Topological data analysis and multiple kernel learning for species identification of modern and archaeological small ruminants

Manon Vuillien, Davide Adamo, Emmanuelle Vila, Agraw Amane, Thierry Argant, Daniel Helmer, Marjan Mashkour, Abdelkader Moussous, Olivier Notter, Elena Rossoni-Notter, et al.

Topological data analysis and multiple kernel learning for species identification of modern and archaeological small ruminants

Manon Vuillien[a], Davide Adamo[a,b], Emmanuelle Vila[c], Agraw Amane[d,e], Thierry Argant[f], Daniel Helmer[c], Marjan Mashkour[g], Abdelkader Moussous[h], Olivier Notter[h], Elena Rossoni-Notter[h], Isabelle Théry[a], Marco Corneli[a,b]

[a] Université Côte d'Azur, UMR 7264 CEPAM, CNRS, Nice, France
[b] Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team, Nice, France
[c] Université Lumière Lyon II, UMR 5133 Archéorient, CNRS, Lyon, France
[d] Department of Microbial Cellular and Molecular Biology, Addis Ababa University, Ethiopia
[e] ILRI Livestock Genetics Program, Addis Ababa, Ethiopia
[f] Eveha études et valorisation archéologique, UMR 5138 ArAr, Lyon, France
[g] AASPE Muséum national d'Histoire naturelle, CNRS, Paris, France
[h] Musée d'Anthropologie préhistorique de Monaco, Monaco

Abstract

The faunal remains from numerous Holocene archaeological sites across southwest Asia frequently include the bones of various wild and domestic ungulates, such as sheep, goats, ibexes, roe deer and gazelles. These assemblages may provide insight into hunting and animal husbandry strategies and offer palaeoecological information on ancient human societies. However, the skeletons of these taxa are highly similar in appearance, which presents a challenge for accurate identification based on their bones. This paper presents a case study to test the potential of topological data analysis (TDA) and multiple kernel learning (MKL) for inter-specific identification of 150 3D astragali belonging to modern and archaeological specimens. The joint application of TDA and MKL demonstrated remarkable efficacy in accurately identifying wild species, with a correct identification rate of approximately 90%. In contrast, the identification of domestic species exhibited a lower success rate, at approximately 60%. This low rate of identification of sheep and goat species is attributed to the morphological variability of domestic breeds. Moreover, while these methods assist in clearly identifying wild taxa from one another, they also highlight their morphological diversity. In this context, TDA and MKL could be invaluable for investigating intra-specific variability in domestic and wild animals. These methods offer a means of expanding our understanding of past domestic animal selection practices and techniques. They also facilitate an investigation into the morphological evolution of wild animal populations over time.

Keywords

Machine learning, Osteology, Herbivores, Zooarchaeology, Topological data analysis, Multiple kernel learning

1. Introduction

Small wild and domestic ungulates are frequently found in Holocene archaeological faunas and are likely to be found together in some contexts. For example, in the Near East, wild and domestic *Caprinae*

(domestic or wild goat and sheep, ibex), roe deer, and gazelle have geographical distributions whose limits in ancient times are poorly defined and partly overlap for specific taxa (Uerpmann 1987). Each of these species shows adaptation to a particular ecological habitat and specific plant resources. Roe deer prefers areas of mixed forest and grassland. Ibex is a mixed feeder (browser and grazer) living in mountainous regions. Gazelle occurs in waterless steppe, semi-arid, and desert environments. These ungulates provide information on hunting and husbandry strategies and palaeoecological information on the climate and environment of ancient societies, the diversity of natural habitats, and regional variations in terms of aridification or vegetation cover capacity, agriculture, and deforestation (Tsahar *et al.* 2009). However, the skeletons of these taxa are very close morphologically, which poses a problem for specific identification based on their bones, as evidenced by numerous methodological studies conducted on these species (Fernandez 2001; Salvagno & Albarella 2017; Sipilä *et al.* 2023; Zeder & Lapham 2010; Zeder & Pilaar 2010) over the past fifty years. Furthermore, specific taxa, such as gazelles, face challenges in distinguishing between species due to the absence of anatomical criteria (Buitenhuis 1988; Gudea & Stan 2012; Peters 1989) or identifying sexual dimorphism (Munro *et al.* 2011) due to their morphological similarity.

Indeed, the taxonomic identification of remains of morphologically related species found in archaeological contexts represents one of the key challenges that zooarchaeologists face. Traditionally, the process of identifying bones or dental remains in archaeology is based on anatomical, morphological and biometric criteria. These taxa are compared with their modern or fossil counterparts documented in modern comparative osteological collections or represented in anatomical atlases. It is often the case that identification criteria are provided by the literature which have been tested and validated on large reference collections. Therefore, these criteria are very likely to be accurate. However, this process can still be challenging due to factors such morphological convergence within closely related species, potential absence of diagnostic criteria, intermediate morphological characteristics, and intra-individual variability. Consequently, the use of anatomical criteria available in the literature to differentiate these species is not always sufficient.

In recent decades, palaeogenetic (Alberto *et al.* 2018; Daly *et al.* 0018; Larsson *et al.* 2024; Lv *et al.* 2022) and palaeoproteomic (Fabrizi *et al.* 2024; Le Meillour *et al.* 2023; Pilaar Birch *et al.* 2019; Prendergast *et al.* 2019; Wadsworth *et al.* 2017) analyses have made a significant contribution to the identification of wild and domestic ungulates remains. Nevertheless, these techniques are not always applicable, as the condition of the faunal remains (*e.g.* poor preservation of DNA and ancient proteins, alteration or modification of the bone surface, lack of reference data) may render them unsuitable for use. Furthermore, they are expensive and can only be used for targeted issues involving a limited number of specimens, not to realize the entire identification of a zooarcheological collection. Moreover, over the past two decades, geometric morphometrics methods (GMM) have been employed in zooarchaeology to document numerous animal species undergoing domestication processes (Cucchi *et al.*, 2021; Cucchi *et al.* 2023; Evin *et al.* 2015), differentiate between morphologically similar taxa such as sheep and goats (Colominas *et al.* 2019; Haruda 2017; Vuillien 2020), and explore species-level variability, such as sheep (Haruda *et al.* 2019; Pöllath *et al.* 2019; Pöllath *et al.* 2019) and deer (Curran 2012). However, GMM are relatively time-consuming, and the observed morphological differences and similarities are based on two- or three-dimensional patterns representing part of the bone being studied and are not a representation of the entire bone.

The utilisation of machine learning (ML) approaches is being explored for the analysis of biological archives, including bone remains of terrestrial (*e.g.* Moclán *et al.* 2019) and marine mammals (*e.g.* Bickler 2021). Recent studies have evaluated the performance of ML and GMM 2D studies on upper and lower molars of modern and fossil mice (Miele *et al.* 2020; Moclán *et al.* 2023). These studies aim to propose novel identification criteria for these taxa to document the dynamics of human settlements and their role in the emergence and spread of the commensal house mouse (Cucchi *et al.* 2020). Another recent study combined classification methods, including artificial neural networks and GMM 2D/3D studies on several teeth and bones of wild and domestic equids and their hybrids (Mohaseb *et al.* 2023). The aim was to increase the identification of archaeological equid species and their hybrids in three archaeological sites located in the Middle East. The majority of ML approaches to the identification of ancient animal species rely on the use of two-dimensional images. Nevertheless, the use of three-dimensional imagery in zooarchaeology and, more generally, in archaeology (Andres *et al.* 2012; Wyatt-Spratt 2022) has become prevalent, thus enabling the construction and development of osteological and archaeological digital reference collections that ML methods can employ. In this context, the investigation of three-dimensional meshes and point clouds describing biological objects is particularly interesting to mathematicians, both in terms of the object itself and for the complex methodological developments that this represents (Botsch *et al.* 2010; Kazhdan *et al.* 2006; Zhao *et al.* 2021). Moreover, the utilisation of 3D imaging, also tried and tested, is pertinent to the issue of species identification, particularly in the context of morphologically similar species such as domestic and wild ruminants.

This contribution aims to explore the potential of ML approaches directly working on 3D scans and, in particular, on point clouds. Although several ML approaches could serve our purpose (*i.e.* the automatic taxonomic identification from 3D point clouds), some features of the available dataset limit the number of "feasible" approaches. In particular, (i) after the acquisition, the 3D bones have different orientations, scales and number points; (ii) the number of bones for each species is quite limited. Due to the first issue (i), it is difficult to successfully come up with a meaningful notion of distance or similarity between two bones. Several techniques were tested to automatically register (*i.e.* reduce to the same pose and find correspondences between points) collections of point clouds (Evangelidis & Horaud 2017; Myorenko & Song 2010) but, due to the difficulty of the task, such methods failed. Instead, the Iterative Closest Point algorithm (Zhang 2021) would successfully register the whole collection. Still, humans are required to intervene in order to fix a set of benchmarks on each bone manually. This task is long and tedious, and an important aim is to avoid any bias that could be introduced by human intervention at this step. The second feature (ii) prevents from exploiting deep learning architectures (such as (Feng *et al.* 2020; Qi *et al.* 2017), etc.), which need a significant amount of data (here, bones per species) to be appropriately trained. Moreover, deep learning methods would require a downsampling preprocess of the point clouds, whose size is prohibitive for standard architectures, which could lead to information losses.

In order to fully use the point clouds and highlight morphological features related to the species, as revealed by GMM studies, Topological Data Analysis approach (TDA, Chazal & Michel 2021) was chosen. TDA is a branch of mathematics that studies the structure and the topological properties of data. It has gained popularity in recent years due to its ability to uncover patterns in datasets that are not easily discernible through traditional ML methods (Calsson *et al.* 2008; Dequeant *et al.* 2008; Nicolau *et al.* 2011). The use of TDA as a descriptor for zooarchaeological bones provides a powerful

method for addressing the complexities of 3D data analysis. Indeed, as previously mentioned, TDA enables the use of the entire 3D scan, preserving the full point cloud and ensuring that no crucial information is lost. Moreover, the invariance of TDA to isometries such as translation, rotation, and reflection makes it particularly suitable for our study, ensuring that the extracted topological features are intrinsic to the 3D bones and not affected by their positioning.

Once the topological features of each (3D scan of a) bone are extracted with TDA, each specimen is classified. However, different topological features induce different notions of similarity between bones. Roughly speaking, although two bones are similar in terms of "connected components", for instance, they might differ in terms of "cycles". Thus, each topological feature (connected components, cycles, holes) is used to construct a so-called "core" matrix. For instance, the entry (i, j) in the kernel matrix of cycles measures how similar bones i and j are, in terms of cycles. The final objective is to assess the impact of each individual kernel/feature on the classification task (taxonomic identification) and possibly discard features that are redundant. This is precisely what multiple kernel learning (MKL, Gönen & Alpaydin 2011) does. A Bayesian formulation of a logistic regression classifier was used and an original stochastic variational inference approach (SVI, Hoffman *et al.* 2013) was developed to i) perform supervised classification of specimens and ii) evaluate the impact of each topological feature.

Research aim

The main contribution of this paper is twofold. Firstly, a machine learning routine is proposed to automatically identify morphologically related animal species, like small ruminant herbivores such as *sheep*, *goat*, *Alpine ibex*, *roe deer* and *gazelle*. This routine relies on topological data analysis. Secondly, an original statistical framework is presented, allowing to *weight* the extracted topological features in such a way as to exploit each one of them (*i.e.* multiple kernel learning).

Three additional remarks are needed. First, although the classification is *automatic*, the expertise of the zooarchaeologist is required for the analysis of the results to construct an analytical framework and gain an understanding of the way statistics work. Second, the above-mentioned species have been chosen for three main reasons: 1) their morphological proximity; 2) their simultaneous presence in certain archaeological contexts; 3) the large number of taxonomic criteria available in the literature to differentiate some of them, such as sheep and goats, compared to the lack of data for others, such as roe deer and gazelle. Third, the model described here is based on whole bones in a good state of preservation.

2. Material and methods

2.1. Astragalus

The anatomical part selected for this study is a short bone, the astragalus, from the tarsal joint of the foot **(Supplementary data 1)**. It preserves very well in archaeological faunas because it is a small bone, particularly compact and robust and rarely broken intentionally due to its low nutritional value (Barone 1976; Popkin *et al.* 2012). These bones presented several anatomical criteria for identifying wild and domestic ungulates discussed by multiple scholars for almost sixty years. The distinction between sheep and goats is well documented (Boessneck *et al.* 1964; Clutton-Brock *et al.* 1990; Fernandez 2001;

Prummel & Frish 1986; Salvagno & Albarella 2017; Zeder & Lapham 2010; Zeder & Pilaar 2010), but still poses problems (Sipilä *et al.* 2023) and represent a challenge as demonstrated by recent GMM (Gaastra et al 2023; Jeanjean *et al.* 2022; Lloveras *et al.* 2022; Pöllath *et al.* 2019; Vuillien 2020) and molecular studies (Jeanjean *et al.* 2023; Le Meillour *et al.* 2020). In addition, there are few criteria for distinguishing between roe deer, gazelle and ibex (Buitenhuis 1988; Crégut-Bonnoure, 2020; Fernandez 2001; Gudea & Stan 2012; Lavocat 1966; Peters 1989).

2.2. 3D models dataset

The dataset included 150 3D complete astragali belonging to five taxa: Alpine ibex (*Capra ibex*), sheep (*Ovis aries*), goat (*Capra hircus*), roe deer (*Capreolus capreolus*) and gazelle *(Gazella cuvieri, Gazella dorcas, Gazella spekei* and *Gazella* sp.) (**Supplementary data 2**). This dataset does not consider the specimens' geographical origin or provenance which do not concern our research topic. Gazelle species are also grouped at the genus level for statistical reasons. The specimens belong to National Museum of Natural History Mammalian and Birds collection of Paris, CEPAM (UMR 7264) and Archéorient (UMR 5133) labs zooarchaeological reference collection, modern sheep and goat collected for the EvoSheep collection (ANR-17-CE27-0004), modern goats form BALUT Laboratory Iran, archeological site of "Grotte de l'Observatoire" Museum of Prehistoric Anthropology of Monaco and archaeological site of "Tell Sheikh Hassan" Archeorient lab (UMR 5133) (**Table 1**). In order to provide a homogeneous group for ML analysis, each taxa is represented by 30 specimens (3D astragali).

Species	Variety	Curator	Number
Alpine ibex *(Capra ibex)*	Archaeological from Southern Alps (*Liguro-Provençal Bassin*)	Museum of Prehistoric Anthropology of Monaco - Archeological site of "Grotte de l'Observatoire"	29
	Modern from Alps	Osteological collection from Thierry Argant (Éveha Lyon, ArAr UMR 5138)	1
Goat (*Capra sp*.)	Modern *Capra nubiana* (zoological specimen)	National Museum of Natural History Paris' Mammalian and Birds collection	2
	Modern domestic goat from France	Osteological collections from Archéorient UMR 5133; National Museum of Natural History Paris' Mammalian and Birds collection	5
	Modern domestic goat from Iran	Osteological collections from Bioarchaeology Laboratory, University of Tehran, Iran	20
	Modern domestic goat from Egypt (zoological specimen)	National Museum of Natural History Paris' Mammalian and Birds collection	2
	Modern feral goat from Crete (*Capra aegagrus cretica*)	Osteological collections from Archéorient UMR 5133	1
Roe deer *(Capreolus*	Modern from France	National Museum of Natural History	30

capreolus)		Paris' Mammalian and Birds collection	
Gazelle (*Gazella sp*.)	Modern *Gazella cuvieri*	National Museum of Natural History Paris' Mammalian and Birds collection	2
	Modern *Gazella dorcas*	National Museum of Natural History Paris' Mammalian and Birds collection	5
	Modern *Gazella spekei*	National Museum of Natural History Paris' Mammalian and Birds collection	1
	Modern *Gazella sp.*	National Museum of Natural History Paris' Mammalian and Birds collection; Osteological collections from Archéorient UMR 5133	7
	Archaeological from Syria (*Gazella cf. subgutturosa?*)	Daniel Helmer - Emmanuelle Vila (UMR 5133 Archéorient)	15
Sheep (*Ovis aries*)	Modern sheep from France	Osteological collections from CEPAM UMR 7264, AASPE UMR 7209 & Archéorient UMR 5133	7
	Modern sheep from Ethiopia	ILRI - Agraw Amane - Emmanuelle Vila - EvoSheep projet (ANR ANR-17-CE27-0004)	23
		Total of 3D astragalus	**150**

Table 1: Summary of sampled modern and archaeological species.

The astragali were scanned using the Artec Spider blue LED surface scanner and Artec Studio reconstruction software (version 16) and EinScan Pro 2X (Figure 1). The 3D models are reconstructed at a resolution between 0.3 and 0.1 mm using a textured polygonal mesh. Meshes are exported in "ASCII.ply" and "obj." archiving format (Vergnieux *et al.* 2017).

2.3. Topological data analysis (TDA)

Although an in-depth presentation of TDA clearly is outside of the scope of this paper (the interested reader is referred to Chazal & Michel 2021), in this section are sketched the main ideas TDA relies on.

The input data here is a collection of 150 3D point clouds, and the aim is to extract some useful information (or features) from each point cloud in order to use it to assign the cloud/bone to a species. In order to extract the features, a "continuous" shape is built from the point cloud by progressively connecting data points that are closer to each other with an edge. Here, the reader can safely consider that two points are close if their Euclidean distance in the 3D space is smaller than a given threshold $\epsilon > 0$. As far as ϵ grows, more and more points are connected and the shape that appears is a collection of simplicial complexes (Figure 2). A simplicial complex can be seen as a higher-dimensional

generalization of a neighboring graph: whereas the latter only includes vertices and edges, the former also contains faces, namely triangles and tetrahedrons. The nested family of simplicial complexes that add to each other is called filtration and as long as the family grows, relevant topological features such as connected components, loops and voids are collected via specific methods, such as persistent homology (PH, (Otter *et al.* 2017; Zomorodian & Carlsson 2004) and stored into the so-called persistence diagrams (PDs), that is described in some detail in the next section. It has been shown (Chazal & Michel 2021) that the topological features provide insights into the underlying structure of the data, making TDA particularly useful for analysing highly dimensional and noisy datasets.

In order to provide the reader with an intuition of what TDA is, the exposition of this pipeline is simplified. However, some remarks are needed:

1. Since the way simplicial complexes are built mainly relies on the relative distance between the points in the cloud, the orientation of the 3D shapes is irrelevant and there is no longer a need to register the collection.
2. Several notions of distance between points can be chosen and several ways of aggregating simplicial complexes (*i.e.* filtrations) exist.
3. The topological features are collected in PDs, but alternatives exist (*e.g.* persistence images, barcodes, etc.).

These few remarks should help the reader to figure out the vast- ness and richness of topological data analysis.

2.3.1. Persistence Diagrams (PDs)

In all the experiments, PDs were created from 3D point clouds based on a particular filtration: the Alpha filtration (GUDHI project 2023; Rouvreau 2023). Its main ingredient is a simplicial complex (the Alpha complex) that is built upon the growing balls mechanism that where sketched in the previous section (Figure 2). As said, as the radius ϵ of each ball increases, more simplices fuse and the topology evolves. Persistent homology then tracks these changes, identifying the *birth* and *death* of topological features such as connected components (0-dimensional features), cycles (1-dimensional), and voids (2-dimensional). These “lifelongs” are then stored and visualised in a 2D diagram, the PD (Supplementary data 3). Each point's coordinates represent the scale (*i.e.* the value of ϵ) at which a feature appears (birth) and disappears (death) along the filtration process. Three different colors correspond to the different topological features: connected components, in red, loops, in blue and voids, in green. Notice that the value ∞ is allowed on the y-axis and a red point takes this value. It means that at the end of the filtration process a single and huge connected component is still alive, whereas all cycles and voids do not exist anymore.

PDs were shown to provide a valuable summary of the point cloud's underlying geometrical structure, being robust to perturbation of the data in the Gromov-Hausdorff metric (Chazal & Michel 2021, Theorem 9). The key underlying intuition is that if two bones have very similar shapes, then also their persistence diagrams will be similar and vice versa. Thus, once all the input point clouds are described by their PDs, next step is to compute a pairwise similarity matrix (kernel), whose entry (i, j) is a non-negative number quantifying the similarity between (the PD of) i and (the PD of) j. The kernel matrix

can then be used as the input for ML algorithms in order to automatically perform taxonomic identification (Section 2.5). Next section describes how the similarity between two PDs can be calculated.

2.4. Discrete optimal transport

To quantify similarities between PDs, several metrics have been developed (Biasotti *et al.* 2011; Efrat *et al.* 2001). It should be stressed that the notions of similarity and distance between PDs are two sides of the same coin. Indeed, in general given two data points x and y if the distance d(x, y) between them is known, a measure of similarity is obtained *via*

$$K(x,y) = e^{-d(x,y)/\lambda} \qquad \text{(1bis)}$$

for any real positive λ. That said, the study of distances between discrete probability distributions, using the Wasserstein distance (Lacombe *et al.*, 2018) from optimal transport has been resorted (OT, Cuturi 2013). Optimal transport provides an intuitive way to quantify the similarity between two probability distributions by considering the minimum cost required to transform one distribution into the other. Here, discrete probability distributions are given consideration. In more details, two persistent diagrams $P_x = \{x_1 \cdots, x_N\}$ and $P_y = \{y_1, \cdots, y_M\}$, with N and M denoting the number of points in each diagram are considered. One can easily define a probability distribution u_x (respectively \mu_y) by putting mass 1/N (1/M) over each point in $P_x(P_y)$.

Definition 1:

The Wasserstein distance of order **p** between $\mathrm{u_x}$ and $\mathrm{u_y}$ is given by

$$W_p(\mu_X, \mu_Y) = \left(\min_{\pi \in \Gamma(\mu_X, \mu_Y)} \sum_{i,j} d(x_i, y_j)\pi_{ij} \right)^{1/p}$$

where $\Gamma(\mu_x, \nu_x)$ denotes the set of joint probability mass functions (also called couplings or transport plans) on P_x \times P_y with marginals $\boldsymbol{\mu}$_x and $\boldsymbol{\nu}$_y, respectively, and $\mathrm{d}(\mathrm{x}, \mathrm{y})$ is the distance between points x and y in the underlying metric space.

Here, $d(x, y)$ are considered to be the Euclidean distance and set p=2. Intuitively, one has to imagine a total mass of 1 is split into N equal portions and distributed over the points of P_x. Now the aim is to carry all this mass from P_x to P_y under the constraint that, at the end of the day, the total mass is uniformly distributed over the points of P_y. The Wasserstein distance quantifies the optimal cost of such an operation, and the optimal transport plan (the one minimising Eq. (1)) tells how much weight one should move from where to where in order to minimise the effort.

However, the PDs gives an additional issue: a green point (for instance) on P_x could be partially or totally transported to a red or a blue point on P_y, whereas the aim is to keep different topological features (connected components, loop and voids) well separated. Thus, a PD is “splitted” into three

PDs, one for each color (see Figure 4) and compute three Wasserstein distances for each pair of bones, one for each topological feature represented in the PDs.

Examples of pairwise kernel similarity matrices can be seen in Figures 3 and 5. Kernels were obtained from the Wasserstein distances via Eq. (1), where λ was set equal to the standard deviation of the corresponding distance matrix. The rows and columns of the kernel matrices correspond to the 150 bones and the *i*-th row and *j*-th column elements denote the Wasserstein similarity between the PDs associated with the *i*-th and *j*-th bones, for the corresponding topological features. The main diagonal is the brightest region of each kernel matrix since each bone is trivially at similarity one (and zero distance) from itself. Darker regions in a matrix correspond to higher distances. To narrow down the focus, the 1 and 2-dimensional features (respectively blue and green points in Figure 3) are only considered, which proved to be more informative.

2.5. Supervised multiple kernel learning

The blueprint of supervised machine learning can be briefly described as follows. Assume a training dataset of {x1, . . . , xN } observations is given, together with labels {y1, . . . , yN }. Here, the i-th observation xi is a 3D scan of a bone, in the form of a point cloud, and its label yi can be seen as an integer ranging from 1 to Q and labelling the species of the bone. As seen in Section 2.2, N = 150 and Q = 5 for us. Next ingredient is a function of the data (the classifier), say fθ, depending on some parameters θ and associating to each observation xi a predicted label fθ(xi) =: yˆi, *i.e.* another integer between 1 and Q. then fθ is "trained" by solving the following minimisation problem

$$\min_{\theta}\left(\sum_{i=1}^{N} L(y_i, \hat{y}_i)\right).$$

where $L(\cdot, \cdot)$ is a loss function. So, roughly speaking, the value of ϑ should be such that the mismatch between the predicted and the actual labels is minimal, on average. Once θ is optimised, the final aim is to be able to correctly predict the label $y*$ of a new test data point $x*$, via $f\theta(x*)$. In order to check that it is actually the case, in these experiments, all dataset (N = 150) has been split into train (Ntrain = 120) and test (Ntest = 30) and the accuracy (*i.e.* the proportion of correctly identified specimens) reported on the test dataset. To assess the robustness of the test accuracy, 50 random train/test splits are performed allowing the computation of a mean accuracy and a standard deviation.

Now, one way to define fθ is to pass through kernel matrices. Several ML classifiers are specifically designed to leverage the representation of the data through kernel matrices. Popular methods include support vector machines (SVM), kernel discriminant analysis (KDA) and kernel logistic regression (KLR). An in-depth description of such methods is outside the scope of this paper and the interested reader is referred to (Hastie *et al.* 2009, Chapters 4,5 and 12).

In this work, the focus is on multi-class KLR with multiple kernels. Indeed, as mentioned above, different types of filtrations and/or topological features lead to different similarity matrices, each of which could contain different discriminant information and relying on a single kernel matrix chosen by

the user could be not the optimal strategy. Multiple kernel learning (MKL) approaches [47] are specifically designed to manage this kind of situation: they allow one to properly weigh the input kernels and possibly discard the useless ones. Formally, D different kernel matrices {K(1), ..., K(D)} with K(d) ∈ RN×N were considered for each d. The aim is for an optimal convex combination of such matrices, defined by

$$K := \sum_{d=1}^{D} \beta_d K^{(d)} \in \mathbb{R}^{N \times N} \text{ with } \beta_d \geq 0 \text{ for each } d. \qquad (2)$$

By adopting a hybrid Bayesian formulation, the kernel weights βd are treated as random variables, following an Exponential prior distribution, whereas the other weights intervening in the KLR are treated as parameters to optimise. This approach allows to revisit MKL in an original way, being a compromise between pure optimization strategies (Rakotomamonjy *et al.* 2008) and fully Bayesian ones (Damoulas & Girolami 2008; Gonen 2012). Moreover, recent developments in importance-weighted stochastic variational inference (Sobolev & Vetrov 2019) have enabled the optimization problem in a fully differentiable manner, via Stochastic Gradient Descent (SGD, Bottou 2010), and perform posterior inference on β1, . . . , βD.

3. Results

3.1 Distance matrices (TDA) and optimal transport distances

The topological dimension 1 (loops) without normalisation indicates that Alpine ibex, roe deer, and gazelle respectively form dense clusters, in contrast to sheep and goat (Figure 3a). Sheep and goat exhibit an intra-specific significant topological heterogeneity, revealed by a marked color gradient from dark blue to yellow. This is mainly due to both the breed factor and the geographical heterogeneity of the specimens studied. For sheep, specimens from Ethiopian breeds (in dark green and yellow) are distinct from specimens from French breeds (in dark blue) (Figure 4). The same is true for goat. Specimens attributed to breeds of Iranian origin (in dark blue) differ from specimens of French breeds (in green and yellow) except for one specimen (number 57) (Supplementary data 4). Furthermore, although ibex form a homogeneous cluster, one individual stands out (dark blue): this is the only modern ibex in the dataset. When comparing species, the same topological dimension reveals a high degree of separation between wild and domestic species. In more details: ibex form a single community; roe dee look quite similar to gazelle (two wild taxa) and sheep often look indistinguishable from goat (two domestic species) despite no apparent link to their breed or geographical origin. Nevertheless, there is a certain topological proximity between roe deer and gazelle, on one side, and specimens of French sheep and goat, on the other (whereas the morphology of gazelle and roe deer is clearly distinguishable from that of African sheep and Asian goat).

Also, the topological dimension 2 (holes) without normalisation highlights differences between Alpine ibexes and other species (Figure 3b). These differences could be explained by the size of the astragalus. Ibex are taller than the other species in the dataset. However, if the size effect were crucial, gazelle, the smallest taxon in the dataset, should be clearly distinguishable from the other taxa. This is not the case. Consequently, although this parameter is undoubtedly important, it does not appear to be the sole factor responsible for the structuring of the dataset. Interestingly, whereas in the previous matrix roe deer and gazelle looked quite similar, here the difference between them is more accentuated

(although an increased similarity with goat appears). This point clearly illustrates why the adoption of several kernel matrices is beneficial to the taxonomic identification at the species level. The topological dimension 1 (loops) with bone normalisation allows for the clear distinction of Alpine ibex, roe deer and gazelle from groups of domestic caprine (sheep and goat) (Figure 5a). This outcome indicates that normalisation does not directly impact classification at the inter-specific level.

Finally, the topological dimension 2 (holes) with bone normalisation might seem of no particular interest (Figure 5b) since the similarities between specimens of each species render it impossible to distinguish between the species in question (except for roe deer, partially). However, as demonstrated in the next section, supervised MKL remarkably benefits from this matrix since it is the only one highlighting similarities between specimens of the same species far from each other in previous representations, especially sheep and goat.

3.2 Classification: Supervised MKL and automatic taxonomic identification of bones

For the supervised MKL part, as mentioned, the whole dataset (N = 150 bones) was randomly divided into 120 (80%) as train dataset and 30 (20%) as test dataset. It is recalled that such a random split is repeated 50 times. Table 2 reports the average test accuracy together with its standard deviation (in small characters). The second column reports the global test accuracy, whereas the remaining columns show the test accuracies for each species.

	$\text{Acc.}_{\text{std.dv.}}$	Alpine ibex	Goat	Roe deer	Gazelle	Sheep
Multiple KLR	$0.811_{0.064}$	$0.973_{0.061}$	$0.623_{0.218}$	$0.927_{0.101}$	$0.897_{0.124}$	$0.637_{0.228}$

Table 2: Average test classification accuracy.

In order to better inspect how the test accuracy behaves as a function of the train/test data split, a Box-and- Whisker plot is provided in Figure 6a. From the bottom to the top, each Box-and-Whisker reports for each species the minimum test accuracy, the lower quartile, the median accuracy, the upper quartile, and the maximum test accuracy. Outliers are represented as single points. The only modern specimen present in the Alpine ibex dataset is the outlier. The gazelle outlier corresponds to two specimens: a modern specimen (specimen number 1992 1844) and an archaeological specimen from Tell Sheikh Hassan (specimen number TSH 7Z23).

The average test accuracy of 81,1% (sensibly higher than 15% that one would obtain from a random assignment) is boosted by the accuracy on the wild species, which our approach can successfully identify. Figure 6a shows that the median test accuracy on wild species is 100%. The mean/median accuracies for sheep and goat are sensibly lower and come with a higher variability. However, as pointed out when describing the kernel matrices in Section 3.1, the non-homogeneous intra-specific blocks for these two species correspond to different breeds.

Table 3 shows the weights β1, .., β4, introduced in Eq. (2) and estimated via importance-weighted stochastic variational inference (IW-SVI). In more detail, our Bayesian inference strategy allows us to sample from the approximate posterior distribution of β1, .., β4, given the data and the other model

parameters. Table 3 reports the posterior mean of each weight with its standard deviation (in small character). Kernel density estimates of weights' posterior distributions can be seen in Figure 6b. The mode of each estimated density (*i.e.* the maximum a posteriori estimate of the corresponding weight) is away from zero, meaning that multiple KLR exploits all the topological features considered, either for normalized or unnormalized point clouds. Interestingly, the most important weight is put on β4, corresponding to the apparently less expressive kernel matrix of Figure 8b. This happens precisely because this kernel matrix is the only one allowing the algorithm to densify the clusters of sheep and goat, respectively. Furthermore, the weight given to each matrix indicates the influence of size on classification. If size was the more discriminating factor, then β1 (non-standardised matrix) should have a greater weight than β3 (standardised matrix). The results show the opposite (Table 3), indicating that bone size is not the more discriminating factor for the final classification.

	Kernel weights (posterior mean)			
	β_1	β_2	β_3	β_4
IW-SWI	0.231 $_{0.007}$	0.233 $_{0.006}$	0.275 $_{0.010}$	0.481 $_{0.020}$

Table 3: Kernel weight's estimates.

This can clearly be seen in Figure 7 which shows the learned kernel matrix K introduced in Eq. (2) and obtained as a linear combination of the Wasserstein kernels, weighted via the optimal betas shown in Table 3. K exhibits brighter diagonal blocks (especially for wild species) and the blocks corresponding to sheep and goats look more similar to dense blocks with respect to what Figures 3 and 5 show.

4. Discussion

4.1. Inter-specific identification: wild VS domestic

About the main research question of the inter-specific identification of wild and domestic small ruminants, TDA proved to be very proficient at correctly identifying wild specimens from a 3D scan of their astragalus, whereas it suffers with the identification of domestic specimens.

In other words, TDA makes it possible to clearly distinguish ibex, roe deer, and gazelle astragalus from sheep and goat astragalus, which underlines its value. Indeed, it is often difficult to distinguish these wild animals from sheep and goats in archaeological contexts where all these species are represented and where anatomical criteria are not discriminating. In contrast, the use of TDA does not address the challenges faced by the zooarchaeological community in distinguishing sheep from goats. Here, the result is dependent on the discriminative capacity of TDA and the dataset; the obtained classification exhibits the intraspecific morphological variability of the selected sheep and goat breeds.

Nevertheless, the fact that the median accuracy is higher than the mean for the wild species and the presence of outliers in Alpine ibex and gazelle is explained by the heterogeneity of our dataset: a high intra-specific variability of a few specimens lowers the mean accuracy. For instance, the single modern Alpine ibex in the collection turns out to be dissimilar from the other ibexes. For example, the unique

modern Alpine ibex that we have in our collection. As we saw in the figures in Section 3, it is dissimilar from the other ibexes; with regard to the train dataset, multiple KLR can safely identify all the archaeological ibex in the test. Conversely, when the modern specimen is included in the test dataset, the classifier fails to identify it (since trained on archaeological ibex dated to the Upper Pleistocene), and turns it into an outlier in the Alpine ibex group. This outcome is consistent with the difficulties encountered by zooarchaeologists in using modern datasets to identify ancient animal populations. This is due to the contrast between the number of current species and subspecies compared to fossil species and their morphological diversity (Crégut-Bonnoure 2020; Crégut-Bonnoure & Fernandez 2018; Urena *et al.* 2018). However, our findings indicate that this approach should be employed with great caution when attempting to address the question of the evolution of morphology. Indeed, upon noting the presence of two outliers within the gazelle group, a more detailed examination reveals that interpreting these differences in relation to the available temporal origin information is challenging. The dissimilarities between modern and archaeological gazelles (Holocene, between 9000 and 6000 BC) are not immediately apparent: TDA makes it difficult to identify variability among the gazelle species. This is likely due to the dataset itself, which does not accurately reflect the morphological variability among the modern species (*Gazella cuvieri, Gazella dorcas, Gazella spekei*). However, we intend to compare available proteomic data (Culley *et al.* 2021; Janzen *et al.* 2021; Le Meillour *et al.* 2020; Le Meillour *et al.* 2023) and future morphological and morphometric studies (Vuillien 2024).

4.2. Beyond the classification problem: is astragalus a good taxonomic marker?

The results obtained for species classification prompt the question of whether the astragalus can be employed as an interspecific identification marker. The geographical provenance of domestic and wild specimens appears to exert a more pronounced impact than the distinction between species. As we demonstrated, TDA cannot see sheep and goats as uniform and separated clusters (within the limits of the explored filtrations) and this fact has some significant consequences for the weighting of the kernels. The misidentification of sheep and goat species is attributed to the morphological variability of domestic breeds and their geographical origin. In addition, the difference in the ibex group observed between the single modern specimen and the other archaeological specimens may be correlated with the morphological evolution of the Alpine ibex over time and would point to the potential of the astragalus as an ecomorphological marker (Barr 2014; DeGusta 2003; Plummer *et al.* 2008). The biometrical and biomolecular studies carried out for this species show its morphotypic diversity over time, linked to environmental changes, the rocky areas frequented and human pressure (Crégut-Bonnoure 2020). However, it is still difficult to identify these factors precisely from faunal remains. TDA could be a valuable tool for exploring morphological variability at the intra-specific level for domestic and wild species and their adaptation across time and the environment, such as recent GMM studies applied to the same modern sheep breeds (Bader *et al.* 2022) and archaeological Alpine ibex populations. In archaeozoology, identifying and classifying of domestic sheep and goat morphotypes is of great importance. Such an analysis can provide insight into the evolution of zootechnical practices, economic development, and human society (Vila *et al.* 2021).

4.3. More robust assessment of the result obtained via TDA

To correctly estimate the potential of the method proposed in this paper, it will be crucial in the future to analyse the same dataset with other approaches either directly based on human expertise to have a “human” accuracy relying on anatomical criteria or automatic, such as GMMs. This paper aimed to test TDA and supervised MKL on a zooarchaeological dataset, nevertheless, TDA features would

certainly express the best of their potential in conjunction with other features, such as anatomical criteria and/or GMM patterns.

Another crucial aspect of machine learning routines in general is *explainability*: to know whether a classifier can be trusted, we need to understand how it works (on this topic, see Rudin 2019). Unfortunately, whereas TDA keeps track of the lifelong of topological features such as connected components, loops, and holes, it is currently not possible to know where these features are located on the surface of a 3D bone. This makes it very difficult, for instance, to assess whether the similarities that our approach detected between French sheep and roe deer/gazelle are based on meaningful and previously unexplored morphological patterns or if they are due to some other geometric components. Although some model agnostic explaining techniques exist (Lundberg & Lee 2017; Ribeiro *et al.* 2016), their use in the context of 3D point clouds, in conjunction with TDA, is not immediate at all. This issue will be addressed in future research to propose new anatomical features. Finally, it will be of interest to test deep learning models. Although requiring a huge amount of training data, such models can obtain impressive results, and their behaviors could be more easily captured than the one of TDA, either via explaining techniques or ad-hoc architectures.

Conclusion and perspectives

This paper mainly focuses on the taxonomic identification of wild and domestic small ruminants from 3D scans of complete astragalus of modern and archaeological specimens. The problem was framed as a supervised learning problem and addressed using TDA, optimal transport and an original inference routine. From one side, the topological features extracted with TDA proved to be very discriminant in classifying wild species (Alpine ibex, roe deer, gazelle). The main strengths of the proposed approach are a median test accuracy of 100% for these species and the fact that our routine is entirely automated (the expert's intervention is only required to analyse the results). On the other hand, TDA/MKL partly failed to identify the modern domestic species goat and sheep. However, an in-depth analysis of the reasons for such difficulty revealed that TDA might be better suited for the intra-specific classification of such species, for which our method seems likely to perceive a lot of detail. Another drawback of TDA methods is their lack of explainability: it is not possible to know which part of the bone contributed the most to the identification. In light of the above remarks, a few avenues for future research can be outlined: (i) the creation of ad-hoc datasets of 3D scans to assess the capabilities of TDA/MKL intra-specific classifiers (ii) combine TDA features to anatomical criteria and GMM morphological patterns; (iii) test deep learning methods on an increasing dataset to test others classification approaches. In conclusion, this research demonstrates the effectiveness of the TDA/MKL methods in extracting authentic biological information that can be interpreted by archaeozoologists, as well as novel information that cannot be detected by traditional anatomical criteria. In this sense, these approaches represent a milestone in the dialogue between two scientific disciplines, mathematics and archaeology, by providing information comparable to that obtained by palaeogenetics.

Acknowledgements

This paper constitutes the doctoral thesis in applied mathematics of the co-first author and as such this project has received financial support from the CNRS through the MITI interdisciplinary programs. It also forms a part of the postdoctoral fellowship "IBEX" supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the

Initiative of Excellence of Université Côte d'Azur under reference number ANR-15-IDEX-01. We would like to thank Arch'AI'Story project (Ministère de l'Enseignement Supérieur et de la Recherche and University Côte d'Azur) for funding this project. We would like to extend our gratitude to the researchers from the Museum of Prehistoric Anthropology in Monaco, the Laboratoire Départemental de Préhistoire du Lazaret in Nice (France), the National Museum of Natural History in Paris (France) and BioArch (UMR 7209, Paris) and Archéorient (UMR 5133, Lyon) research laboratories for their invaluable contributions. We are grateful to the ANR EvoSheep project (ANR-17-CE27-0004) for providing access to the modern sheep and goat collection collected in 3D during the first postdoctoral position of the first author to use part of it in this study. We are grateful to the collective research project '"Paleoecology of the Lazaret Cave: human-environment interactions on the coast of the meridional Alps during the late Middle Pleistocene (MIS 6)," granted by the DRAC PACA (French Ministry of Culture) that supported work in Pleistocene archaeological collection and provided the access to the 3D Einscan. We would like to express our gratitude to Emmanuel Desclaux for his support during the 3D modelling of Alpine ibex collection. We also express our gratitude to the "Imagery platform in bioarchaeology" coordinated by Thomas Cucchi at the BioArch laboratory. We would also like to thank Joséphine Lesur for her assistance and authorisation to provide samples of the modern gazelle and roe deer collection held at the National Museum of Natural History of Paris.

Declaration of interest: none

References

Adcock, A., Carlsson, E. and Carlsson, G. 2013. The ring of algebraic functions on persistence bar codes. *arXiv preprint arXiv:1304.0530*.

Alberto, F.J., Boyer, F., Orozco-terWengel, P., Streeter, I., Servin, B., De Villemereuil, P., Benjelloun, B., Librado, P., Biscarini, F., Colli, .L, Barbato M., Zamani W., Alberti A., Engelen S., Stella A., Joost S., Ajmone-Marsan P., Negrini R., Orlando L., Reza Rezaei H., Naderi S., Clarke L., Flicek P., Wincker P., Coissac E., Kijas J., Tosser-Klopp G., Chikhi A., Bruford W. M., Taberlet P., Pompanon F. 2018. Convergent genomic signatures of domestication in sheep and goats. *Nature communications* 9(1): 813. DOI: 10.1038/s41467-018-03206-y

Andrés, AN., Pozuelo, FB., Marimón, JR. and de Mesa Gisbert, A. 2012. Generation of virtual models of cultural heritage. *Journal of Cultural Heritage* 13(1): 103–106.

Bader, C., Mallet, C., Chahoud, J., Amane, A., De Cupere, B.., Berthon, R, Lavenne, F., Mohaseb, A., Davoudi, H., Albesso, M., Fathi H., Vuillien M., Lesur J. Helmer D., Gourichon L., Hanotte O., Mashkour M., Vila E., Cucchi T. 2022. Are petrous bones just a repository of ancient biomolecules? Investigating biosystematic signals in sheep petrous bones using 3D geometric morphometrics. *Journal of Archaeological Science: Reports* 43: 103447. DOI : 10.1016/j.jasrep.2022.103447.

Barone, R. 1976. *Anatomie comparée des mammifères domestiques*. Ostéologie deuxième édition revue et augmentée. Vigot. Paris.

Barr, WA. 2014. Functional morphology of the bovid astragalus in relation to habitat: Controlling phylogenetic signal in ecomorphology. *Journal of Morphology* 275(11): 1201–1216. DOI: https://doi.org/10.1002/jmor.20279.

Biasotti, S., Cerri, A., Frosini, P. and Giorgi, D. 2011. A new algorithm for computing the 2-dimensional matching distance between size functions. *Pattern Recognition Letters* 32(14): 1735–1746.

Bickler, SH. 2021. Machine Learning Arrives in Archaeology. *Advances in Archaeological Practice* 9(2): 186–191. DOI: https://doi.org/10.1017/aap.2021.6.

Boessneck, J., Müller, H-H. and Teichert, M. 1964. *Osteologische Unterscheidungsmerkmale zwischen Schaf (Ovis aries Linné) und Ziege (Capra hircus Linné)*. Verlag nicht ermittelbar.

Bonneel, N., Rabin, J., Peyré, G. and Pfister, H. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51: 22–45.

Botsch, M., Kobbelt, L., Pauly, M., Alliez, P. and Lévy, B.. 2010. *Polygon mesh processing*. CRC press.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. 2010. Springer. pp. 177–186.

Buitenhuis, H. 1988. *Archeozoölogisch onderzoek langs de Midden-Eufraat: onderzoek van het faunamateriaal uit zes nederzettingen in Zuidoost-Turkije en Noord-Syrië daterend van ca. 10.000 BP tot 1400 AD*.

Carlsson, G., Ishkhanov, T., De Silva, V. and Zomorodian, A. 2008. On the local behavior of spaces of natural images. *International journal of computer vision* 76: 1–12.

Chazal, F. and Michel, B. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence* 4: 667963.

Clutton-Brock, J., Dennis-Bryan, K., Armitage, PL. and Jewell, PA. 1990. Osteology of the Soay sheep. *Bulletin british Museum natural History Zoology* 1(56): 1–56.

Colominas, L., Evin, A., Burch, J., Campmajó, P., Casas, J., Castanyer, P., Carreras, C., Guardia, J., Olesti, O., Pons, E., Tremoleda, J. and Palet, J-M. 2019. Behind the steps of ancient sheep mobility in Iberia: new insights from a geometric morphometric approach. *Archaeological and Anthropological Sciences* 11(9): 4971–4982. DOI: https://doi.org/10.1007/s12520-019-00837-0.

Crégut-Bonnoure, E. 2020. *Les Ovibovini, Caprini et Ovini (Mammalia, Artiodactyla, Bovidae, Caprinae) du Plio-Pléistocène d'Europe*. BAR Publishing International Series. BAR International Series.

Crégut-Bonnoure, E. and Fernandez, P. 2018. Perspectives morphométriques et phylogéniques du genre Capra au Pléistocène (Mammalia, Artiodactyla, Caprinae). *Quaternaire* 29(3): 243–254.

Cucchi, T., Domont, A., Harbers, H., Evin, A., Alcàntara Fors, R., Saña, M., Leduc, C., Guidez, A., Bridault, A., Hongo, H., Price M., Peters J., Briois F., Guilaine J. Vigne J.-D. 2021. Bones geometric morphometrics

illustrate 10th millennium cal. BP domestication of autochthonous Cypriot wild boar (*Sus scrofa circeus nov*. ssp). *Scientific Reports* 11(1): 11435. DOI: 10.1038/s41598-021-90933-w

Cucchi, T., Harbers, H., Neaux, D., Balasse, M., Garbé, L., Fiorillo, D., Bocherens, H., Drucker, D., Zanolli, C., Cornette, R., Arbogast R.-M., Bréhard S., Bridault A. Gourichon L., Guilaine J., Manen C., Perrin T., Schafberg R., Tresset A., Vigne J.-D., Herrel A. 2023. 4500 years of morphological diversification in Western Europe wild boars (*Sus scrofa*) and the consequences of the Neolithic transition. *Quaternary Science Reviews* 309: 108100. DOI: 10.1016/j.quascirev.2023.108100

Cucchi, T., Papayianni, K., Cersoy, S., Aznar-Cormano, L., Zazzo, A.., Debruyne, R, Berthon, R., Bălăşescu, A., Simmons, A., Valla, F., Hamilakis Y., Mavridis F. Mashkour M., Darvish J., Siahsarvi R., Biglari F., Petrie A. C., Weeks L., Sardari A., Maziar S., Denys C., Orton D., Jenkins E., Zeder M., Searle J. B., Larson G., Bonhomme F., Auffray J.-C., Vigne J.-D. 2020. Tracking the Near Eastern origins and European dispersal of the western house mouse. *Scientific reports* 10(1): 8276. DOI: 10.1038/s41598-020-64939-9

Culley, C., Janzen, A., Brown, S., Prendergast, M.E.., Shipton, C., Ndiema, E, Petraglia, M.D., Boivin, N. and Crowther, A. 2021. Iron Age hunting and herding in coastal eastern Africa: ZooMS identification of domesticates and wild bovids at Panga ya Saidi, Kenya. *Journal of Archaeological Science* 130: 105368. DOI: https://doi.org/10.1016/j.jas.2021.105368.

Curran, SC. 2012. Expanding ecomorphological methods: geometric morphometric analysis of Cervidae post-crania. *Journal of Archaeological Science* 39(4): 1172–1182.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26, NIPS'13: Proceedings of the 27th International Conference on Neural Information Processing Systems: 2292-2300.

Daly, K.G., Maisano Delser, P., Mullin, V.E., Scheu, A., Mattiangeli, V., Teasdale, M.D., Hare, A.J., Burger, J., Verdugo, M.P., Collins, M.J., Kehati, R., Erek, C.M., Bar-Oz, G., Pompanon, F., Cumer, T., Çakırlar, C., Mohaseb, A.F., Decruyenaere, D., Davoudi, H., Çevik, Ö., Rollefson, G., Vigne, J.-D., Khazaeli, R., Fathi, H., Doost, S.B., Rahimi Sorkhani, R., Vahdati, A.A., Sauer, E.W., Azizi Kharanaghi, H., Maziar, S., Gasparian, B., Pinhasi, R., Martin, L., Orton, D., Arbuckle, B.S., Benecke, N., Manica, A., Horwitz, L.K., Mashkour, M. and Bradley, D.G. 2018. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science* 361(6397): 85–88. DOI: https://doi.org/10.1126/science.aas9411.

Damoulas, T. and Girolami, MA. 2008. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics* 24(10): 1264–1270.

DeGusta, D. and Vrba, E. 2003. A method for inferring paleohabitats from the functional morphology of bovid astragali. *Journal of Archaeological Science* 30(8): 1009–1022. DOI: https://doi.org/10.1016/S0305-4403(02)00286-8.

Dequeant, M.-L., Ahnert, S., Edelsbrunner, H., Fink, T.M., Glynn, E.F., Hattem, G., Kudlicki, A., Mileyko, Y., Morton, J., Mushegian, A.R., Lior P. Rowicka M., Shiu A., Sturmfels B., Pourquié O. 2008. Comparison of pattern detection methods in microarray time series of the segmentation clock. *PLoS One* 3(8): e2856. DOI: 10.1371/journal.pone.0002856

Efrat, A., Itai, A. and Katz, M.J. 2001. Geometry helps in bottleneck matching and related problems. *Algorithmica* 31: 1–28.

Evangelidis, G.D. and Horaud, R. 2017. Joint alignment of multiple point sets with batch and incremental expectation-maximization. *IEEE transactions on pattern analysis and machine intelligence* 40(6): 1397–1410.

Evin, A., Girdland Flink, L., Balasescu, A., Popovici, D., Andreescu, R., Bailey, D., Mirea, P., Lazar, C., Boroneant, A., Bonsall, C., Strand Vidarsdottir, U., Brehard, S., Tresset, A., Cucchi, T., Greger, L. and Dobney, K.. 2014. Unravelling the complexity of domestication: a case study using morphometrics and ancient DNA analyses of archaeological pigs from Romania. *Philosophical Transactions B* 370(1660): 1–8. DOI: https://doi.org/10.1098/rsb.2013.0616.

Fabrizi, I., Flament, S., Delhon, C., Gourichon, L., Vuillien, M., Oueslati, T., Auguste, P., Rolando, C. and Bray, F. 2024. Low-Invasive Sampling Method with Tape-Disc Sampling for the Taxonomic Identification of Archeological and Paleontological Bones by Proteomics. *Journal of Proteome Research* acs.jproteome.4c00083. DOI: https://doi.org/10.1021/acs.jproteome.4c00083.

Feng, M., Zhang, L., Lin, X., Gilani., SZ and Mian, A. 2020. Point attention network for semantic segmentation of 3D point clouds. *Pattern Recognition* 107: 107446.

Fernandez, H. 2001. *Ostéologie comparée des petits ruminants eurasiatiques sauvages et domestiques (genres Rupicapra, Ovis, Capra et Capreolus): diagnose différentielle du squelette appendiculaire*. Thèse de Doctorat. Muséum d'histoire naturelle Genève, Université de Genève.

Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N.T.H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D.J., Tavenard, R., Tong, A. and Vayer, T. 2021. POT: Python Optimal Transport. *Journal of Machine Learning Research* 22(78): 1–8.

Gaastra, JS. 2023. Corrigendum to "Domesticating details: 3D geometric morphometrics for the zooarchaeological discrimination of wild, domestic and proto-domestic sheep (*Ovis aries*) and goat (*Capra hircus*) populations" [J. Archaeol. Sci. 151 (2023) 105723]. *Journal of Archaeological Science* 153: 105768. DOI: https://doi.org/10.1016/j.jas.2023.105768.

Gönen, M. 2012. Bayesian efficient multiple kernel learning. *arXiv preprint arXiv:1206.6465*.

Gönen, M. and Alpaydın, E. 2011. Multiple kernel learning algorithms. *The Journal of Machine Learning Research* 12: 2211–2268.

Gudea, A. and Stan, F. 2012. The discriminative macroscopical identification of the bones of sheep (*Ovis aries*), goat (*Capra hircus*) and roe deer (*Capreollus capreollus*). 2. Elements of the hindlimb skeleton. *Bulletin UASMV-Ve Med* 69: 132–139.

Haruda, AF. 2017. Separating sheep (*Ovis aries* L.) and goats (*Capra hircus* L.) using geometric morphometric methods: an investigation of Astragalus morphology from late and final Bronze age central asian contexts. *International Journal of Osteoarchaeology* 27(4): 551–562.

Haruda, A.F., Varfolomeev, V., Goriachev, A., Yermolayeva, A. and Outram, A.K. 2019. A new zooarchaeological application for geometric morphometric methods: Distinguishing *Ovis aries* morphotypes to address connectivity and mobility of prehistoric Central Asian pastoralists. *Journal of Archaeological Science* 107: 50–57. DOI: https://doi.org/10.1016/j.jas.2019.05.002.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hoffman, M.D., Blei, D.M., Wang, C. and Paisley, J. 2013. Stochastic variational inference. *Journal of Machine Learning Research*.

Janzen, A., Richter, K.K., Mwebi, O., Brown, S., Onduso, V., Gatwiri, F., Ndiema, E., Katongo, M., Goldstein, S.T., Douka, K. and Boivin, N. 2021. Distinguishing African bovids using Zooarchaeology by Mass Spectrometry (ZooMS): New peptide markers and insights into Iron Age economies in Zambia Adams, JW (ed.). *PLOS ONE* 16(5): e0251061. DOI: https://doi.org/10.1371/journal.pone.0251061.

Jeanjean, M., Haruda, A., Salvagno, L.., Schafberg, R, Valenzuela-Lamas, S., Nieto-Espinet, A,. Forest, V., Blaise, E., Vuillien, M., Mureau, C. and Evin, A. 2022. Sorting the flock: Quantitative identification of sheep and goat from isolated third lower molars and mandibles through geometric morphometrics. *Journal of Archaeological Science* 141: 105580. DOI: https://doi.org/10.1016/j.jas.2022.105580.

Jeanjean, M., McGrath, K., Valenzuela-Lamas, S., Nieto-Espinet, A., Schafberg, R., Parés-Casanova, P.M., Jiménez-Manchón, S., Guintard, C., Tekkouk, F., Ridouh, R., Mureau, C. and Evin, A. 2023. ZooMS confirms geometric morphometrics species identification of ancient sheep and goat. *Royal Society Open Science* 10(9): 230672. DOI: https://doi.org/10.1098/rsos.230672.

Kazhdan, M., Bolitho, M. and Hoppe, H. 2006. Poisson surface reconstruction. In: *Proceedings of the fourth Eurographics symposium on Geometry processing*. 2006. p.

Lacombe, T., Cuturi, M. and Oudot, S. 2018. Large scale computation of means and clusters for persistence diagrams using optimal transport. *Advances in Neural Information Processing Systems* 31.

Larsson, M.N., Morell Miranda, P., Pan, L., Başak Vural, K., Kaptan, D., Rodrigues Soares, A.E., Kivikero, H., Kantanen, J., Somel, M., Özer, F., Johansson A. M., Stora J., Günther T. 2024. Ancient sheep genomes reveal four Millennia of North European short-tailed sheep in the Baltic Sea region. *Genome Biology and Evolution* evae114. DOI: 10.1093/gbe/evae114

Lavocat, R. 1966. *Atlas de Prehistoire. Tome III, Faunes Et Flores Prehitoriques de L'Europe Occidentale*. Boubee et Cie.

Le Meillour, L., Zazzo, A., Zirah, S., Tombret, O., Barriel, V., Arthur, K.W., Arthur, J.W., Cauliez, J., Chaix, L., Curtis, M.C., Gifford-Gonzalez, D., Gunn, I., Gutherz, X., Hildebrand, E., Khalidi, L., Millet, M., Mitchell, P., Studer, J., Vila, E., Welker, F., Pleurdeau, D. and Lesur, J. 2023. The name of the game: palaeoproteomics and radiocarbon dates further refine the presence and dispersal of caprines in eastern and southern Africa. *Royal Society Open Science* 10(11): 231002. DOI: https://doi.org/10.1098/rsos.231002.

Le Meillour, L., Zirah, S., Zazzo, A., Cersoy, S., Détroit, F., Imalwa, E., Lebon, M., Nankela, A., Tombret, O., Pleurdeau, D. and Lesur, J. 2020. Palaeoproteomics gives new insight into early southern African pastoralism. *Scientific Reports* 10(1): 14427. DOI: https://doi.org/10.1038/s41598-020-71374-3.

Lloveras, L., Rissech, C., Davis, S. and Parés-Casanova, P.M. 2022. Morphological Differences between Sheep and Goat Calcanea Using Two-Dimensional Geometric Morphometrics. *Animals* 12(21): 2945.

Lundberg, S.M. and Lee, S-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Lv, F.-H., Cao, Y.-H., Liu, G.-J., Luo, L.-Y., Lu, R., Liu, M.-J., Li, W.-R., Zhou, P., Wang, X.-H., Shen, M., and others. 2022. Whole-genome resequencing of worldwide wild and domestic sheep elucidates genetic diversity, introgression, and agronomically important loci. *Molecular biology and evolution* 39(2): msab353. DOI: 10.1093/molbev/msab353

Miele, V., Dussert, G., Cucchi, T. and Renaud, S. 2020. Deep learning for species identification of modern and fossil rodent molars. *BioRxiv* 2020–08.

Moclán., A, Domínguez-García, Á.C., Stoetzel, E., Cucchi, T., Sevilla, P. and Laplana, C. 2023. Machine Learning interspecific identification of mouse first lower molars (genus Mus Linnaeus, 1758) and application to fossil remains from the Estrecho Cave (Spain). *Quaternary Science Reviews* 299: 107877.

Moclán, A., Domínguez-Rodrigo, M. and Yravedra, J. 2019. Classifying agency in bone breakage: an experimental analysis of fracture planes to differentiate between hominin and carnivore dynamic and static loading using machine learning (ML) algorithms. *Archaeological and Anthropological Sciences* 11: 4663–4680.

Mohaseb, A.F., Cornette, R., Zimmermann, M.I., Davoudi, H., Berthon, R., Guintard, C., Cucchi, T., Hanot, P., Mohandesan, E., Eisenmann, V., Peters J. Mashkour M. 2023. Predictive use of modern reference osteological collections for disentangling the shape of Eurasian equid cheek teeth and metapodials in archaeological material. *International Journal of Osteoarchaeology* 33(5): 938–954. DOI: 10.1002/oa.3255

Munro, N.D., Bar-Oz, G. and Hill, A.C. 2011. An exploration of character traits and linear measurements for sexing mountain gazelle (*Gazella gazella*) skeletons. *Journal of Archaeological Science* 38(6): 1253–1265.

Myronenko, A. and Song, X. 2010. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence* 32(12): 2262–2275.

Nicolau, M., Levine, A.J. and Carlsson, G. 2011. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences* 108(17): 7265–7270.

Otter, N., Porter, M.A., Tillmann, U., Grindrod, P. and Harrington, H.A. 2017. A roadmap for the computation of persistent homology. *EPJ Data Science* 6: 1–38.

Peters, J. 1989. Osteomorphological features of the appendicular skeleton of gazelles, genus *Gazella* Blainville 1816, bohor reedbuck, redunca redunca (Pallas, 1767) and bushbuck, *Tragelaphus scriptus* (Pallas, 1766). *Anatomia, Histologia, Embryologia* 18(2): 97–113.

Pilaar Birch, S.E., Scheu, A., Buckley, M. and Çakırlar, C. 2019. Combined osteomorphological, isotopic, aDNA, and ZooMS analyses of sheep and goat remains from Neolithic Ulucak, Turkey. *Archaeological and Anthropological Sciences* 11(5): 1669–1681. DOI: https://doi.org/10.1007/s12520-018-0624-8.

Plummer, T.W., Bishop, L.C. and Hertel, F. 2008. Habitat preference of extant African bovids based on astragalus morphology: operationalizing ecomorphology for palaeoenvironmental reconstruction. *Journal of Archaeological Science* 35(11): 3016–3027. DOI: https://doi.org/10.1016/j.jas.2008.06.015.

Pöllath, N., Alibert, P., Schafberg, R. and Peters, J. 2018. Striking new paths–Distinguishing ancient Ovis orientalis from its modern domestic descendant (Karakul breed) applying Geometric and traditional Morphometric approaches to the astragalus. In: *Archaeozoology of the Near East XII. Proceedings of the 12th International Symposium of the ICAZ Archaeozoology of Southwest Asia and Adjacent Areas Working Group, Groningen Institute of Archaeology, June 14-15 2015, University of Groningen, the Netherlands*. 2018. pp. 207–225.

Pöllath, N., Schafberg, R. and Peters, J. 2019. Astragalar morphology: Approaching the cultural trajectories of wild and domestic sheep applying Geometric Morphometrics. *Journal of Archaeological Science: Reports* 23: 810–821. DOI: https://doi.org/10.1016/j.jasrep.2018.12.004.

Popkin, P.R.W., Baker, P., Worley, F., Payne, S. and Hammon, A. 2012. The Sheep Project (1): determining skeletal growth, timing of epiphyseal fusion and morphometric variation in unimproved Shetland sheep of known age, sex, castration status and nutrition. *Journal of Archaeological Science* 39(6): 1775–1792. DOI: https://doi.org/10.1016/j.jas.2012.01.018.

Prendergast, M.E., Janzen, A., Buckley, M. and Grillo, K.M. 2019. Sorting the sheep from the goats in the Pastoral Neolithic: morphological and biomolecular approaches at Luxmanda, Tanzania. *Archaeological and Anthropological Sciences* 11(6): 3047–3062. DOI: https://doi.org/10.1007/s12520-018-0737-0.

Project, TG. 2023. *GUDHI User and Reference Manual*. 3.9.0. GUDHI Editorial Board.

Prummel, W. and Frisch, H-J. 1986. A guide for the distinction of species, sex and body side in bones of sheep and goat. *Journal of Archaeological Science* 13(6): 567–577. DOI: https://doi.org/10.1016/0305-4403(86)90041-5.

Qi, C.R., Yi, L., Su, H. and Guibas, L.J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30.

Rakotomamonjy, A., Bach, F., Canu, S. and Grandvalet, Y. 2008. SimpleMKL. *Journal of Machine Learning Research* 9: 2491–2521.

Ribeiro, M.T., Singh, S. and Guestrin, C. 2016. 'Why should i trust you?' Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016. pp. 1135–1144.

Rouvreau, V. 2023. Alpha complex. In: *GUDHI User and Reference Manual*. 3.9.0. GUDHI Editorial Board. p.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1(5): 206–215.

Salvagno, L. and Albarella, U. 2017. A morphometric system to distinguish sheep and goat postcranial bones Mihlbachler, MC (ed.). *PLOS ONE* 12(6): e0178543. DOI: https://doi.org/10.1371/journal.pone.0178543.

Sipilä, I.M.V., Steele, J., Dickens, L. and Martin, L. 2023. Bones of contention: a double-blind study of experts' ability to classify sheep and goat astragali from images. *Archaeological and Anthropological Sciences* 15(12): 187. DOI: https://doi.org/10.1007/s12520-023-01865-7.

Sobolev, A. and Vetrov, D.P. 2019. Importance weighted hierarchical variational inference. *Advances in Neural Information Processing Systems* 32

Tsahar, E., Izhaki, I., Lev-Yadun, S. and Bar-Oz, G. 2009. Distribution and Extinction of Ungulates during the Holocene of the Southern Levant Hansen, DM (ed.). *PLoS ONE* 4(4): e5316. DOI: https://doi.org/10.1371/journal.pone.0005316.

Uerpmann, H-P. 1987. *The ancient distribution of ungulate mammals in the Middle East*. Beihefte zum Tübinger Atlas des Vorderen Orients Reihe A. Naturwissenschaften. Reichert, Wiesbaden.

Vila, E., Abrahami, P., Albesso, M.., Amane, A, Bader, C., Berthon, R., Bouzid, S., Bradley, D., Breniquet, C., Chahoud, J., Cucchi, T., Davoudi, H., De Cupere, B., Escarguel, G., Estrada, O., Gourichon, L., Helmer, D., Huangfu, W., Lesur, J., Mashkour, M., Michel, C., Mohaseb, A., Orlando, L., Pompanon, F., Studer, J. and Vuillien, M. 2021. EVOSHEEP: the makeup of sheep breeds in the ancient Near East. *Antiquity* 95(379): e2. DOI: https://doi.org/10.15184/aqy.2020.247.

Vuillien, M. 2020. *Systèmes d'élevage et pastoralisme en Provence et dans les Alpes méridionales durant la Protohistoire : Nouvelles perspectives en archéozoologie*. Thèse de doctorat de Préhistoire. CEPAM, UMR 7264, Université Côte d'Azur, 687 p.

Vuillien, M. 2024. *Archéozoologie et Machine Learning : vers une collaboration d'avenir ?* 2024, Bioarcheologies, Le Blog, Available at https://doi.org/10.58079/VZQ7.

Wadsworth, C., Procopio, N., Anderung, C., Carretero, J.-M., Iriarte, E., Valdiosera, C., Elburg, R., Penkman, K. and Buckley, M. 2017. Comparing ancient DNA survival and proteome content in 69 archaeological cattle tooth and bone samples from multiple European sites. *Journal of Proteomics* 158: 1–8. DOI: https://doi.org/10.1016/j.jprot.2017.01.004.

Wyatt-Spratt, S. 2022. After the revolution: a review of 3D modelling as a tool for stone artefact analysis. *Journal of Computer Applications in Archaeology* 5(1).

Zeder, M.A. and Lapham, H.A. 2010. Assessing the reliability of criteria used to identify postcranial bones in sheep, *Ovis*, and goats, *Capra*. *Journal of Archaeological Science* 37(11): 2887–2905. DOI: https://doi.org/10.1016/j.jas.2010.06.032.

Zeder, M.A. and Pilaar, S.E. 2009. Assessing the reliability of criteria used to identify mandibles and mandibular teeth in sheep, *Ovis*, and goats, *Capra*. *Journal of Archaeological Science* 37(2): 225–242. DOI: https://doi.org/10.1016/j.jas.2009.10.002.

Zhao, T., Alliez, P., Boubekeur, T., Busé, L. and Thiery J.-M. 2021. Progressive discrete domains for implicit surface reconstruction. In: *Computer Graphics Forum*. Wiley Online Library. pp. 143–156. DOI: 10.1111/cgf.14363

Zomorodian, A. and Carlsson, G. 2004. Computing persistent homology. In: *Proceedings of the twentieth annual symposium on Computational geometry*. 2004. pp. 347–356.

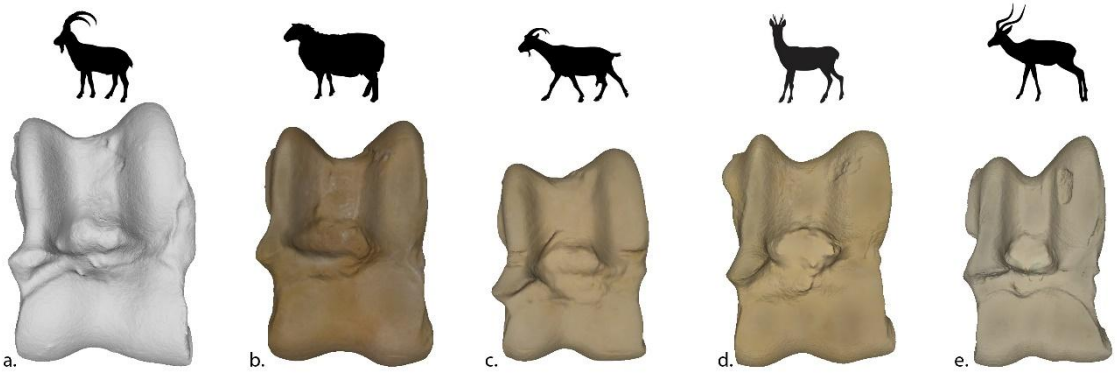

Figure 1: 3D astragalus presented in dorsal view of (a) Alpine ibex (*Capra ibex*), (b) sheep (*Ovis aries*), (c) goat (*Capra hircus*), (d) roe deer (*Capreolus capreolus*) and (e) gazelle (*Gazella* sp.). Alpine ibex astragalus is untextured.

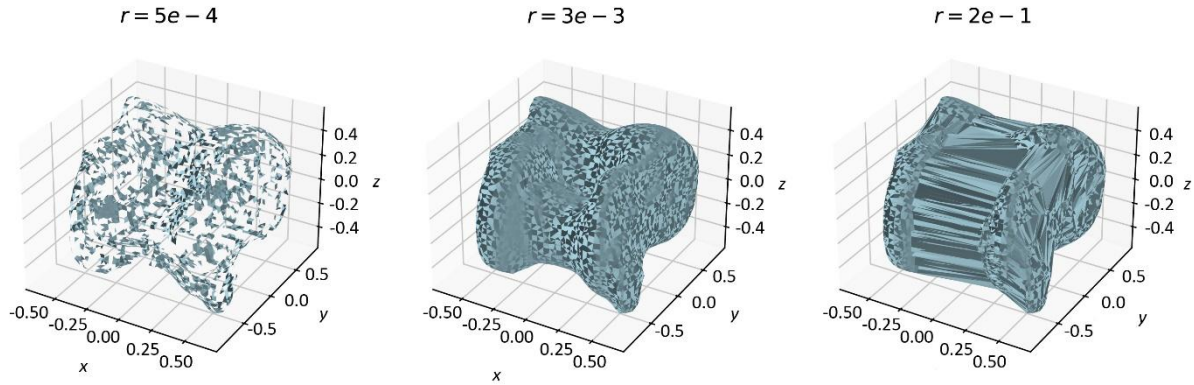

Figure 2: Evolution of the simplicial complex of a 3D astragalus (reference name Obs_1997_187) during the Alpha filtration process, highlighting the impact of varying radius thresholds r. (Left) At this stage,

the bone's structure is partially reconstructed, with multiple connected components and some topological cycles, 1-dimensional features, visible. (Center) The Alpha complex has merged into a single connected component, capturing the overall structure of the bone. The cycles from the previous step have been filled (died), and the complex closely approximates the 3D shape of the bone. (Right) At this advanced stage, most topological cycles have disappeared, and the complex over-reconstructs the bone's shape. The filtration process concludes once all topological voids, 2-dimensional features, are filled.

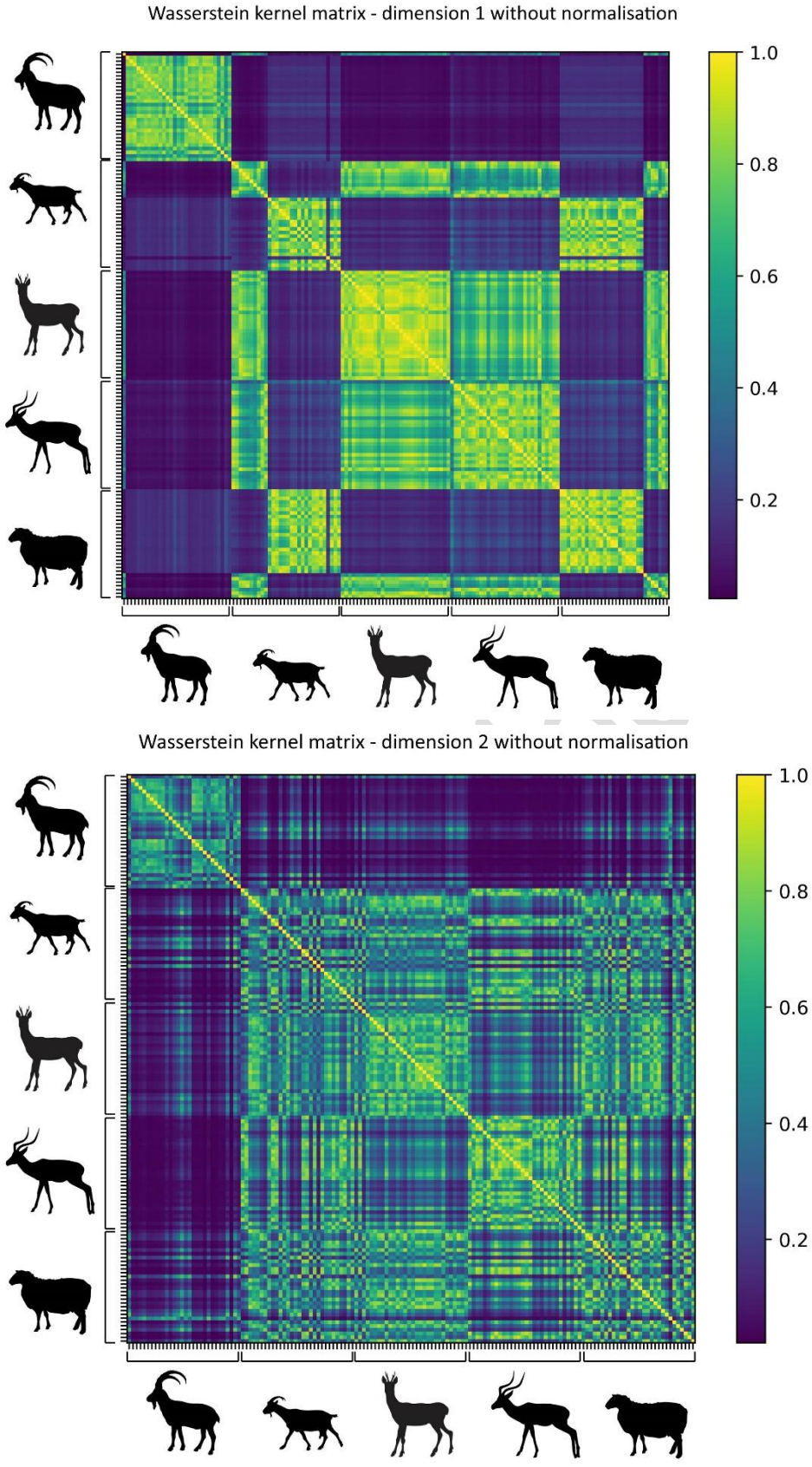

Figure 3: Wasserstein kernel matrices without bone's normalisation. Up: topological dimension 1; Down: topological dimension 2. For both the matrices the color code, indicated by the colorbar on the right of the matrix, represents pairwise similarity within the range [0, 1]. Yellow cells (similarity equals

to 1), such as those along the diagonal, signify that the *x* and *y* bones are identical. As the color shifts towards blue, the bones exhibit increasing dissimilarity (similarity approaching 0).

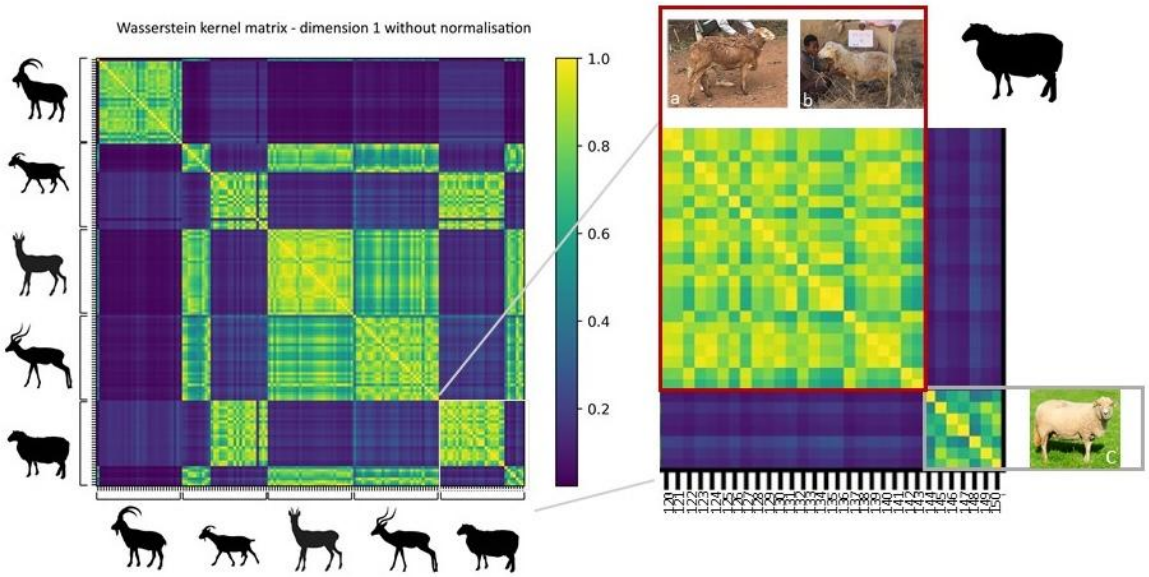

Figure 4: Illustration of topological dissimilarities observed in sheep in the Wasserstein kernel matrices without bone's normalisation (dimension 1). Picture of sheep breed "Bonga" (a) and "Menz" (b) from Ethiopia © A. Amane / E. Vila. c) Picture of sheep breed "Landes de Bretagne" from France ©H. Ronné https://www.ecomusee-rennes-metropole.fr/le-mouton-des-landes/

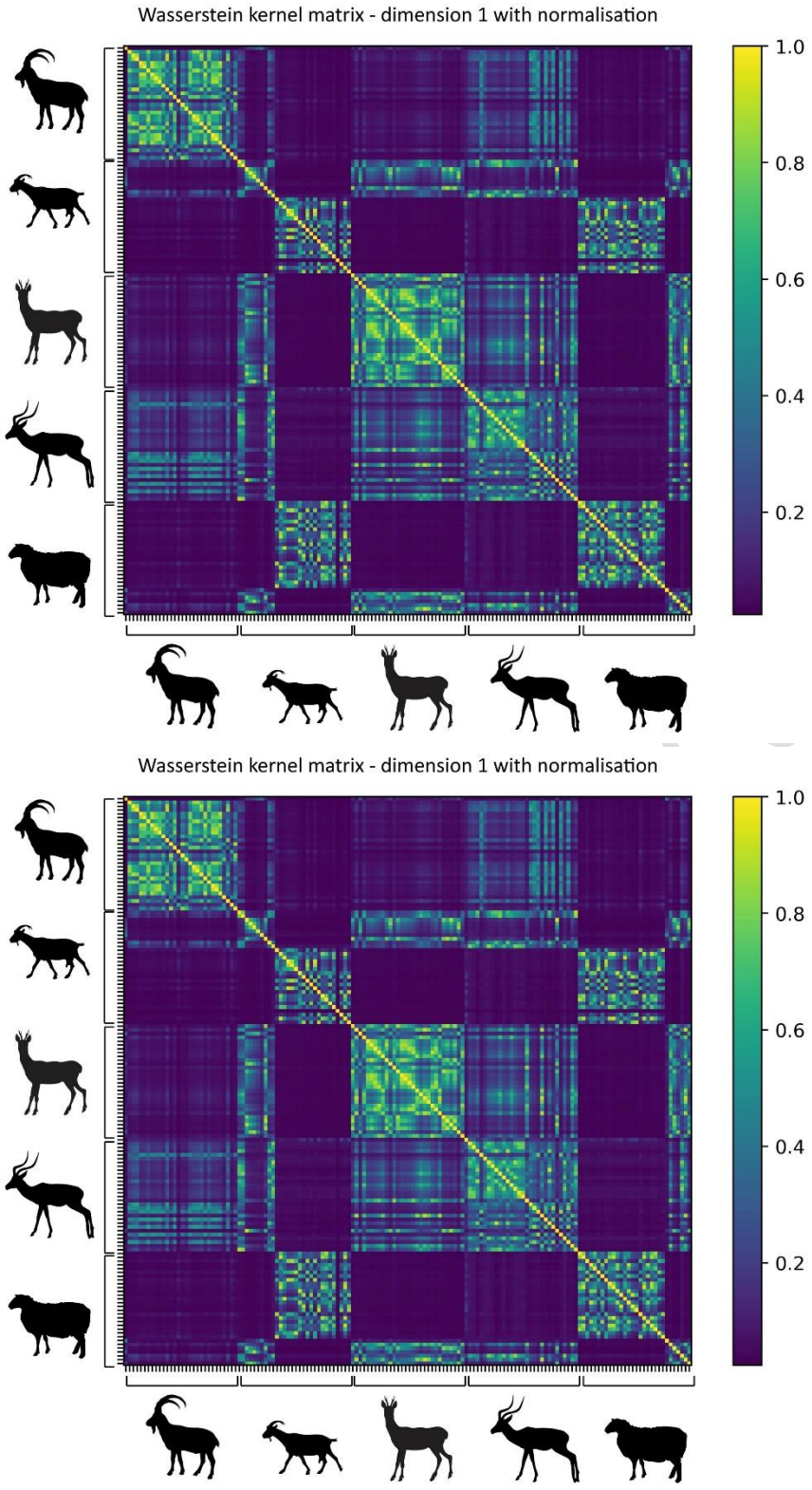

Figure 5: Wasserstein kernel matrices with bone's normalisation. Up: topological dimension 1; Down: topological dimension 2. For both the matrices the color code, indicated by the colorbar on the right of the matrix, represents pairwise similarity within the range [0, 1]. Yellow cells (similarity equals to 1), such as those along the diagonal, signify that the *x* and *y* bones are identical. As the color shifts towards blue, the bones exhibit increasing dissimilarity (similarity approaching 0).

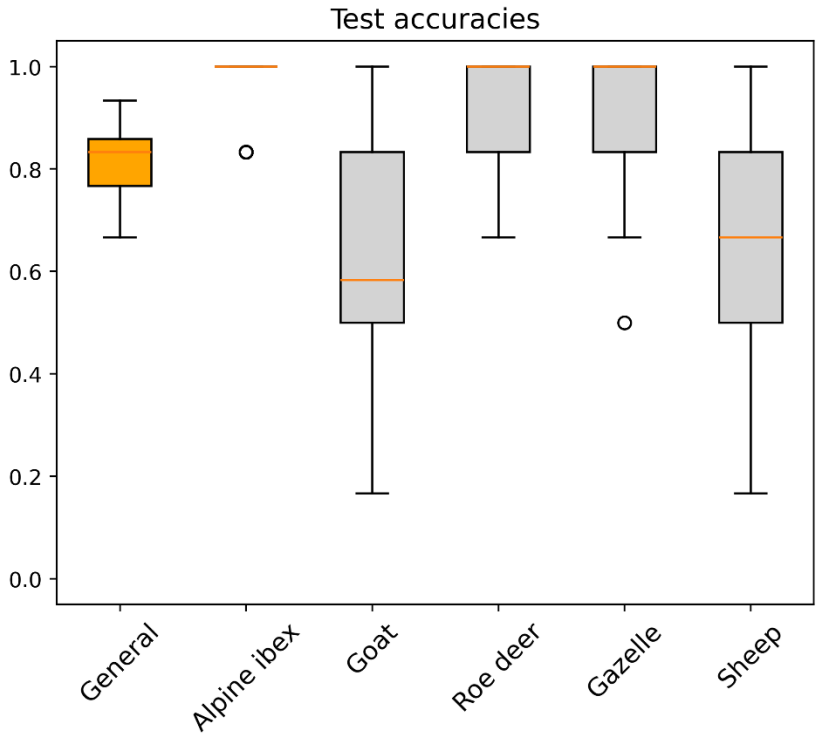

Figure 6a : Boxplot average test classification accuracies. The first column (leftmost) represents the average performance over the five classes, while the others show the average accuracy for each specie. The central line (orange) within each box represents the median accuracy, while the lower and upper edges of the box correspond to the first (Q1) and third quartiles (Q3), respectively, indicating the interquartile range (IQR). The whiskers extend to the minimum and maximum values within (1.5*IQR) from Q1 and Q3. Beyond this range are considered outliers and are shown as individual markers.

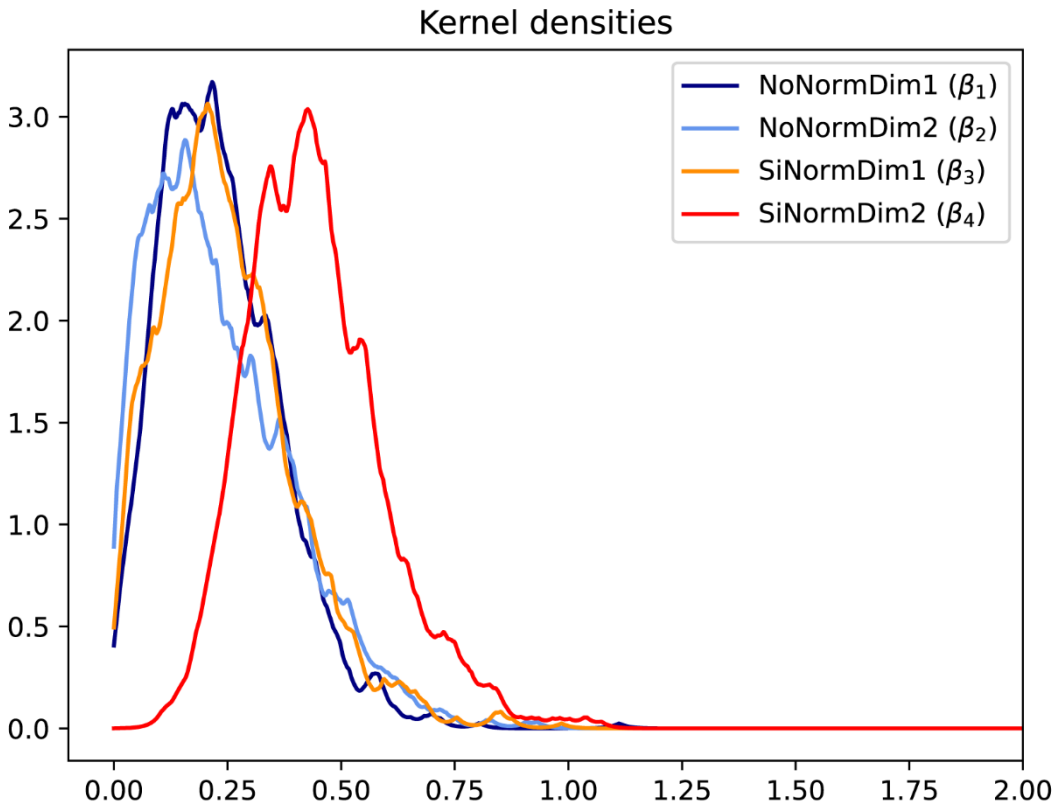

Figure 6b : Kernel densities estimates. Different colors correspond to the weights of the four Wasserstein kernel matrices obtained via TDA. X-axis corresponds to weights values, while the y-axis indicates the estimated density, reflecting the relative frequency of occurrence. Peaks in the density curves show the most probable values of the weights, while the spread of each distribution provides insight into the variability and uncertainty of the estimates.

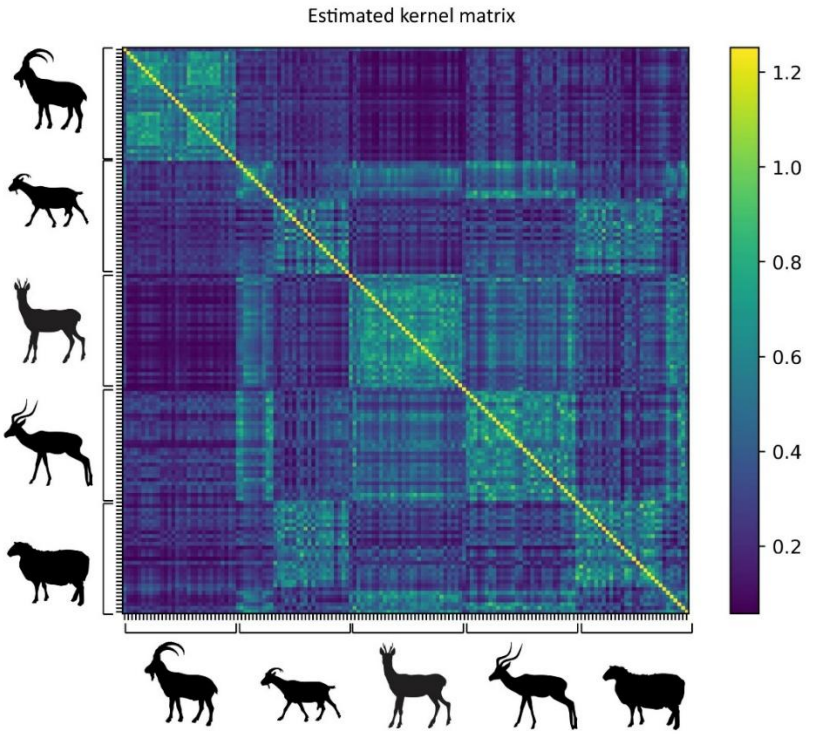

Figure 7: Learned kernel matrix for a single data split. It represents the optimal linear combination, where each input kernel matrix is weighted by its corresponding estimated weight value. The color code follows the same interpretation as in Figures 3 and 5.