme, respectively. Supplementary Figures S1-S7 present the corresponding confusion matrices. All the post GPA-PCA analyses were carried out in python using the following modules: Numpy (46), Pandas (75), Scikit-learn (89), DEAP (40) (32) and Scipy (113). We note that while NN is not considered a traditional machine-learning method, it is commonly listed as such in the literature (106) and, for simplicity, will be referred to as such here.

Visualisation of results

To visualise results, we used two-dimensional t-distributed Stochastic Neighbour Embedding (t-SNE) (110) for dimensionality reduction and PCA (full Procrustes projection). t-SNE aims to reduce the Kullback-Leibler divergence between the joint probability of the embedded and the high-dimensional data via translating similarities between data points into joint probabilities (110). We used two more tools in visualisation. First, convex hulls were used to define clusters of samples belonging to one taxon. Second, the decision boundary of classifiers was projected into the low-dimensional embedded space and presented with Voronoi tessellation (76) with a five-fold CV. A 1NN classifier trained on the results of the 2NN classifier was used to classify a two-dimensional mesh grid spanning the space of the plot to create the tessellation and contours used for depicting the decision boundaries.

Kernel Density Estimation (KDE) to evaluate distributions

Kernel Density Estimation (KDE) is a non-parametric statistical method used to estimate the probability density function of a random variable. This technique involves smoothing a set of observations by creating a continuous probability density function based on a chosen kernel function (typically a bell-shaped one). The kernel function is then centred at each observation and used to estimate the probability density at that point. The estimated density is calculated by summing up the kernel functions centred at each observation point and dividing the result by the total number of observations. The bandwidth controls the smoothness of the density estimate (50). We used KDE to estimate and depict the density of the outlier scores attributed to samples by outlier detection methods in the corresponding section.

Outlier detection methods

Outlier detection requires distinguishing novel samples from samples belonging to known taxa. For that, we used the Local Outlier Factor (LOF) (18) algorithm, an unsupervised method that computes the local density deviation of a given data point with

respect to its neighbours. LOF considers the samples that have a substantially lower density than their neighbours as outliers. The score of the outliers (outlier) detection method was visualised with circles around samples with radii proportional to the absolute value of the scores (eq1. and Table3). Outlier candidates were evaluated using local outlier scores (proportional to the degree to which a data point deviates from its local neighbourhood, indicating its potential as an outlier) and KDE plots. For each case, we plotted the KDE plots and a scatter plot of the t-SNE dimensions. We plotted circles around the samples, with a radius calculated according to Eq. 1 wherein $i = 1, \dots, n$ is the index of the samples in the dataset, and *score* is the outlier score according to the outlier detection method of choice.

$$\text{Eq. 1.}\ radius_i = \frac{\max_i\big(score(x_i)\big) - score(x_i)}{\max_i\big(score(x_i)\big) - \min_i\big(score(x_i)\big)}, i = 1, \dots, n$$

In other words, the radius around each sample is proportional to its LOF score. The final decision on whether a sample is an outlier is made by comparing the LOF scores of the sample along with other samples belonging to the potentially new taxa.

In addition to the LOF algorithm, two additional algorithms were tested. First, Isolation Forest (70), an unsupervised machine-learning algorithm used for outlier detection. It works by creating a set of isolation trees, where each tree isolates an instance in the dataset by recursively splitting the feature space. Instances that require fewer splits to be isolated are considered more anomalous, while normal instances require more splits to be isolated (70). Second, One-class SVM (101), a machine-learning algorithm used for outlier detection. It works by creating a hyperplane that separates the normal instances from the outliers in the dataset. The algorithm learns a boundary that encloses a region in the feature space where most of the normal instances lie. Instances that fall outside this boundary are considered outliers (101).

Feature selection analysis to evaluate landmarks' informativeness

We used a genetic algorithm as a wrapper feature selection method (68) (51) (72) (60) with a logistic regression classifier. The algorithm was implemented using the Python module 'sklearn-genetic' (21) based on the Python module *DEAP* (40) (32). The operator for creating the next generation in each iteration was '*VarAnd,*' and the type of mating between every two individuals was '*Uniform Crossover*'. The initial population size was set to 1000. The algorithm stops when no fitter individual is produced after five generations. We used accuracy as the objective function of the optimisation task.

Estimating the citation number of PCA tools

Very conservatively, we estimate that, as of August 2023, between 18.000 to 32,900 morphometric studies employed PCA without employing a conventional classification tool of the field (LDA): 'landmarks AND morphometrics AND PCA -LDA' (18,000), 'morphometrics AND PCA -LDA' (32,900). We also estimated the citation count (12511) of PCA tools based on Google Scholar's citation count using the following searches: '"geomorph" AND PCA -LDA' (6,390) (6), '"morphoJ" AND PCA -LDA' (3770) (65), '"TPS (Thin-Plate Spline) series" AND PCA -LDA' (898) (100), 'morphologika and PCA -LDA' (584) (83), '"MorphpoTools" and PCA -LDA' (147) (66), '"Morphpo Tools" and PCA -LDA' (41), 'EVAN Toolbox and PCA -LDA' (124) (www.evan-society.org), respectively.

Limitations

PCA has a wide range of applications in geometric morphometrics where different variations of it are used. In some methods PCA is used in combination with or integrated in other methods as is the case with Phylogenetic Principal Component Analysis (Phy-PCA) (92) and Phylogenetically Aligned Component Analysis (PACA) (27). We did not investigate and address all these applications; however, the same approach presented here can be taken to investigate them.

Data and scripts availability

All our data and scripts that can replicate our results and figures are available via GitHub: https://github.com/NimaMohseni/Morphometric-Analysis

Acknowledgement

EE and NM were partially funded by the Swedish Research Council award to EE (2020-03485). The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lund, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

We would like to extend our heartfelt gratitude to Paul O'Higgins for his invaluable support throughout this research. His provision of landmark data, prompt responses to our

queries, and meticulous manuscript review significantly enriched the quality of this work, reflecting his passion and expertise in the field of morphometrics.

Declaration of interests

In accordance with ethical guidelines and the nature of our research, we hereby declare that the authors have no conflicting interests to disclose pertaining to this paper. There are no financial, personal, or professional affiliations that could potentially influence the findings or conclusions presented in this paper.

References

1. Meet Nesher Ramla Homo: New form of human found [press release]. The Times. 2021.
2. A Bitter Archaeological Battle Is Rocking Tel Aviv University [press release]. Haaretz2021.
3. A Previously Unknown Type of Ancient Human Has Been Discovered in The Levant [press release]. Science Alert2021.
4. Fossilised bones found in Israel could belong to mystery extinct humans [press release]. The Guardian2021.
5. Adams D, Cardini A, Monteiro L, O'higgins P, Rohlf F. Morphometrics and phylogenetics: principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. Journal of human evolution. 2011;60(2):240-243.
6. Adams DC, Otárola-Castillo E. geomorph: an R package for the collection and analysis of geometric morphometric shape data. Methods in ecology and evolution. 2013;4(4):393-399.
7. Arsuaga JL, Carretero J, Marti I, Gracia A. Cranial remains and long bones from Atapuerca/Ibeas (Spain). Journal of Human Evolution. 1991;20(3):191-230.
8. Bacon AM. Principal components analysis of distal humeral shape in Pliocene to recent African hominids: the contribution of geometric morphometrics. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists. 2000;111(4):479-487.
9. Bardua C, Felice RN, Watanabe A, Fabre A-C, Goswami A. A practical guide to sliding and surface semilandmarks in morphometric analyses. Integrative Organismal Biology. 2019;1(1):obz016.
10. Baylac M, Villemant C, Simbolotti G. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. Biological Journal of the Linnean Society. 2003;80(1):89-98.
11. Behnamian S, Esposito U, Holland G, Alshehab G, Dobre AM, Pirooznia M, et al. Temporal population structure, a genetic dating method for ancient Eurasian genomes from the past 10,000 years. Cell reports methods. 2022;2(8).
12. Bellman R. Dynamic programming. Princeton, New Jersey: Princeton University Press; 1957.
13. Berner D. Size correction in biology: how reliable are approaches based on (common) principal component analysis? Oecologia. 2011;166(4):961-971.
14. Biró Z, Lanszki J, Szemethy L, Heltai M, Randi E. Feeding habits of feral domestic cats (Felis catus), wild cats (Felis silvestris) and their hybrids: trophic niche overlap among cat groups in Hungary. Journal of Zoology. 2005;266(2):187-196.
15. Bookstein FL. Morphometric tools for landmark data. Cambridge, UK: Cambridge University Press; 1997.
16. Bookstein FL. The inappropriate symmetries of multivariate statistical analysis in geometric morphometrics. Evolutionary Biology. 2016;43:277-313.
17. Bookstein FL. Pathologies of between-groups principal components analysis in geometric morphometrics. Evolutionary Biology. 2019;46(4):271-302.
18. Breunig MM, Kriegel H-P, Ng RT, Sander J, editors. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD international conference on Management of data; 2000.
19. Brodersen KH, Ong CS, Stephan KE, Buhmann JM, editors. The balanced accuracy and its posterior distribution. 2010 20th international conference on pattern recognition; 2010: IEEE.
20. Brophy JK, Moggi-Cecchi J, Matthews GJ, Bailey SE. Comparative morphometric analyses of the deciduous molars of Homo naledi from the Dinaledi Chamber, South Africa. American Journal of Physical Anthropology. 2021;174(2):299-314.
21. Calzolari M. sklearn-genetic: sklearn-genetic 0.3.0. Zenodo. 2020.
22. Cardini A. Less tautology, more biology? A comment on “high-density” morphometrics. Zoomorphology. 2020;139(4):513-529.
23. Cardini A, Polly PD. Cross-validated between group PCA scatterplots: a solution to spurious group separation? Evolutionary Biology. 2020;47(1):85-95.

24. Chari T, Pachter L. The specious art of single-cell genomics. BioRxiv. 2022:2021.2008. 2025.457696.
25. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
26. Cobb SN, O'Higgins P. Hominins do not share a common postnatal facial ontogenetic shape trajectory. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2004;302(3):302-321.
27. Collyer ML, Adams D. Phylogenetically aligned component analysis. Methods in Ecology and Evolution. 2021(12):359–372.
28. Cooke SB, Terhune CE. Form, function, and geometric morphometrics. The Anatomical Record. 2015;298(1):5-28.
29. Cope ED. The origin of the fittest: Essays on evolution: D. Appleton; 1886.
30. Da Silva FO, Fabre A-C, Savriama Y, Ollonen J, Mahlow K, Herrel A, et al. The ecological origins of snakes as revealed by skull evolution. Nature Communications. 2018;9(1):376.
31. Das R, Wexler P, Pirooznia M, Elhaik E. The origins of Ashkenaz, Ashkenazic Jews, and Yiddish. Frontiers in genetics. 2017;8:87.
32. De Rainville F-M, Fortin F-A, Gardner M-A, Parizeau M, Gagné C, editors. Deap: A python framework for evolutionary algorithms. Proceedings of the 14th annual conference companion on Genetic and evolutionary computation; 2012.
33. Delson E, Harvati K, Reddy D, Marcus LF, Mowbray K, Sawyer G, et al. The Sambungmacan 3 Homo erectus calvaria: a comparative morphometric and morphological analysis. The Anatomical Record: An Official Publication of the American Association of Anatomists. 2001;262(4):380-397.
34. Donoho DL. High-dimensional data analysis: The curses and blessings of dimensionality. AMS math challenges lecture. 2000;1(2000):32.
35. Dryden IL, Mardia KV. Statistical shape analysis. New York, NY: John Wiley & Sons, Ltd; 1998.
36. Dryden IL, Mardia KV. Statistical shape analysis: with applications in R: John Wiley & Sons; 2016.
37. Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nature communications. 2014;5(1):3513.
38. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Scientific reports. 2022;12(1):1-35.
39. Flury B. Common principal components & related multivariate models: John Wiley & Sons, Inc.; 1988.
40. Fortin F-A, De Rainville F-M, Gardner M-AG, Parizeau M, Gagné C. DEAP: Evolutionary algorithms made easy. The Journal of Machine Learning Research. 2012;13(1):2171-2175.
41. Franklin D, Freedman L, Milne N, Oxnard C. Geometric morphometric study of population variation in indigenous southern African crania. American Journal of Human Biology. 2007;19(1):20-33.
42. Friedman ST, Price SA, Hoey AS, Wainwright PC. Ecomorphological convergence in planktivorous surgeonfishes. Journal of Evolutionary Biology. 2016;29(5):965-978.
43. Garrod DAE, Bate DMA. The Stone Age of Mount Carmel: Excavations at the Wady El-Mughara;[report of the Joint Expedition of the British School of Archaeology in Jerusalem and the American School of Prehistoric Research 1929-34]: Clarendon Press; 1937.
44. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006;63(1):3-42.
45. Gunz P, Mitteroecker P, Bookstein FL. Semilandmarks in three dimensions. Modern morphometrics in physical anthropology. 2005:73-98.
46. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357-362.
47. Hart PE, Stork DG, Duda RO. Pattern classification: Wiley Hoboken; 2000.

48. Harvati K. Quantitative analysis of Neanderthal temporal bone morphology using three-dimensional geometric morphometrics. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists. 2003;120(4):323-338.
49. Harvati K, R÷ ding C, Bosman AM, Karakostis FA, GrŘn R, Stringer C, et al. Apidima Cave fossils provide earliest evidence of Homo sapiens in Eurasia. Nature. 2019;571(7766):500-504.
50. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction: Springer; 2009.
51. Haupt RL, Haupt SE. Practical genetic algorithms: John Wiley & Sons; 2004.
52. Haykin S. Neural networks and learning machines, third edition: Pearson Education India; 2009.
53. Hershkovitz I, May H, Sarig R, Pokhojaev A, Grimaud-Hervé D, Bruner E, et al. A middle pleistocene homo from Nesher Ramla, Israel. Science. 2021;372(6549):1424-1428.
54. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics. 2015;2015.
55. Ho TK, editor Random decision forests. Proceedings of 3rd international conference on document analysis and recognition; 1995; Montreal, QC, Canada: IEEE.
56. Ho TK. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence. 1998;20(8):832-844.
57. Ioannidou M, Koufos GD, de Bonis L, Harvati K. A new three-dimensional geometric morphometrics analysis of the Ouranopithecus macedoniensis cranium (Late Miocene, Central Macedonia, Greece). American Journal of Physical Anthropology. 2019;170(2):295-307.
58. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013.
59. Jones N. Mysterious skull fossils expand human family tree - but questions remain. Nature. 2021;595(7865):20.
60. Jung M, Zscheischler J. A guided hybrid genetic algorithm for feature selection with expensive cost functions. Procedia Computer Science. 2013;18:2337-2346.
61. Kelleher JD, Mac Namee B, D'arcy A. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies: MIT press; 2020.
62. Kendall DG. Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London mathematical society. 1984;16(2):81-121.
63. 300,000-year-old skull found in China unlike any early human seen before [press release]. CNN2023.
64. Klingenberg CP, McIntyre GS. Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with Procrustes methods. Evolution. 1998;52(5):1363-1375.
65. Klingenberg CP. MorphoJ: an integrated software package for geometric morphometrics. Molecular ecology resources. 2011;11(2):353-357.
66. Koutecký P. MorphoTools: a set of R functions for morphometric analysis. Plant Systematics and Evolution. 2015;301:1115-1121.
67. Lague MR. The pattern of hominin postcranial evolution reconsidered in light of size-related shape variation of the distal humerus. Journal of human evolution. 2014;75:90-109.
68. Lanzi PL, editor Fast feature selection with genetic algorithms: a filter approach. Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97); 1997; Indianapolis, IN, USA: IEEE.
69. A new human species? Mystery surrounds 300,000-year-old fossil [press release]. nature2023.
70. Liu FT, Ting KM, Zhou Z-H. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD). 2012;6(1):1-39.
71. Lockwood CA, Lynch JM, Kimbel WH. Quantifying temporal bone morphology of great apes and humans: an approach using geometric morphometrics. Journal of Anatomy. 2002;201(6):447-464.
72. Lu J, Zhao T, Zhang Y. Feature selection based-on genetic algorithm for image annotation. Knowledge-Based Systems. 2008;21(8):887-891.

73. MacLeod N. Morphometrics: History, development methods and prospects. 动物分类学报. 2017;42(1):4-33.
74. Marom A, Rak Y. Comment on “A Middle Pleistocene Homo from Nesher Ramla, Israel”. Science. 2021;374(6572):eabl4336.
75. McKinney W, editor Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference; 2010; Austin, TX, USA.
76. Migut M, Worring M, Veenman CJ. Visualizing multi-dimensional decision boundaries in 2D. Data Mining and Knowledge Discovery. 2015;29(1):273-295.
77. Mitteroecker P, Gunz P, Bernhard M, Schaefer K, Bookstein FL. Comparison of cranial ontogenetic trajectories among great apes and humans. Journal of human evolution. 2004;46(6):679-698.
78. Mitteroecker P, Gunz P, Bookstein FL. Heterochrony and geometric morphometrics: a comparison of cranial growth in Pan paniscus versus Pan troglodytes. Evolution & development. 2005;7(3):244-258.
79. Mitteroecker P, Gunz P. Advances in geometric morphometrics. Evolutionary Biology. 2009;36:235-247.
80. Mitteroecker P, Bookstein F. Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. Evolutionary Biology. 2011;38:100-114.
81. Nazri A, Agbolade O, Yaakob R, Ghani AA, Cheah YK. A novel investigation of the effect of iterations in sliding semi-landmarks for 3D human facial images. BMC bioinformatics. 2020;21(1):1-10.
82. Nielsen F. Introduction to HPC with MPI for Data Science: Springer; 2016.
83. O'HIGGINS P, Jones N. Facial growth in Cercocebus torquatus: an application of three-dimensional geometric morphometric techniques to the study of morphological variation. The Journal of Anatomy. 1998;193(2):251-272.
84. O'Higgins P, Collard M. Sexual dimorphism and facial growth in papionin monkeys. Journal of Zoology. 2002;257(2):255-272.
85. Oliveira R, Godinho R, Randi E, Alves PC. Hybridization versus conservation: are domestic cats threatening the genetic integrity of wildcats (Felis silvestris silvestris) in Iberian Peninsula? Philosophical Transactions of the Royal Society B: Biological Sciences. 2008;363(1505):2953-2961.
86. Oxnard C, O’Higgins P. Biology clearly needs morphometrics. Does morphometrics need biology? Biological Theory. 2009;4(1):84-97.
87. Oxnard CE. One biologist's view of morphometrics. Annual Review of Ecology and Systematics. 1978;9(1):219-241.
88. Palci A, Lee MS. Geometric morphometrics, homology and cladistics: review and recommendations. Cladistics. 2019;35(2):230-242.
89. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-2830.
90. Perez SI, Bernal V, Gonzalez PN. Differences between sliding semi-landmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. Journal of anatomy. 2006;208(6):769-784.
91. Academic locked out of lab over fossil feud [press release]. The Times2021.
92. Polly PD, Lawing AM, Fabre A-C, Goswami A. Phylogenetic principal components analysis and geometric morphometrics. Hystrix. 2013;24(1):33.
93. Pozharskiy A, Aubakirova K, Gritsenko D, Tlevlesov N, Karimov NZ, Galiakparov N, Ryabushkina N. Genotyping and morphometric analysis of Kazakhstani grapevine cultivars versus Asian and European cultivars. Genetics and Molecular Research. 2020;19(1):gmr18482-gmr18482.
94. Quilodrán CS, Montoya-Burgos JI, Currat M. Harmonizing hybridization dissonance in conservation. Communications Biology. 2020;3(1):391.
95. Randi E. Detecting hybridization between wild species and their domesticated relatives. Molecular ecology. 2008;17(1):285-293.
96. Rasmussen CE, Nickisch H. Gaussian processes for machine learning (GPML) toolbox. The Journal of Machine Learning Research. 2010;11:3011-3015.

97. Reich D. Who we are and how we got here: Ancient DNA and the new science of the human past: Oxford University Press; 2018.
98. Rohlf FJ. Stereograms in numerical taxonomy. Systematic Biology. 1968;17(3):246-255.
99. Rohlf FJ. On the use of shape spaces to compare morphometric methods. Hystrix-the Italian Journal of Mammalogy. 2000;11(1).
100. Rohlf FJ. The tps series of software. Hystrix. 2015;26(1):9-12.
101. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural computation. 2001;13(7):1443-1471.
102. Slice DE. Modern morphometrics in physical anthropology: Springer Science & Business Media; 2006.
103. Smith WJ. Rapid assessment of the extent of feral introgression in British and Irish Rock Dove Columba livia populations. Ibis. 2023.
104. Sneath PH. Thirty years of numerical taxonomy. Systematic Biology. 1995;44(3):281-298.
105. Sokal RR. The principles and practice of numerical taxonomy. Taxon. 1963:190-199.
106. Taunk K, De S, Verma S, Swetapadma A, editors. A brief review of nearest neighbor algorithm for learning and classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS); 2019: IEEE.
107. Thomson JA. On growth and form. Nature Publishing Group UK London; 1917.
108. Have scientists discovered a new species of HUMAN? Ancient skull belonging to a child with no chin who lived 300,000 years ago suggests our family tree 'needs another branch' [press release]. The Daily Mail2023.
109. Torkkola K. Feature extraction by non-parametric mutual information maximization. Journal of machine learning research. 2003;3(7/8):1415-1438.
110. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(11).
111. Velez AD, Quam R, Conde-Valverde M, Martínez I, Lorenzo C, Arsuaga JL. Geometric morphometric analysis of the bony labyrinth of the Sima de los Huesos hominins. Journal of Human Evolution. 2023;174:103280.
112. Verardi A, Lucchini V, Randi E. Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (Canis lupus) by admixture linkage disequilibrium analysis. Molecular ecology. 2006;15(10):2845-2855.
113. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020;17(3):261-272.
114. Wu X, Pei S, Cai Y, Tong H, Zhang Z, Yan Y, et al. Morphological and morphometric analyses of a late Middle Pleistocene hominin mandible from Hualongdong, China. Journal of Human Evolution. 2023;182:103411.
115. Zelditch ML, Swiderski DL, Sheets HD. Geometric morphometrics for biologists: a primer: academic press; 2012.