Engels, I, Burnett, A, Robert, P, Pironneau, C, Abrams, G, Bouwmeester, R, Van der Plaetsen, P, Di Modica, K, Otte, M, Straus, LG, Fischer, V, Bray, F, Mesuere, B, De Groote, I, Deforce, D, Daled, S and Dhaenens, M

Classification of Collagens via Peptide Ambiguation, in a Paleoproteomic LC-MS/MS-Based Taxonomic Pipeline

https://researchonline.ljmu.ac.uk/id/eprint/26694/

Article

Journal of proteome research

Open Access

pubs.acs.org/jpr

Article

Classification of Collagens via Peptide Ambiguation, in a Paleoproteomic LC-MS/MS-Based Taxonomic Pipeline

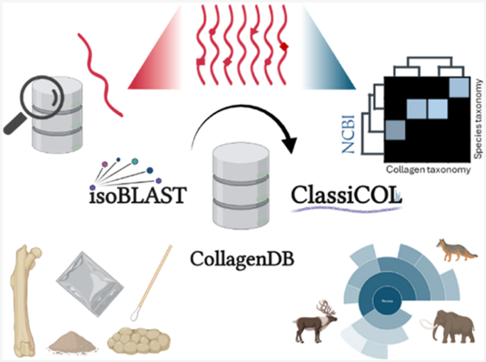
Ian Engels,[§§] Alexandra Burnett,[§§] Prudence Robert, Camille Pironneau, Grégory Abrams, Robbin Bouwmeester, Peter Van der Plaetsen, Kévin Di Modica, Marcel Otte, Lawrence Guy Straus, Valentin Fischer, Fabrice Bray, Bart Mesuere, Isabelle De Groote, Dieter Deforce, Simon Daled,[§§] and Maarten Dhaenens*,[§§]

Cite This: J. Proteome Res. 2025, 24, 1907−1925

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Liquid chromatography−mass spectrometry (LC-MS/MS) extends the matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) Zooarcheology by Mass Spectrometry (ZooMS) "mass fingerprinting" approach to species identification by providing fragmentation spectra for each peptide. However, ancient bone samples generate sparse data containing only a few collagen proteins, rendering target−decoy strategies unusable and increasing uncertainty in peptide annotation. To ameliorate this issue, we present a ZooMS/MS data pipeline that builds on a manually curated Collagen database and comprises two novel algorithms: isoBLAST and ClassiCOL. isoBLAST first extends peptide ambiguity by generating all "potential peptide candidates" isobaric to the annotated precursor. The exhaustive set of candidates created is then used to retain or reject different potential paths at each taxonomic branching point from superkingdom to species, until the greatest possible specificity is reached. Uniquely, ClassiCOL allows for the identification of taxonomic mixtures, including contaminated samples, as well as suggesting taxonomies not represented in sequence databases, including extinct taxa. All considered ambiguity is then graphically represented with clear prioritization of the potential taxa in the sample. Using public as well as in-house data acquired on different instruments, we demonstrate the performance of this universal postprocessing and explore the identification of both genetic and sample mixtures. Diet reconstruction from 40,000-year-old cave hyena coprolites illustrates the exciting potential of this approach.



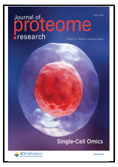
KEYWORDS: *paleoproteomics, proteomics, archeology, bioinformatics, ZooMS, ZooMS/MS, mass spectrometry, Belgium, isoBLAST, ClassiCOL*

■ INTRODUCTION

Bone morphology-based species classification has been the state of the art for both paleontological and zooarcheological research for decades. However, when specimens are degraded, fragmented, and/or fractured to a point where this methodology can no longer be used, paleoproteomics has proven to be an excellent candidate to tackle this challenge.[1] Compared to DNA, proteins typically survive longer, particularly in biomineralized matrices like bone, enamel or eggshells.[2−7] Together with the low sample amount required, straightforward and increasingly automatable sample preparation and relatively low cost, this makes paleoproteomics a very appealing approach.[8−11]

When analyzing ancient bone samples, often only collagen proteins remain detectable, especially COL1A1/COL1A2, which are the most abundant and stable proteins, estimated to constitute over 90% of protein in bone, *in vivo*.[5,12] This abundance inspired the first successful ancient protein methodologies relying on MALDI-TOF MS, to create species-specific Peptide Mass Fingerprints (PMF).[13−17] This method is generally referred to as Zooarcheology by Mass Spectrometry (ZooMS), and can be performed at very high sample throughputs. More recently, methods have been adopted that use Liquid Chromatography tandem Mass Spectrometry (LC-MS/MS) for species determination because this provides intrinsically more information-rich data. We refer to such approaches as ZooMS/MS. For example, SPIN (Species by Proteome INvestigation) relies on curated Peptide-Spectrum Matches (PSMs) to accurately classify species, using a custom database containing the most commonly found proteins in archeological mammalian bones.[8]

Received: October 30, 2024
Revised: February 7, 2025
Accepted: March 5, 2025
Published: March 13, 2025



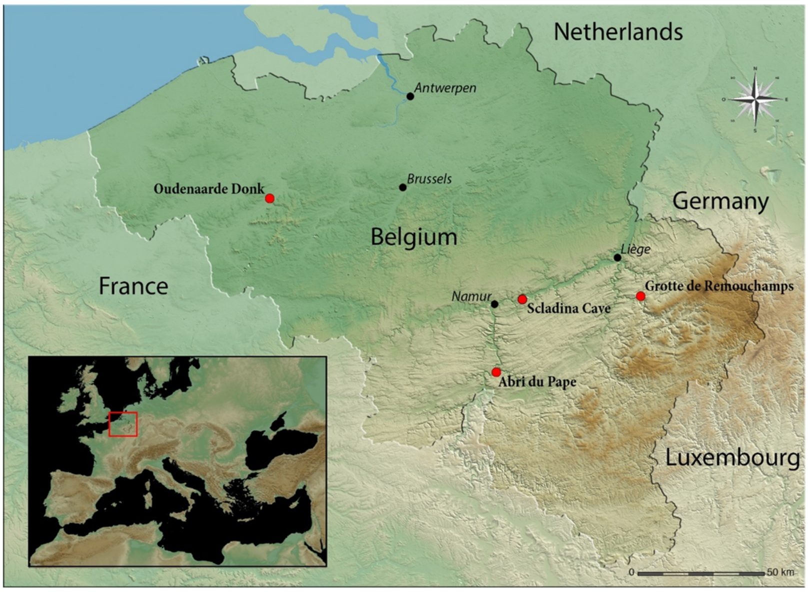


Figure 1. Geographical location of the Belgian archeological sites. Oudernaarde-Donk (OD), Scladina cave, Grotte de Remouchamps (RMC) and Abri du Pape (ADP) are represented by red dots. Map generated with global mapper.

Essentially, an LC-MS/MS system extends on the MALDI-TOF approach by measuring not only the intensity and m/z of the intact peptide (precursor) masses but also providing retention time (t_R) and a fragmentation pattern for each peptide.[18,19] In overview, the spectra that are generated from peptides are translated into a simplified, i.e., peak-picked, list of ion coordinates including at least the precursor mass and fragment ions with their respective intensities.[20] Peptide sequences are then obtained via scoring algorithms that perform a search against a translated DNA sequence database. Generally, these report the best peptide hit for each spectrum in the data.[21] To threshold true identifications from false ones, a probabilistic approach was developed by Matrix Science (Mascot),[22] and later also integrated in MaxQuant.[23,24]

As MS data became increasingly complex, a strategy to empirically assess the False Discovery Rate (FDR) at a given score threshold (rather than computing it probabilistically) was established through the target−decoy approach,[25,26] as an alternative to algorithms like PeptideProphet[27] that fit the discriminant score distributions of true and false hits based on fixed weights of the different metrics or scores. To do so, the sequences in the database are reversed or scrambled into nonsense sequences and added to the database. When searching against this database, the score can be cut off at a threshold below which a given portion of decoy sequences (and thus false discoveries) have been identified, In a later stage, these score distributions of "true" and especially "known false" annotations became a means to improve the weights that were given to specific features of the scoring algorithm. The first algorithm using this was Percolator, the first widely adopted, support vector-based machine learning strategy in proteomics, still being used to date.[28,29] Further increases in data richness obtained through MS sparked the latest class of search engines with the introduction of more elaborate machine learning approaches, which rely on neural networks and accurate predictions to extract ever-larger numbers of peptide identifications from the data.[30,31] Still, the target−decoy approach is pivotal for all these algorithms, since they need to be trained on what is right and what is wrong in order to improve the feature weights used to score the peptide-to-spectrum matches.

This presents paleoproteomics with a conundrum: ancient bone samples generate sparse data sets containing only a few hundred peptides derived from only a few proteins, effectively disabling the use and efficiency of target−decoy strategies for FDR estimation and potential score improvement. In other words, where the simplicity of the protein composition enables fast and effective PMF using MALDI, it actually complicates LC-MS/MS data interpretation. Fortunately, probabilistic search engines like Mascot and MaxQuant are still suitable for sparse peptide identification and are therefore the preferred choice for most proteomic challenges involving data sparsity. Still, the data consists of peptides that can be derived from the orthologous collagen sequences of a plethora of organisms, producing lists of very similar and often indistinguishable peptide candidates for each spectrum.[32,33] This ambiguity is not easily resolved because it is caused by positional isomers and isobaric changes that do not affect the score, such as posttranslational modification (PTM) combined with amino acid substitutions. This is further aggravated by the taphonomic processes that lead to low quality spectra.[34,35] Ultimately, this ambiguity impairs accurate postprocessing tools like Unipept,[36] which infers unique peptides to species at their last common ancestor (LCA), and leads to incorrect classifications by an otherwise highly performant tool.

As the end goal of ZooMS(/MS) is to identify the species of origin rather than the proteins inside the sample, we here embrace and even extend the ambiguity in MS/MS search outputs to obtain an exhaustive set of Potential Peptide Candidates (PPCs). This increases the chance of obtaining the correct peptide sequence for each spectrum in an otherwise redundant list of PPCs. This new axiom is then leveraged by rejecting the respective peptide sequences during species classification, i.e., postannotation, as opposed to trying to improve the annotation itself. In turn, this means that the approach is compatible with any prior search engine. Therefore, an anti-Occam's razor[37] based algorithm is used to follow the NCBI taxonomic tree and, at every branching point, discard the branches that are mere subsets of the other and do not contribute unique peptides. We demonstrate the efficiency of this approach on several public data sets and in-house-generated data, showing that this universal postprocessing tool enables the identification of different species from single bone fragments as well as mixtures derived from their remains (e.g., from dust in bags or coprolites). Moreover, we show that this methodology holds equally high potential for identifying protein-containing paint and glue binders in cultural heritage samples.

■ METHODS

Archeological Contexts of Oudenaarde-Donk (OD), Abri du Pape (ADP) and Remouchamps (RMC)

A single *Rangifer tarandus* (reindeer) element from Grotte de Remouchamps was sampled. Located in Southern Belgium within the province of Liège near the Amblève River (Figure 1), the cave has been known since the 18th century and attracted the attention of various prospectors. Rahir[38] and later Dewez[39] conducted the latest and most complete research on the archeological material uncovered during their excavations. The material, lithic, and faunal remains are associated with Ahrensburgian occupation(s) dated from ca. 12,700/12,500 to ca. 11,400/11,200 cal BP, corresponding to most of the Younger Dryas and the early Preboreal.[40]

The rock shelter of Abri du Pape is located in the province of Namur, in southern Belgium, on the right bank of the Meuse River (Figure 1). Initial excavations at the site were conducted by Ph. Lacroix in 1988, who uncovered archeological deposits in a sondage. This early work was followed by a series of major excavations throughout the 1990s, led by Leotard, Otte, Straus and a multidisciplinary team.[41−44] The site revealed an extensive sedimentary sequence documenting human occupations spanning from the Mesolithic to the present day. The 14 faunal remains sampled for our analyses come from the Mesolithic layers, dated from ca. 9918 to ca. 8459 cal BP, and were initially analyzed by Gautier.[42]

Finally, the 42 faunal samples from Oudenaarde-Donk were excavated from Neo 1, one of the 10 sites identified at this location on the left bank of the Middle Scheldt (Figure 1).[45,46] Several faunal remains, among human remains (not analyzed here), were directly radiocarbon-dated. The ^{14}C dates of the faunal remains range from 6177 to 3178 cal BP, indicating a Middle Neolithic to Early Bronze Age origin.[47,48]

Bone powder was obtained from the exterior surface of the bones during cleaning prior to collagen extraction for stable isotope analysis. Drilling was undertaken with diamond-tipped drill bits which were changed and cleaned between each use. Cleaning was twice performed by 5 min sonication in Milli-Q water followed by 5 min sonication in 70% ethanol. A maximum of 5 mg of powder was separated from the cleaning layers and taken forward for protein extraction.

Sample Information Scladina Cave

Scladina Cave is located on the right bank of the Meuse Valley, between Andenne and Namur (Figure 1). The cave was discovered by amateur archeologists in 1971 and has been under a permanent scientific archeological program since the early 80s.[49] The cave yielded numerous archeological occupations mostly by Neanderthals, as well as some evidence of early anatomically modern humans in northwestern Europe.[50] The stratigraphic sequence is composed of no less than 120 layers for an approximate cumulated thickness of around 15 m, which covers the Holocene up to at least the late Middle Pleistocene.[51,52] Although the site is known for its archeological occupations, the sediments at Scladina Cave have preserved an impressive quantity of bone and dental remains, mainly belonging to cave bears (*Ursus spelaeus*), enabling tracing of their evolution and adaptation in well-controlled stratigraphic contexts.[53,54] Among these cave bear remains, several bone fragments were used as retouchers by Neanderthals, documenting specific interactions between Neanderthals and carnivores.[55,56]

The samples from Scladina Cave were excavated from a variety of stratigraphic layers.[52,57] Tooth dentine from *Mammuthus primigenius* (wooly mammoth; SC1997−150−1) originated from layer Z1, as did the mandible of the *Crocuta crocuta spelaea* (spotted cave hyena); for both specimens, dust from within the container was sampled in addition to the dentine/bone fragments themselves. The *Coelodonta antiquitatis* bone (wooly rhino; SC1995−279−475) excavated from Unit 6 was sampled in a similar manner. Additionally, bone dust samples were collected from *Rupicapra rupicapra* (chamois; SC1984−543−3) and *Dama dama* (European fallow deer; SC1986−1270−224) excavated from Unit 5. Bone dust was also collected from layers V grise - Vb of *Panthera pardus* (leopard; SC1982−284−1). Bone samples from Units 4 included *Felis sylvestris* (wildcat; SC1983−57−33, Unit 4A) and *Alopex lagopus* (arctic fox; SC1983−80−2, Unit IV). Finally, dentine from a *Megaloceros giganteus* (Irish elk; SC2011−210−1, Layer 1B-RA) tooth specimen from layer 1B-RA and bone dust from *Lynx lynx* (Eurasian lynx; SC2002−699−5, Layer 39) were sampled.

The plastic bag containing the coprolite samples was sampled by swabbing the interior of the bag with a sterile swab. The coprolites were excavated from layer T-RO that has yielded evidence for an Aurignacian occupation dated between 40,150−37,500 cal BP.[50]

Protein Extraction

Each sample was demineralized in 600 μL of 0.6 M HCl (Chemlab, CL05.0312.1000) for 24 h at 25 °C, while shaking at 750 rpm (Eppendorf Thermomixer comfort). The samples were pelletized via centrifugation after which the supernatant fraction was removed and stored at −20 °C as a back-up. The pellet was washed with ice-cold acetone (Sigma-Aldrich, 179124−1L) and resuspended in extraction buffer (5% SDS (Invitrogen, 15553−027) + 50 mM TEAB (Sigma-Aldrich, 90360−100 ML)). DTT (Chemlab, CL00.0481.0025) was added to a final concentration of 20.8 mM to reduce the disulfide bridges for 30 min at 37 °C in the dark. Next, MMTS (Sigma-Aldrich, 64306−10 ML) was added to a final concentration of 20 mM to alkylate the sulfide groups for 10 min at room temperature in the dark. The denatured proteins were precipitated with phosphoric acid (Chemlab, CL00.0605.1000) at pH 1.

Proteins were trapped on HiPure Viral Mini columns (Magen Biotechnology, China; C13112) after addition of 165 μL binding/washing buffer (100 mM TEAB in 90% Methanol (Chemlab, CL00.1377.1000)). The columns were centrifuged for 30 s (4000 rpm, 25 °C, Eppendorf centrifuge 5417R) between each of the three washing steps to elute the binding/washing buffer, and the first elution was reloaded on the column to decrease potential protein loss. The proteins were digested on-filter with 1 μg trypsin/Lys-C (ProMega, V5073) in 40 μL of 50 mM TEAB overnight at 37 °C. The peptides were eluted from the column with 30 μL of 50 mM TEAB, followed by 30 μL of 0.1% formic acid (FA) (Biosolve, 2324) and finally 30 μL of

50% acetonitrile (Chemlab, CL00.0194.1000). All three elution steps were performed with a 1 min incubation and centrifugation step. The samples were vacuum-dried and resuspended in 0.1% FA for LC-MS/MS analysis.

LC-MS/MS Data Acquisition via ZenoTOF MS

The samples were analyzed using a Waters Acquity MClass UPLC system coupled with a Sciex ZenoTOF 7600 mass spectrometer in data-dependent acquisition mode. The peptide samples were trapped on a YMC triart C18 guard column, 3 μm, 5 × 0.3 mm and separated on a YMC Triart C18, 3 μm, 150 × 0.3 mm analytical column using an optimized nonlinear 20 min gradient of 1.5 to 36% solvent B (0.1% FA in acetonitrile) in solvent A (0.1% FA in water). Precursor scans (TOF-MS) were acquired for 0.1s over a mass range of 300−1600 m/z. Up to 40 precursors with an intensity threshold of 150 counts per second, a dynamic exclusion of 6 s after 2 occurrences and a charge state between 2 and 5 were fragmented per cycle using collision induced dissociation. The fragment spectra were acquired for 0.015s over a mass range of 100−2000 m/z, resulting in a cycle time of 0.920s.

Manual Curation of Collagen Databases

The collagen database (CollagenDB) which sits at the heart of the ClassiCOL pipeline contains a manually curated list of protein sequences collected from the UniProt and the NCBI protein repositories. Here, for each species (n = 614), all 45 types of collagen were extracted if they were present in one of the repositories and downloaded in FASTA format (Data S1). Protein sequences were collected from the NCBI data repository between the 30th of October 2023 and the 28th of July 2024, by blasting protein sequences from closely related species via BLASTp (NCBI). All FASTA files were combined into a single file (sequences n = 22,017), which was submitted on the MASCOT server. This FASTA file is a living document to which proteins are added continuously when new species are submitted into the aforementioned repositories.

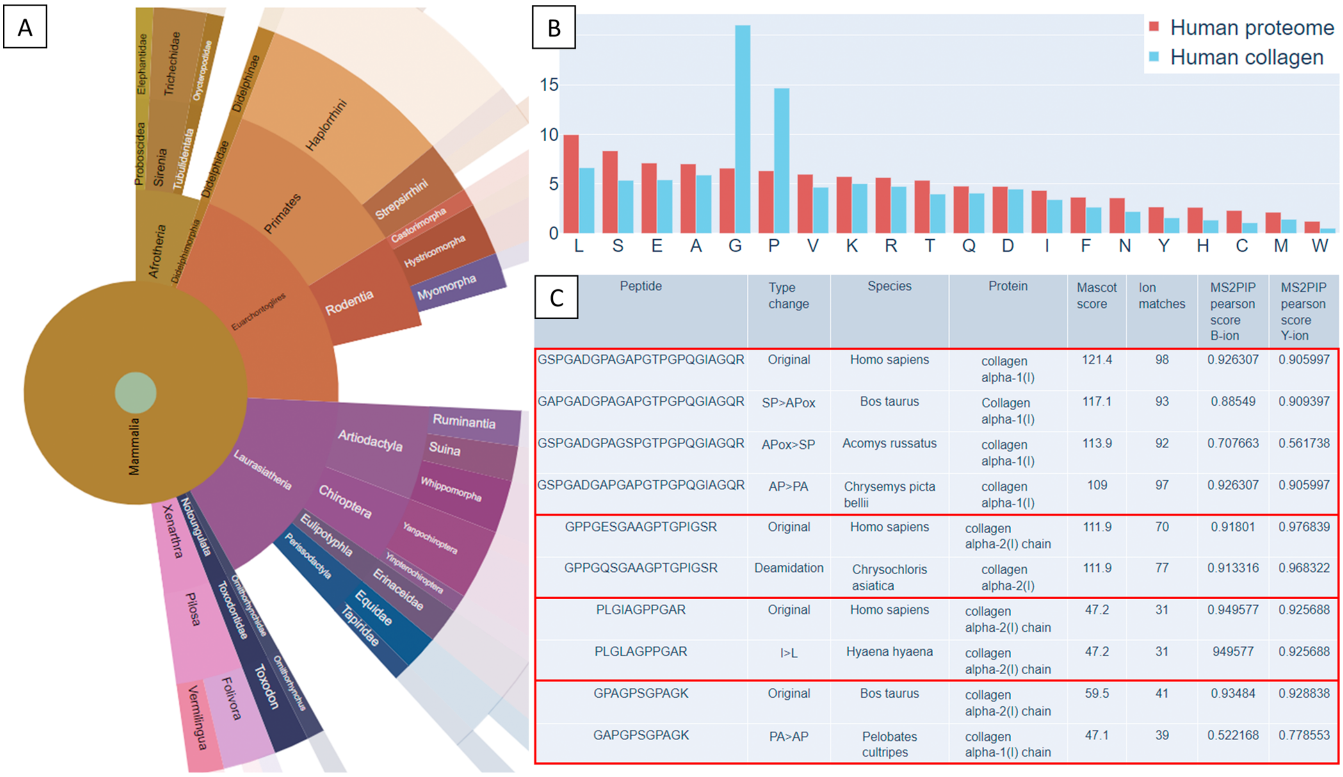
The isoBLAST and ClassiCOL Pipeline Explained

The isoBLAST and ClassiCOL pipeline is programmed as follows:

1. As input for isoBLAST, a Mascot CSV, MaxQuant TXT result file, or a generic CSV containing a peptide list with modifications, is required. From the input file the peptides, protein names, modifications, location of the modifications and spectrum titles are extracted. The protein names are used to filter out any keratin, trypsin and lys-C contaminants. All duplicate peptides are filtered out, so each peptide is only considered once. Next, all modifications are matched with the UniMod database to extract their mass changes. With this set of considered PTMs, the "isobaric output" is built, containing all possible isobaric peptides with up to two local isobaric modifications, including any PTMs directly extracted from the input file.
2. Database selection: by default, the CollagenDB containing all curated collagen sequences is used by the isoBLAST algorithm, yet the user can reduce the search space to one or more taxonomic levels. Users can also choose to search with alternative databases.
3. After extraction of all information from the search engine output file and the selected database, the isoBLAST algorithm generates a comprehensive list of potential peptide candidates. For each of the peptides three scenarios are possible:
    a. the peptide exactly matches a peptide in the CollagenDB.
    b. the peptide matches a sequence in the database that has ambiguity (B, Z, X amino acid annotations): the ambiguity is flagged for downstream processing.
    c. the peptide does not match any sequence. In this case, the mass is theoretically calculated, and the algorithm slides over all protein sequences (i.e., not only considering tryptic peptides), looking for an isobaric match. This needs to exactly match the peptide mass accompanied by combinatorial masses of the PTMs under investigation. Then, if the original peptide ends in a K or an R, it filters out only those candidates, otherwise all candidates (tryptic and semitryptic alike) are retained. All potential peptide−protein sequence matches are returned and aligned, allowing for gaps in either the measured or the theoretical peptide sequence. Then the algorithm will try to resolve each mismatch or gap in the alignment with the isobaric output. Only peptides that can be turned into the protein sequence match by introducing local isobaric switches are retained.
4. The list of candidate peptides is purified by discarding:
    a. all peptides that are isobaric to trypsin or Lys-C peptides;
    b. all single hit wonders (1 peptide to 1 protein);
    c. all flagged ambiguous peptides if no evidence for them is found in another species in the database, thus removing bias toward more completely sequenced closely related species.
5. Building the "collagen taxonomic" tree:

    To cope with the annotation bias of the proteins in the database, the ClassiCOL algorithm aligns (BLOSUM90) and calculates the distance between all probable proteins that are still under consideration. These distances are used to build the collagen tree.
6. Building the NCBI taxonomic tree:

    Next, the species taxonomic tree is built according to NCBI taxonomy using the taxoniq package, using only species for which peptides are still under consideration. The algorithm starts at the last common ancestor of all species that are under consideration. At every branching point, the algorithm looks to see if there is a difference in protein and/or peptide content between two branches. At multifurcations, bifurcation is "enforced" through the creation of a pseudobranch that contains all the species most related and bifurcates that away from the species that is least related. Note that for this, only peptide candidates in the data are considered, i.e., not genetic relatedness. So, for each iteration two branches are considered: when a branch is considered a subset of another branch, this branch is discarded; when both show signs of uniqueness, both are retained, and the sample is considered to be a mixture. The algorithm halts at a higher taxonomic level when no difference between the branches is found (see Figure 3 for what this implies to the user). After each potential end point has been found, the algorithm filters out all species which have less than 10% coverage of the total amount of collagen peptides in the



Peptide	Type change	Species	Protein	Mascot score	Ion matches	MS2PIP pearson score B-ion	MS2PIP pearson score Y-ion
GSPGADGPAGAPGTPGPQGIAGQR	Original	Homo sapiens	collagen alpha-1(I)	121.4	98	0.926307	0.905997
GAPGADGPAGAPGTPGPQGIAGQR	SP>APox	Bos taurus	Collagen alpha-1(I)	117.1	93	0.88549	0.909397
GSPGADGPAGSPGTPGPQGIAGQR	APox>SP	Acomys russatus	collagen alpha-1(I)	113.9	92	0.707663	0.561738
GSPGADGAPGAPGTPGPQGIAGQR	AP>PA	Chrysemys picta bellii	collagen alpha-1(I)	109	97	0.926307	0.905997
GPPGESGAAGPTGPIGSR	Original	Homo sapiens	collagen alpha-2(I) chain	111.9	70	0.91801	0.976839
GPPGQSGAAGPTGPIGSR	Deamidation	Chrysochloris asiatica	collagen alpha-2(I)	111.9	77	0.913316	0.968322
PLGIAGPPGAR	Original	Homo sapiens	collagen alpha-2(I) chain	47.2	31	0.949577	0.925688
PLGLAGPPGAR	I>L	Hyaena hyaena	collagen alpha-2(I) chain	47.2	31	949577	0.925688
GPAGPSGPAGK	Original	Bos taurus	collagen alpha-2(I) chain	59.5	41	0.93484	0.928838
GAPGPSGPAGK	PA>AP	Pelobates cultripes	collagen alpha-1(I) chain	47.1	39	0.522168	0.778553

Figure 2. Demonstration of taxonomic uncertainties from a ZooMS/MS output file. (A). Unipept[36] sunburst view on Mascot result of *Homo sapiens* reference sample of Rüther et al.[8] showing uniqueness of peptide stretches on different taxonomic levels (Figure S1). Sunburst plot is colored via phylogenetic relatedness. (B). Histogram showing the percentage amino acid distribution between the human reference proteome (UP000005640), collected from Uniprot (red) and solely collagen sequences extracted from the same human proteome from Uniprot (blue). (C). Demonstrative overview of *Homo sapiens* collagen peptides, alongside the top results with similar Mascot scoring. Due to an isobaric switch the protein of origin and species of origin can change. Unfortunately, intensity prediction and the accompanying MS^{2}PIP[63] score does not always change either, compromising this second-pass solution to further filter the peptide results.

sample. Also, all species are discarded that demonstrate >80% subset of a mixture of two other species.

7 Now, for each of the possible species a Bray−Curtis score is assigned based on the peptide content of the species compared against the total collagen peptide list (see Figure 3). Additionally, all peptides that show signs of in-sample decay are given a lower weight during scoring.

8 The ClassiCOL algorithm generates interactive sunburst plots for easy visualization of the results, as well as CSVs and other result files. The first sunburst plot depicts all the species that could not be filtered out during the taxonomic classification, color-coded by taxon score. A second sunburst plot additionally shows all of the taxa that are known to be missing from the protein databases, excluding those which are not relevant to the scored taxa, in gray. (*Id est*, if *Capra hircus* is one of the scored output taxa and the Ovis taxonomic branch has been discarded, *Ovis gmelini* will not appear in gray.) A heatmap shows the peptide count of species-protein matches in a bicluster analysis, and a barplot shows the count of unique peptide-to-species matches in the highest-scoring taxa. A line graph visualizes the scores given to the highest-scoring taxa before and after rescoring (see below). The first CSV output lists all scoring peptides and species within a sample, and another CSV lists the top outputs for each sample in a batch submission (when applicable).

9 Finally, a rescoring process is performed. From the result file, the top hits are reconsidered and rescored by the Bray−Curtis score while considering only this subset of peptides - de facto changing the ratio of the numerator and denominator and resulting in a more resolved scoring. Such rescoring has the added benefit of helping to resolve whether the sample is a genetic or a physical mixture, based on uniqueness among the top scoring hits (as described below). These results are visualized as a line plot of the score and the rescore as well as a barplot showing the overlap of unique peptides. Common peptides between the top results are discarded for the rescoring.

Peptide Identification

Raw datafiles were peak-picked by the MSConvert peak-picking algorithm into MGF file format.[58] The MGF files were submitted into MASCOT Daemon (version 2.8.2, Matrix Science, London, U.K.), and searched against the manually curated CollagenDB and a Universal Contaminants database (n = 381)[59] The enzyme was set to semiTrypsin with a maximum of 1 missed cleavage. Methylthio (C, + 45.987721 Da) was added as a fixed modification, and Deamidation (NQ, + 0.984016 Da), and Oxidation (MP, + 15.994915 Da) were added as variable modifications. The fragment error was set to 50 ppm and the peptide error tolerance was set to 10 ppm.

For species inference analysis, the Mascot search results were extracted in CSV data format restricted to a significance threshold of p < 0.01.

Publicly Available Data Sets

Raw datafiles were collected from the PRIDE ProteomeXchange repository with identifiers PXD024487,[8] PXD031386[60] and PXD042536,[61] and were processed in the same way as the in-house-generated DDA data sets. Morphological and analytical species classifications were downloaded from their respective papers.

■ RESULTS

Establishing CollagenDB, an Extensively Curated Collagen Database for ZooMS/MS

The most prominent obstacle in developing LC-MS/MS-based paleoproteomic approaches is the low number of peptides and their derived spectra, i.e., data sparseness, and in the case of ZooMS/MS, the exclusive focus on collagen proteins. We have compiled a database (CollagenDB) that is comprised of 221 species of Mammalia, supplemented by Reptilia (n = 45), Osteichthyes (n = 157), Chondrichthyes (n = 11), Aves (n = 162), Amphibia (n = 15), and Cephalopoda (n = 3), from which all collagen homologues, up to 45 per species, were downloaded in FASTA format from both the UniProt and NCBI repositories (Data S1). Hereby, the diversity of species searched is significantly extended compared to general ZooMS workflows, where this is still a limiting factor.[17] This collagen database contains all available collagen proteins (beyond the conventional COL1A1 and COL1A2) and will be continuously maintained on the ClassiCOL GitHub page as new species are submitted to the repositories.

This CollagenDB is at the heart of the proposed ZooMS/MS pipeline and is used at all three steps: the database search, isoBLAST, and ClassiCOL. Outputs from more constrained database searches can still be rescued through the isoBLAST ambiguation described below, albeit to a less precise taxonomic level, as also explained below.

Mapping the Ambiguity Problem in Conventional ZooMS/MS Searches

Collagen databases differ substantially from conventional proteome databases because of their highly similar tryptic peptide composition. When using databases consisting of highly similar protein sequences, conventional search engines like Mascot provide taxonomically ambiguous results similar to the more elaborately described protein inference problem[62] (Figure 2). Therefore, this equivalent "species inference problem" makes species classification very challenging.

Another layer of ambiguity is added when peptides cannot be distinguished from one another due to isobaric changes, which can lead to identification of false positive, species-specific, unique peptides (Figure 2A). In ZooMS/MS, isobaric changes include **(i)** a switch of two consecutive amino acids (positional isomer), **(ii)** a single residue isobaric switch including isoleucine-to-leucine (only resolvable in MS^n acquisitions), and N or Q deamidation versus D or E unmodified amino acids, **(iii)** monodipeptide changes e.g., AG to Q and **(iv)** combined PTM and amino acid switches, e.g., the isobaric alanine-to-serine substitution (+ 15.994915 Da) which can occur due to a nearby hydroxyproline (-PS- to -P$_{ox}$A-). The chance of correctly identifying isobaric shifts further decreases when the general amino acid distribution in the database is greatly shifted compared to full proteomes, for which these search algorithms are designed. This is certainly the case for collagenous samples, which are strongly enriched in glycine and proline (Figure 2B).

This issue is profound, as many amino acid differences between species involve isobaric changes. This has been recognized by Buckley et al.,[6] who suggested that the isobaric shifts within peptide sequences could only be resolved by an in-depth investigation of the tandem spectra.[6] Later, Rüther et al.[8] suggested to automate this process by calculating additional scores to filter the highest quality peptide matches, in contrast to Gilbert et al.[61] who have explored the capability of MS^3, wherein MS/MS fragments are further fragmented. Alternatively, standard first-hit-export files have been used, yet these contain a random selection of correct and incorrect annotations making it futile to use well-established species inference postprocessing algorithms such as Unipept.[36] On the other hand, exports containing all ambiguous peptide-to-species matches completely impair the use of a species-specific peptide selection.

Generally, these isobaric changes will barely affect the scoring, especially if the resolving fragment is low in abundance and the scoring will not be affected if the resolving fragment is absent (Figure 2C). Still, recent advances in intensity prediction could theoretically help to prioritize the different options identified by the search engine. We verified this by comparing the Pearson correlation of the measured intensities of the fragments with an MS^{2}PIP intensity prediction[63] for the original sequence and one of the aforementioned allowed isobaric changes by isoBLAST. Figure 2C shows how for several peptides from Figure 2A, Mascot scores of isobaric peptides are very close and could easily switch ranks based on slight changes in the spectral quality. Unfortunately, even machine learning-based algorithms that accurately predict the y-ion and b-ion intensities cannot always resolve these isobaric changes in peptide sequences.

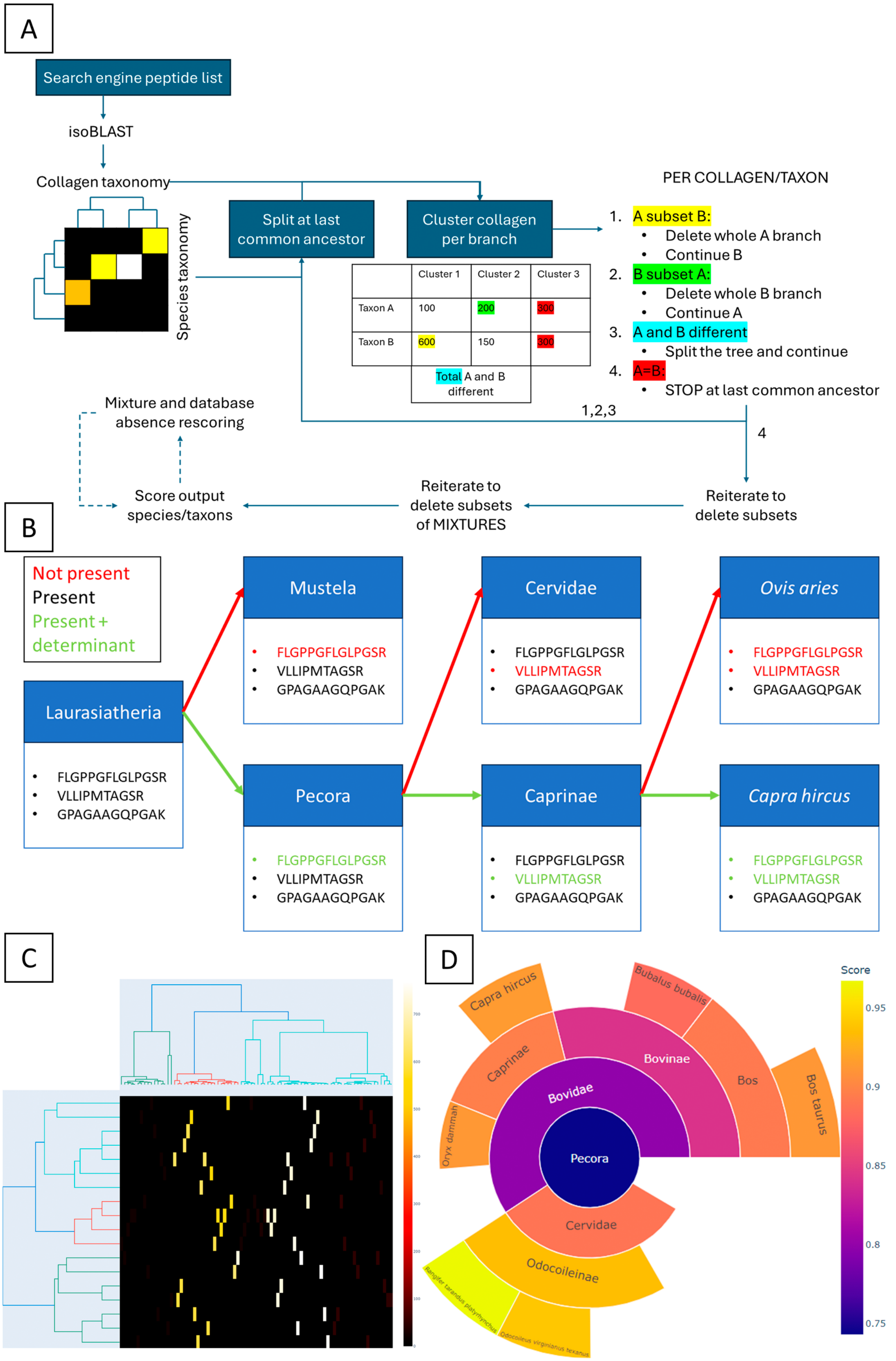
We consequently strategized a taxonomic classification based on a comprehensive list of peptides obtained through ambiguation with isoBLAST and subsequent classification with ClassiCOL, an anti-Occam's razor-inspired algorithm for decision-making at each taxonomic level.

Collagen Peptide Ambiguation through the Novel isoBLAST Approach

Inferring proteins is not the goal of ZooMS/MS approaches. Rather, it is the taxon (ideally down to the species) that must be inferred. To this end, we propose a novel approach, applicable to any ZooMS/MS search result output, irrespective of the instrument or search engine used.

Many search engines do not allow the user to export ambiguous results (and several spectra could potentially be matched with more than the maximum of ten peptide candidates with similar scores that are displayed in Mascot). Given that many tryptic peptides match to a multitude of different species entries in the database, the correct answer for a given spectrum is therefore frequently not exported. To maximize the chance that for each spectrum, the correct PSM is among the list of ambiguous annotations, we developed isoBLAST.

Conventionally, when peptide sequences need to be attributed to an organism, this is done through a BLAST algorithm, which uses an evolutionary background scoring matrix to assess the similarity to protein sequences in the database. However, when looking for all the peptides in a database that could explain a given spectrum, there is no use in allowing for non-isobaric changes, since the precursor mass is accurately measured (e.g., at 10 ppm mass error) and a non-isobaric change would additionally shift the masses of the rest of the b- and y-fragment ion series either side of the amino acid change. Therefore, isoBLAST searches the CollagenDB only for



	Cluster 1	Cluster 2	Cluster 3
Taxon A	100	200	300
Taxon B	600	150	300
	Total A and B different		

Figure 3. Overview of the ClassiCOL pipeline. (A) Schematic of the ClassiCOL algorithm workflow starting from a search engine output file. The numbers represented in the table are randomly chosen to highlight the possible difference between two taxonomic branches. (B). Schematic overview of peptide reuse from independent mutations after the speciation event, which are important for classification over different taxonomic levels. The peptides represented in the scheme are true peptides, from different spectra, that were found in the goat reference data set from Rüther et al.[8] (C).

Figure 3. continued

Visual representation of the collagen-species bicluster. The cluster is zoomed in at the Pecora family level on a sample from *Rangifer tarandus* (RMC42); the lighter the color in the heatmap, the greater the coverage for that collagen sequence. The y-axis represents the species taxonomic tree, the x-axis represents the homology between collagen sequences, colored by relatedness. (D). Output sunburst plot of the *Rangifer tarandus* sample, where higher scores represent higher likelihood that the named taxon approximates the sample content. All ambiguity is retained in the final output, whether originating from the isoBLAST approach or directly from the search engine results. The interactive version can be found as Figure S2.

isobaric changes, i.e., local, mono- to dipeptide amino changes that do not change the observed mass and are expected to have little or no impact on the scoring. As described above, these include (i) a positional isomer, (ii) isoleucine-to-leucine and N or Q deamidation (to D or E), (iii) mono- to dipeptide changes and (iv) the isobaric -PS- to -P_{ox}A- or any other isobaric changes involving a defined PTM.

Importantly, nonconsideration of relevant PTMs during the database searches can lead to misannotated sequences, "wrong" proteins, and "wrong" taxonomic assignments. Following this, the isoBLAST outcome will also differ from that produced when the original peptide is correctly annotated, with the same knock-on classification effects. It is necessary to carry out the initial database search with proline hydroxylation as a variable modification in order for this original peptidoform variation to be taken into account by isoBLAST when calculating the isoBLAST peptides, in part due to the maximum calculation of two isobaric switches per peptide. As the number of original peptidoforms has an impact on the eventual species identification based on amount of evidence (i.e., different peptidoforms in the isoBLAST outcome), there will be a tangible impact on the non/inclusion of relevant PTMs in the original database search. isoBLAST will only consider PTMs that occur in the input CSV file as relevant in the production of isoBLAST peptides: functionally this means that while P(ox)A can always be isoBLASTed into PS, PS cannot be isoBLASTed into P(ox)A when a search without PTMs is submitted.

Note that the initial search for each of the samples presented in this manuscript is done against the CollagenDB. That said, users can also provide any (correctly parsed) CSV output of searches against a more constrained database like e.g., Swissprot with a given taxon filter. In case of the latter, isoBLAST will increase ambiguity using the CollagenDB and coordinately extend the considered organism selection beyond the constrained database used during the initial probabilistic database search. If clear evidence of unexpected organisms is found in this way, the process is best reiterated on a database search performed against the CollagenDB. Without the use of CollagenDB, PPCs will likely be missed that could have an important impact on the final depth of taxonomic classification.

Taxonomic Classification of isoBLAST Output with ClassiCOL

In order to tackle the species inference problem, a novel taxonomic classification algorithm was developed specifically for the ambiguated data, since this ambiguation changes the prior assumption for classification. Conventionally, 100% correct annotations are assumed, whereby tryptic peptides can be found that are unique to a given taxon or species, as is envisioned in the Unipept approach. Yet, if the search algorithm cannot distinguish between two equally scoring peptides, it will propose only one, leaving the possibility of branching off into an entirely wrong lineage of taxonomy and moving the ambiguity downstream from the search algorithm to the classification algorithm. Instead, we resolve the ambiguity through classification. Additionally, semitryptic peptides are omnipresent in paleoproteomics data due to high levels of protein taphonomy, and are often not supported in downstream approaches, except for the Unipept tool. This currently results in the loss of valuable and informative peptides for taxonomic classification.

Now the axiom becomes that for each spectrum present in the data, the correct explanation is included, in an otherwise redundant list. Therefore, the algorithm (Figure 3A) starts by building a bicluster based on the available sequences. We create a taxonomic tree representing all species (according to NCBI taxonomy) matching at least 2 peptides to at least one protein sequence in CollagenDB, i.e., no "single hit wonders" are considered. Next, at every taxonomic branching point from the last vertebrate common ancestor to species, the two branches are compared one to one and can either be discarded or retained: when a branch cannot be distinguished from its counterpart because it is a subset of the other, it is discarded, yet when a difference at the peptide or protein level distinguishes it, the branch is retained (Figure 3B). When both branches have uniqueness to them, both are considered and the algorithm splits its search toward both branches separately, i.e., it considers the sample to be a mixture of ≥ 2 species.

We account for protein missingness in the following way: for collagens which are absent from a part of the taxonomic space, (i.e., some species are absent from some collagen clusters), that collagen cluster is not used to make a distinction in the taxonomic classification process. This rule is ignored once the species level is reached, as the protein distance scores of closely related species are more commonly (near-)equal. When branches cannot be further separated based on their peptide (and thus protein) content, (i.e., the peptides within both branches are identical) the algorithm halts at the taxonomic level of that branching point (Figure 3A). Notably, after every split, all initial peptides are reconsidered, having the benefit of retaining peptide sequences that originated from independent mutations that occurred after the speciation event (Figure 3B). In Unipept-like classifications, such peptides are usually plotted to the lowest common ancestor of the two species that express it, while all other species in that branch might not have this sequence. Because of the repeated decision-making at every taxonomic branch, these peptides are rescued and have proven to contribute to the taxonomic classification.

Figure 3C visualizes this approach with a simplified heatmap, exported from the ClassiCOL tool during every analysis. On the *x*-axis, all considered collagen sequences in the result file are clustered through a sequence homology matrix ("Collagen taxonomy"). On the *y*-axis, we depict the taxonomic tree representing all species from NCBI Taxonomy matching at least 2 peptides to at least one protein sequence ("Species taxonomy"). It is through this matrix that the anti-Occam's razor-inspired algorithm finds the species or taxonomic level that can best explain the comprehensive peptide-to-spectrum match space created through isoBLAST.

As for any classification algorithm, a single unique peptide difference is sufficient to retain a species. Therefore, multiple

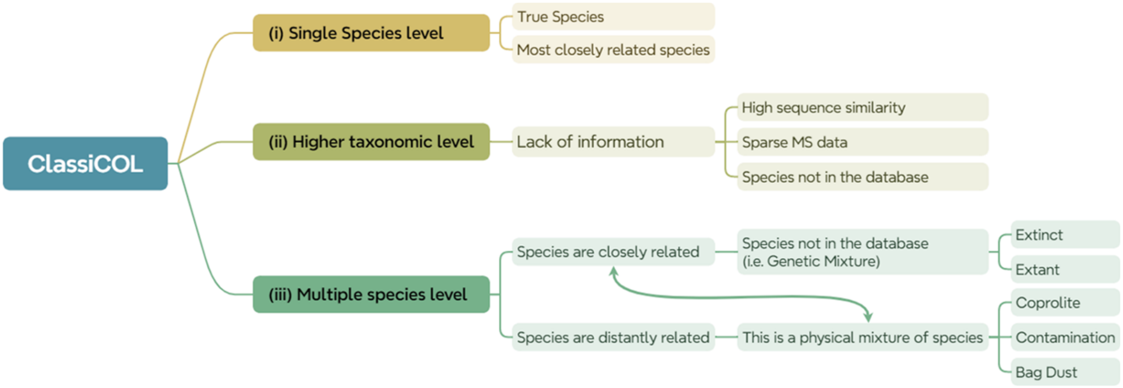

ClassiCOL
(i) Single Species level
True Species
Most closely related species
(ii) Higher taxonomic level
Lack of information
High sequence similarity
Sparse MS data
Species not in the database
(iii) Multiple species level
Species are closely related
Species not in the database
(i.e. Genetic Mixture)
Extinct
Extant
Species are distantly related
This is a physical mixture of species
Coprolite
Contamination
Bag Dust

Figure 4. Overview of the ClassiCOL outcome scenarios. This schematic (created with Xmind) shows all potential outcome scenarios of the ClassiCOL algorithm in the form of a decision tree. (i) Depicts the scenario of the identification of a single species, which can be the true species or the most closely related species in the database, (ii) shows how to interpret an outcome specific to a higher taxonomic level than the species level, and (iii) visualizes how to infer a physical and/or a genetic mixture from the ClassiCOL output files. We have included "coprolite" and "bag dust" as real-life examples of physical sample mixtures analyzed with ClassiCOL.

taxonomic levels and different species will be retained in the final output, reflecting the ambiguity provided by the Mascot search and isoBLAST algorithms. To facilitate user interpretation, an adapted Bray–Curtis pairwise distance dissimilarity[64] score is calculated by comparing the possible peptides of each species to the entire set of collagen peptides in the sample.

$$\text{classification score} = 1 - \left(\sum |u_i - v_i| / \sum |u_i + v_i|\right)$$

where u is a binary vector showing presence-absence of peptides for each species, and v represents a vector of all collagen peptides in the sample. In other words, it computes the distance between the species array and the sample array, adapted to 1- Bray–Curtis distance to calculate the similarity. Thus, the more a species contributes to the sample proportionally, the more likely it is for the sample to have originated from said species. Finally, an interactive sunburst plot is provided to the user as an output that still captures the underlying ambiguity, yet is color-coded according to the score to facilitate interpretation (Figures 3D and S2).

Resolving taxonomic outcomes will necessarily require the synergy of all context that is available to the analyst, including the sample itself (whether it is one sample or a mixture), the non-considered species in the database, the archeological context (location, chronology, climate; informing possible taxa), outcomes of other molecular analyses, etc. Each of these should inform the analyst toward greater accuracy of the final outcome. This necessary metadata was the impetus behind the formulation of an outcome interpretation decision tree (Figure 4). When applying ClassiCOL, three outcome scenarios are possible: **(i)** one single species was granted a higher score than all others, which indicates a true species match or a single-species match to the closest related extant species in the database; **(ii)** the algorithm displays the lowest non-species level taxonomy where lower branches cannot be further separated based on their peptide content, e.g., at the genus level, indicating that any member of that genus is considered to be an equally likely outcome; **(iii)** several different species are deemed to be similarly likely outcomes. In this third case, two separate explanations can be offered, depending on the relatedness of the species. First, when the two species are distantly related, this can reflect the output of a physical mixture of species present in the sample, which may also derive from contamination. Second, both species are very closely related, yet explicitly depicted as having evidence for being in the sample (as opposed to the algorithm stopping the branching process at a higher taxonomic level as in (ii)). This happens when the species to which the sample belongs is not represented in the CollagenDB, either because the sample species is extinct, or the sample is from an extant species which is not in the protein database, for example a roe deer (*Capreolus capreolus*). Irrespectively, we consider these to be "genetic mixtures" in the perspective of the database used.

Users should note that the classification algorithm has been programmed with the purpose of taxonomic identification based on all collagen isoforms, as these proteins have been known to persist over the longest periods of time and in high abundances. However, the algorithm does not necessarily rely on collagens alone, opening up the possible addition of non-collagenous proteins (NCPs) to the database in the future. Alternatively, a non-collagen custom database could be used. In fact, NCPs are typically less evolutionarily conserved and can be annotated more efficiently with an order-, family-, genus- or species-specific database after the bone sample has been classified to said level on the basis of collagen content,[65] although we do not extend on this strategy here. In this context, it is also relevant to note that the computational time increases linearly with both the number of peptides in the result file and the number of considered species in the CollagenDB, emphasizing the crucial role of bioarcheological preassessment. In the event that the user deploys prior knowledge to restrict the taxonomic outcome of a sample (e.g., to Mammals), the algorithm will still consider a maximum of 15 species from each of the Vertebrate Classes that do not belong to that taxonomy (e.g., vertebrates other than Mammals) as entrapment validation sequences, i.e., known false targets.[66] If the user chooses to restrict the results to taxa within the Pecora infraorder, the algorithm will still choose other non-Pecora members of the Mammal vertebrate class as entrapment targets, which maintains the possibility of false targets both closely and distantly related to the sample species.

Algorithm Performance on Public Data Sets

Different LC-MS/MS techniques and search algorithms have been reported for species identification using ZooMS/MS in recent years, all using a different (user-driven) decision-making process, including – most recently – MS^3 spectra.[61] Therefore, to increase user-friendliness to all users, regardless of expertise level, the Collagen Classification (ClassiCOL) algorithm was developed as a common postprocessing pipeline that requires

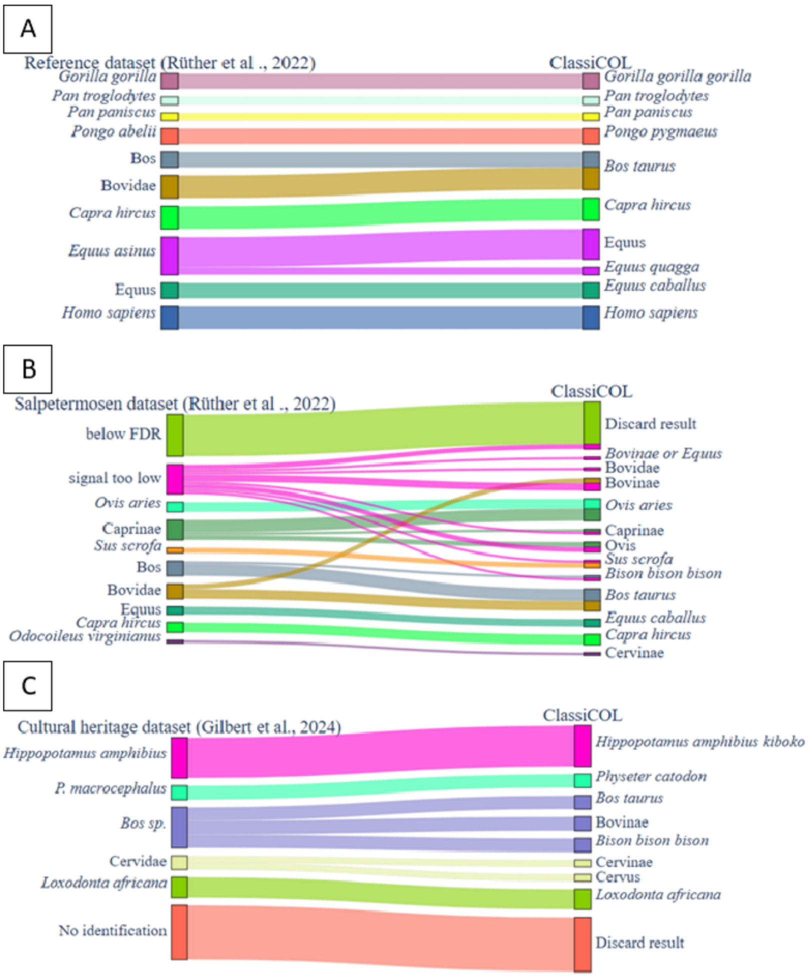


Figure 5. Similar, more precise classifications are computed by the ClassiCOL pipeline compared to the original papers' proteomics results. Sankey diagrams show the taxonomic classification approximation made by the ClassiCOL pipeline in comparison to the proteomics outcome of (A) the reference collection and (B). the Salpetermosen data sets from Rüther et al.,[8] and (C). a cultural heritage data set from Gilbert et al.[61] Colors are specific to the classification of the original paper per taxonomic level, and are directed to the classification made by the ClassiCOL algorithm. We observe that ClassiCOL can also rescue outcomes from samples with poor data quality, as seen for the (B) Salpetermosen "signal too low" input.

only a list of identified peptides as input, including PTMs and their respective localization. In order to demonstrate its potential and applicability toward a variety of different LC-MS/MS acquisition methods and sample types, we processed several publicly available data sets. In the process, we demonstrate how the different outcome scenarios can be interpreted (Figure 4). All individual interactive sunburst plots and an overview table of all the scorings for the public data sets are compiled as Data S2 and S3. As shown in Figure 5, our annotations agree well with the published interpretations.

First, we demonstrate the performance of the tool on a reference data set produced by Rüther et al.[8] (Figure 5A). Briefly, they describe a workflow wherein they made a site-specific difference matrix, only considering single amino acid mutation sites that differ between two species. This matrix was then used to score LC-MS/MS data after it was mapped to a multiple protein sequence alignment from multiple species in their database. They combined this metric with peptide intensities, peptide counts and precursor counts, giving more weight to higher quality MS/MS spectra to overcome ambiguity. Our results concur with their original proteomic identifications; both are in line with the reference collection bone morphologies. Still, we identified one species differently as being *Pongo pygmaeus* (Bornean orangutan), which matches the morphological identification and not the *Pongo abelii* (Sumatran orangutan) identification from Rüther et al.'s LC-MS/MS workflow (Figure S3). In this case, this derives from the fact that the correct protein sequence was not present in the database used in the original paper. On the other hand, we identified one of the reference samples as originating from *Equus quagga* (zebra) rather than *Equus asinus* (donkey), the latter of which was morphologically determined (Figure S4A). In this specific case, the recycling of peptides at each taxonomic decision level (Figure 3B) resulted in a high score given to a single peptide specific to zebra and horse but not to donkey, while all other peptides retrieved from the donkey sample are shared between zebra and donkey, therefore zebra was a more likely outcome (Figure S4B). Lastly, we managed to find evidence for three equine species when analyzing two mules in the reference collection data set, which corresponds to their mixed parentage from both *Equus caballus* (horse) and *Equus asinus* (Figures 4ii and S5). This particular form of hybrid genetic relatedness can be further interrogated through peptide-species match overlaps, as discussed later.

Next, we reanalyzed the samples excavated at the Salpetermosen site in Denmark[8] (Figure 5B). Because a lot of

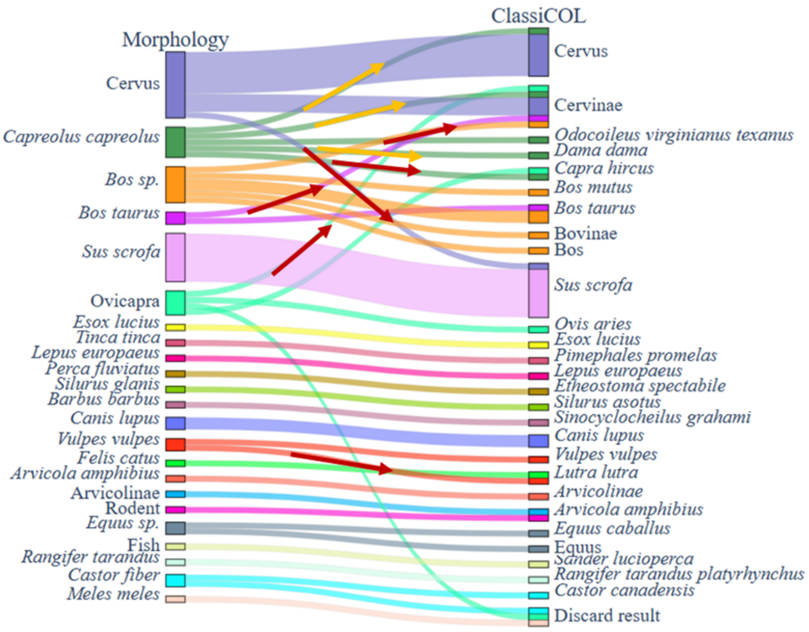


Figure 6. Analysis via ClassiCOL showed a few dissimilarities with bone morphology for in-house-processed samples. Connection graph shows the classification made by ClassiCOL on samples from ADP, RMC and OD. Red arrows indicate changes of classification to distantly related taxa based on the proteomics data. Yellow arrows show discrepancies with the morphological estimation within the same family.

archeological samples yield little high-quality data, several of these samples were classified as "signal too low" in the original manuscript. However, with our approach, leveraging the increased ambiguity that inevitably results from the low-quality data, we were able to classify several additional samples which all matched the morphological species classification (Figure 5B and Data S2 and S3). Furthermore, samples that were previously classified to the subfamily level (e.g., Caprinae), are now classified to the species level, again matching the morphology and demonstrating the performance of the algorithm (Figure S6). Additionally, one sample that was initially classified as Bos returned a likely match to a kind of bison, which could be derived from a species not in our database, e.g., *Bison bonasus* (European bison), given the age of the sample and the geographical location of the excavation site (Figure S7).

Lastly, there is also a lot of interest in identifying paint binders and bone- or hide glue in cultural heritage collections.[67−70] Therefore, we tested the applicability of the workflow in cultural heritage science, first focusing on data from a paper that identified the species of origin from bone and ivory museum pieces from the Smithsonian Institution, produced by Gilbert et al..[61] (Figure 5C). Here we were able to identify the species of origin without going into each spectrum separately and consulting the MS3 spectra created by fragmenting fragments from the MS/MS spectrum, as was necessitated by the original approach. Furthermore, we were able to pinpoint some samples to a more specific taxonomic level compared to the original manuscript. In fact, one sample was identified to have originated from a species of Bison, which was not considered in the paper (Figure S8). Also, as stated in the original paper, one of the artifacts was covered in glue, explaining the *Bos taurus* (cattle) identification of an ivory object.[61] This shows that our approach can also classify species based on collagen types other than I and II, such as COL3A1, which is a marker protein for hide-based glues which are commonly investigated in heritage paleoproteomics[70] (Figure S9).

We refer to Data S3 for a full overview of this validation on public data. Overall, incorrect species classification and annotation to an irrelevantly high taxonomic level were almost exclusively found in samples displaying a score below 0.6 for a single species and/or with fewer than 20 unique collagen peptides (peptidoforms). These criteria were therefore used as fixed thresholds to avoid ambiguous classification.

Classification of a Wider Selection of In-House-Processed Samples

To extend on the taxonomic coverage, we extracted proteins from extinct and ancient species from Belgian archeological sites and acquired the data through LC-MS/MS. First, a number of Ahrensburgian, Mesolithic, and Neolithic samples were selected from the Belgian archeological sites of Remouchamps ($n = 1$),[40] L'Abri du Pape ($n = 14$),[42] and Oudenaarde Donk (Neo 1) ($n = 41$),[47] including fish remains and microfauna. Similarly to the validation on public data, almost all sample results matched the initial morphological classifications, yet with greater taxonomic depth and with a few exceptions (see red arrows in Figure 6). In particular, one sample that was morphologically classified as *Cervus elaphus* (red deer; ADP006) came out as *Sus scrofa* (pig/boar) (Figure S10). Morphological reinspection of the proximal epiphysis revealed that it was indeed an exceptionally large wild boar specimen. Another example involved a bone originally classified as *Vulpes vulpes* (fox, ADP0014), which was classified by ClassiCOL as *Lutra lutra* (otter), an identification which was again confirmed through secondary morphological classification (Figure S11). Conversely, we also identified a sample as *Cervus elaphus* which was morphologically determined to be *Capreolus capreolus* (ADP0011), but in this case the thickness of the cortical bone excluded both adult and juvenile red deer (Figure S12). Thus, these very closely related species of cervids could not be disentangled by the algorithm, which can be largely credited to the absence of *Capreolus capreolus* in the database and highly similar collagen sequences within the family Cervidae, an issue aggravated by the poor coverage of collagen peptides in this sample.

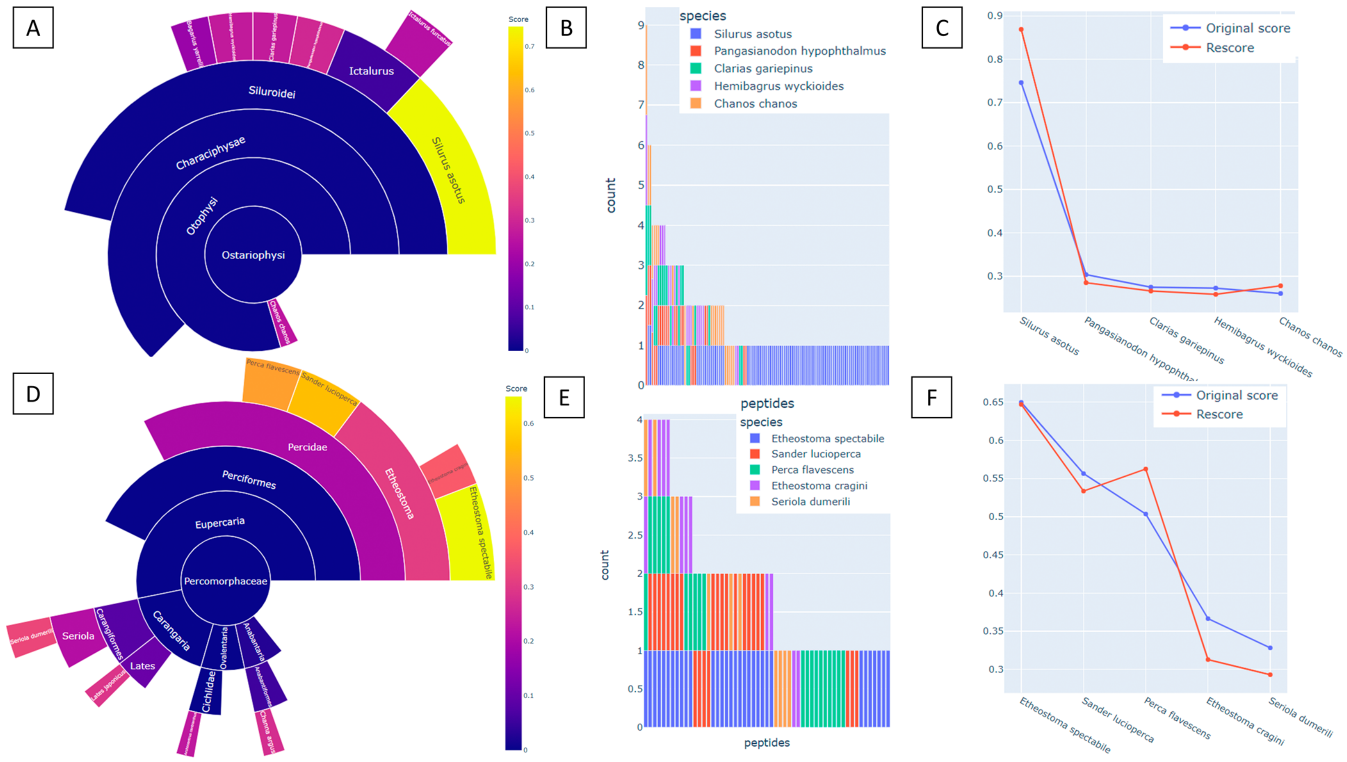


Figure 7. Rescoring highlights instances of genetic relatedness when the true species is not in the database. A−C. A fish bone morphologically identified to the genus *Silurus* (OD0004) analyzed with ClassiCOL. The barplot (B) and line graph (C) visualize high uniqueness for *Silurus asotus* (Amur catfish) both before and after rescoring, with a large drop-off for other species scores. The species *Silurus asotus* is likely the closest related species in the database to the species of origin—*Silurus glanis* (Wels catfish), based on the geographical location and time period. D−F. Rescoring of morphologically estimated *Perca fluviatus* (European perch; OD0002) with ClassiCOL shows potential for genetic relatedness with other members of the family Percidae. Barplot (E) shows a significant nonoverlapping difference between sample peptides shared with *Etheostoma specabile* (orangethroat darter), *Sander lucioperca* (sander) and *Perca flavescens* (yellow perch). Line graph (C) demonstrates similar rescoring distributions, suggesting that the true species is not included in the fish database, but can be found in the family Percidae.

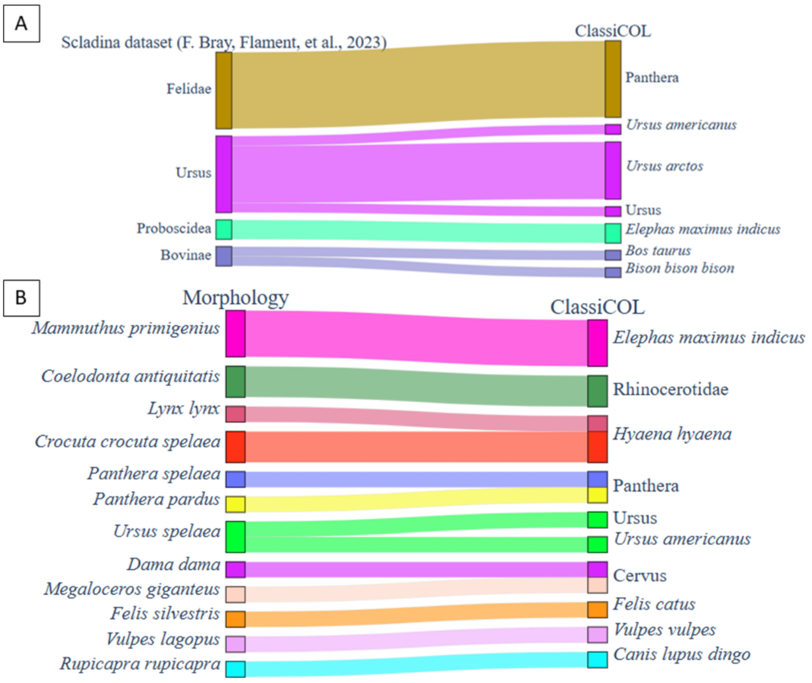


Figure 8. Performance of ClassiCOL on extinct mammals. (A) Reanalysis of the publicly available data of Bray et al.[60] provides identifications matching the original publication, with algorithmic taxa identification where manual curation was needed in the original paper. (B). ClassiCOL's performance on in-house-processed extinct mammal samples excavated from the same site as Bray et al., namely Scladina Cave,[60] from which both bag dust and single bone samples were analyzed.

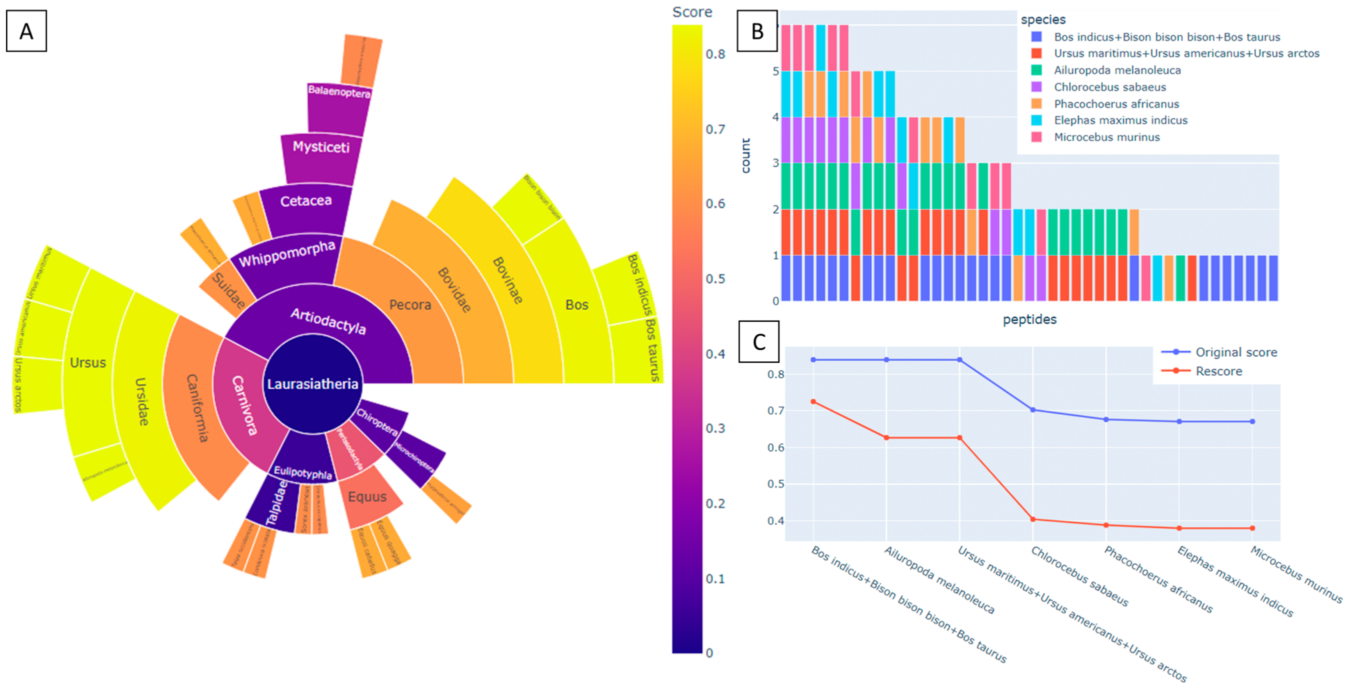


Figure 9. Analysis of a cave hyena coprolite via ClassiCOL with subsequent mixture reanalysis. (A). Output of ClassiCOL shows high scores for two distantly related species, indicating a sample mixture. (B). Barplot depicts 1 isoBLAST peptide difference among bears and high uniqueness for both *Ursus* and *Bovinae*. (C). Rescoring showed no drop-off between bear groups nor was a drop-off seen for Bovinae, a drop-off is noticeable after Ursus. These results show that the sample is comprised of both ursine and bovine bone fragments.

Some classification discrepancies were impossible to morphologically reassess as the bone fragments were too small, too degraded, or too similar to closely related species to give an accurate reassessment (OD024, OD08, OD042, OD029 and OD31). (Figures S13−S17)

Regarding the fish fragments, we saw that several of the bones that were morphologically identified as fish were indeed classified as fish, yet often to a higher taxonomic level (scenario Figure 4ii), incentivizing us to expand the database to 168 species of "fish" (including Actinopterygians and Chondrichthyans). The same analysis was performed, and gave a more accurate species approximation compared to when the species of origin was not considered in the first search (scenario Figure 4i). Indeed, all actinopterygian specimens were identified to have originated from the morphologically estimated genus or species, when included in the database (Figures S18−S23). Notably, several fish genera are not yet present in any database. These species returned results indicating a mixture of close relatives under the same taxonomic family. Next, we assessed the performance of the algorithm on actual mixtures and on species not included in the database, including extinct animals.

Challenging Samples: Rescoring Mixtures and Unsequenced Species

With the tool's performance well-established on bones that were sampled specifically for the identification of species, we turned to applications in paleoproteomics that require the identification of different species from the same sample (Figure 4iii). The ability of the ClassiCOL approach to retain multiple species outputs per sample, coupled with the scoring approach, enables inference of physical as well as genetic mixtures, including extinct species.

First, to better resolve cases of extant and extinct species not present in the database, we strategized a data-driven selection of the top five candidate taxa represented in the sunburst output, followed by a recalculation of the Bray−Curtis score on the discriminant peptides only, i.e., a second-pass search. From the in-house-processed fish remains, we isolated a one-species and a multiple-species sunburst (Figure 7A,D), isolated the unique peptides (Figure 7B,E) and recalculated the Bray−Curtis score (Figure 7C,F) in a second-pass classification. For a single high scoring species, the second-pass rescoring can still offer a higher score drop-off to increase the certainty of the annotation (Figure 7C). Yet, in the case that a species is not in the CollagenDB, this drop-off is less pronounced, i.e., a gradual decline in scoring is obtained, approaching a more detailed relatedness of a genetic mixture. Thus, while not obligatory for single species, this extra algorithmic step leads to a higher resolution in classification, much like second-pass Percolator rescoring can improve peptide annotation in database searches,[29] especially when dealing with physical and/or genetic mixtures.

Next, we assessed how this approach reflects extinct species. Therefore, we reanalyzed a public data set from Bray et al.,[60] which was focused on >100,000-year-old remains excavated in Scladina Cave in Belgium (Figure 8A). Here as well, all classifications matched the taxonomic level presented in the manuscript, yet most samples were classified to the species level present in the database, i.e., an extant relative. Strikingly, several samples resulted in a sunburst plot that looked more like a mixture. As an example, a cave bear bone resulted in an ambiguous annotation of ursine species, which is not the same as the algorithm stalling at a higher taxonomic level. Still, this shows that extinct animals that are not in the database will be challenging to interpret, mimicking sample mixtures, and giving ambiguous results of two closely related species at best.

In parallel to the Ahrensburgian, Mesolithic and Neolithic samples, we processed and analyzed several extinct animal specimens from Scladina Cave, the same site as Bray et al..[60] Again, we observed that the species classification agreed with the

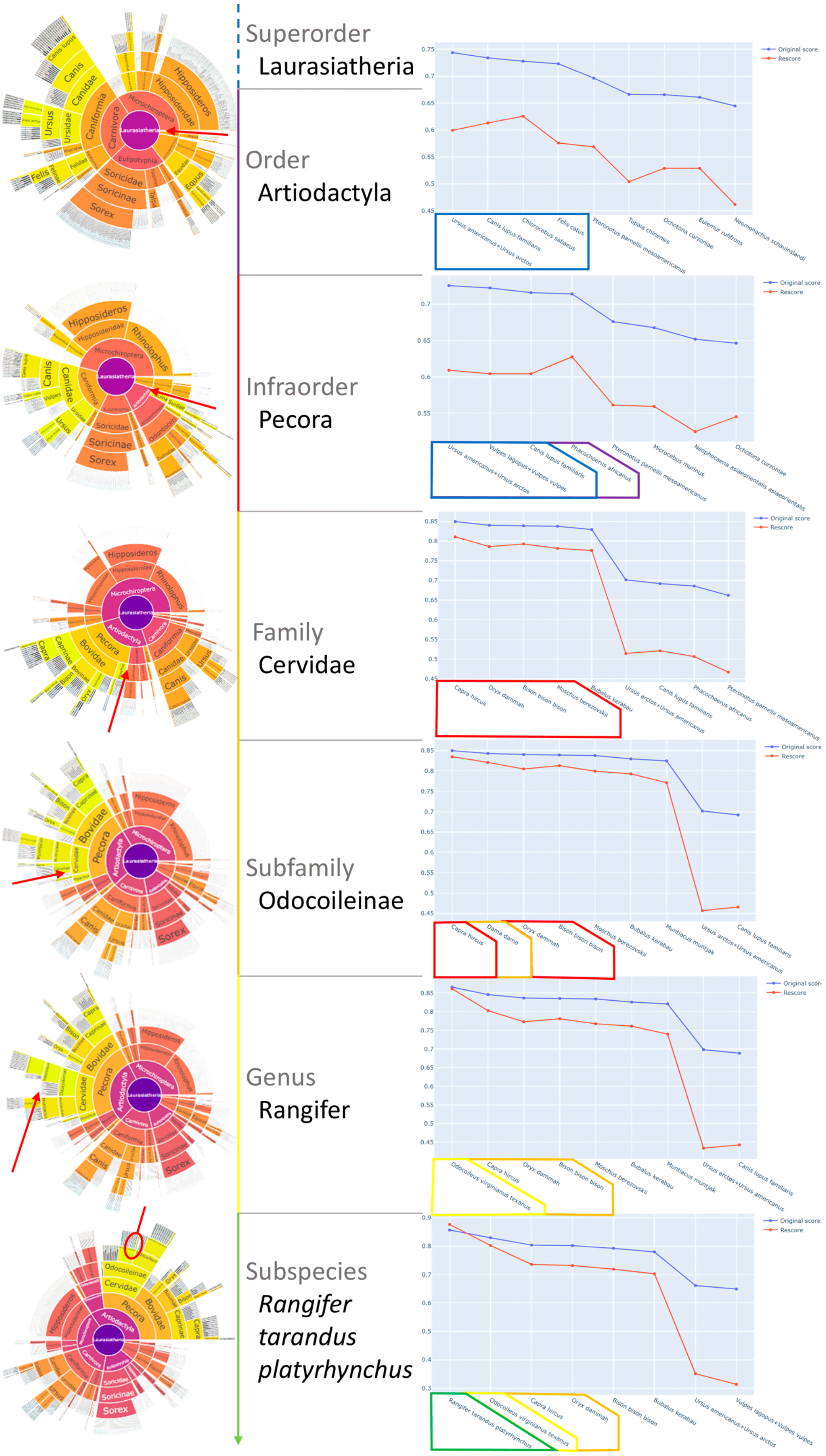


Figure 10. Algorithmic behavior in response to taxonomic missingness. A single LC-MS/MS file, derived from a morphologically determined *Rangifer tarandus* specimen, was searched using MASCOT, each time sequentially restricting the CollagenDB by consecutively removing the Artiodactyla, Pecora, Cervidae, Odocoileinae and Rangifer taxonomic levels. The output csv files were fed into the ZooMS/MS pipeline (isoBLAST and

Figure 10. continued

ClassiCOL), alongside their respectively restricted databases. The sunburst plots (left) show the taxonomic classification result, with the explicit visualization of missingness linked to each candidate's taxonomic lineage, i.e., taxa from the NCBI taxonomy classifier that were not considered during the ClassiCOL classification. Red arrows indicate most relevant level of missingness for this sample. For each of the taxa absence databases (middle), the scored and rescored results are shown (right). The highest-scoring results against each of the databases are highlighted and color-coded according to the taxonomic level in the middle, based on their relationship (last common ancestor) with the morphologically estimated *R. tarandus*.

morphological classification (Figure 8B). Here, we were able to classify *Coelodonta antiquitatis* only to the family Rhinocerotidae because of the absence of wooly rhino and the closest living relative in the database (Figure S24). However, due to the geographic location of the site and Upper Pleistocene dating of the bone, wooly rhino was the only possible outcome. Again, this emphasizes that interdisciplinary expertise will always be required for robust classification.

Additionally, two samples had to be morphologically reassessed due to discrepancies with the ClassiCOL output. The samples formerly identified as being *Lynx lynx* and *Rupicapra rupicapra* were classified by ClassiCOL as Hyaena hyaena (striped hyena) and Canis lupus dingo (dingo) respectively (Figures S25 and S26). The *Hyaena hyaena* identification was confirmed morphologically as the mandible most likely originated from a cave hyena cub, and the Canis identification was confirmed due to shape, curvature, and muscle attachments on the bone of the juvenile specimen. For this analysis, when possible, we sampled the bone both directly and indirectly as disintegrated bone dust in the storage bag, each of which produced highly covered collagen sequences (COL1A1 and COL1A2) and taxonomically classified the sample correctly according to the morphology. In other words, both direct invasive sampling and swabbed bone dust from containers can be used for analysis. This limiting of destructive sampling is an exciting development as minimally destructive sampling is important for responsible heritage curation and paleoproteomics as a research field.[71]

Physical mixtures are of great interest, because bags or boxes commonly contain several bones that can derive from different species, which could theoretically all be identified in a single run. Another challenging prospect is the analysis of coprolites, which can contain mixtures of extinct animals. Therefore, we analyzed the dust from a bag containing cave hyena coprolites (*Crocuta crocuta spelaea*) excavated in Scladina Cave (Layer T-RO, early Aurignacian). These paleofeces preserve information on carnivore meals and potentially collagen bone fragments from multiple extinct animals, as hyenas are opportunistic scavengers. Figure 9A shows the sunburst plot depicting distantly related species that point toward a physical mixture and closely related species within each of these branches, pointing to a genetic mixture, in turn potentially indicative of extinct species. During mixture rescoring, we therefore allowed the algorithm to take more than the default five and thus to include all possible outcome species (Figure 9B). Via the aforementioned rescoring, the potential candidates in the mixture were confirmed (Figure 9C)—all three of its replicates contained two high-scoring distantly related species: *Bison bonasus* and *Ursus spelaeus* (Figure 9).

Algorithmic Performance When Dealing with Taxonomic Missingness

To explore the impact of taxonomic missingness on searches, for example in the case that an entire taxonomic branch to which a target species belongs is extinct, we carried out a taxonomic exclusion experiment in which progressively higher taxonomic levels were excluded from CollagenDB in the pipeline. We start from a morphologically identified *Rangifer tarandus* bone, from the Remouchamps archeological site. In this approach, we remove increasingly higher taxonomic levels along the *R. tarandus* taxonomic branch and assess the outcome in each case. From top to bottom, Figure 10 displays what happens if the taxonomic resolution in CollagenDB increases (lower taxonomic levels are added) to mimic all possible situations wherein respectively the order, infraorder, family, subfamily, genus and subspecies of a given sample are not present in the database. With increasing taxonomic specificity made available in CollagenDB, the new result always contains the highest-scoring answer, which increases confidence in the classification with each iteration. The absence of organisms which are listed in NCBI taxonomy but which do not have collagen protein counterparts for classification (i.e., are not present in CollagenDB) is visualized in the gray slices in the sunbursts. The gray slices are present at the lowest taxonomic level for each branch, provided that the branch is one of the scored outcomes and has not been entirely discarded during classification. Given that this experiment is carried out on a sample from a single bone, the effect of the taxonomic exclusions is to increase the evolutionary distance between the target species and the closest in-database relative. As well as non-sequenced extant species, this outcome approximates the analysis of extinct species, which appear in the output as genetic mixtures. The color-coding of the candidate taxa in the line graphs according to the level of shared ancestry with *R. tarandus* demonstrates that the highest-scoring hits are increasingly aberrant from the expected outcome.

DISCUSSION

Our novel ZooMS/MS pipeline builds on the most extensively curated collagen database to date (CollagenDB), and combines isoBLAST ambiguation and ClassiCOL taxonomic classification to approximate or identify the species of origin for archeological bone samples starting from the search results of any LC-MS/MS run. These analyses are intrinsically more information-rich than a ZooMS analysis because they contain peptide fragment data, yet the ambiguity in search results has long hampered their routine implementation. Therefore, we here embrace and extend the ambiguity found in searches against an extensive manually curated CollagenDB, using our isoBLAST approach, to ensure that the correct annotation for each spectrum is always present. With this new premise, the ClassiCOL taxonomic classifier is then applied to isolate the species that explains most of the potential peptide candidates in the isoBLAST result space. By comparing the results of our algorithm to several publicly available data sets, we have demonstrated that the ClassiCOL algorithm output matches existing results, routinely gives more precise species identifications and can rescue poor quality data. In contrast to other approaches, ClassiCOL is not reliant on unique peptide stretches, but can differentiate between species based on overall peptide content, provided through a search engine output file or peptide list in CSV format, including the reconsideration of peptide stretches representing independent mutations after speciation events. As the CollagenDB consists of all collagen types, heritage samples containing bone or hide glues can also be analyzed through the same pipeline without any adaptations. Finally, our approach shows promise to resolve genetic mixtures, i.e., species not in the database, whether extant or extinct, as well as different species in a physical mixture, as demonstrated on dust from coprolites.

The main limitation of ClassiCOL is the interpretability of the final sunburst plots. Therefore, we provide examples of the three main scenarios of outcomes, guide the user by including a decision tree, and provide a benchmark experiment for users to interpret the outcome when taxonomic branches are absent from CollagenDB. Still, the very fact that all potential peptide candidates are considered during classification, in combination with sample complexity and degradation, sometimes leads to complex sunburst plots that require interrogation. ClassiCOL is, elementally, an analytical algorithm developed to guide decision-making, and we advise that all analyses should be posteriorly assessed by a zooarcheologist.

■ ASSOCIATED CONTENT

Data Availability Statement

All data are available in the main text or the Supporting Information. All original analyses were performed with ClassiCOL version 1.0.0 which is made available on GitHub (https://github.com/EngelsI/ClassiCOL). The in-house-generated Sciex WIFF, MGF and mzTab files have been deposited in the PRIDE repository (https://www.ebi.ac.uk/pride/) with identifier PXD055222. Other supplementary data, the sunbursts and other output files are also deposited on the PRIDE repository in. ZIP format.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.4c00962.

All figures referenced in the main text, including: a Unipept sunburst output from reference *Homo sapiens* data (Figure S1); an interpretation of the ClassiCOL pipeline heatmap and sunburst output plots of a *Rangifer tarandus* sample (Figure S2); sunburst identifications of the public datasets and in-house datasets (Figure S3-S5, S7-S26); and a figure showing the use of non-common collagen proteins and re-use of peptides on different taxonomic levels resulting in more precise classifications (Figure S6). A detailed extraction protocol which was used for the in-house samples is also included. (PDF)

The ClassiCOL output files from the taxonomic missingness experiment, with sunburst plots supplied in .html format and output peptide lists as CSVs. (ZIP)

Additional files, including the FASTA file of the CollagenDB, the output overview file from all experiments performed (XLSX), and a list of proteins included in the CollagenDB (XLSX). Further sunburst plots (html) output peptide lists (CSV) are made available in the PRIDE repository under accession PXD055222. (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Maarten Dhaenens − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent 9000, Belgium;* orcid.org/0000-0002-9801-3509; Email: maarten.dhaenens@ugent.be

Authors

Ian Engels − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent 9000, Belgium;* orcid.org/0009-0003-1059-7541

Alexandra Burnett − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent 9000, Belgium; ArcheOs Laboratory for Biological Anthropology, Faculty of Arts and Philosophy, Ghent University, Ghent 9000, Belgium;* orcid.org/0000-0002-2810-0500

Prudence Robert − *ArcheOs Laboratory for Biological Anthropology, Faculty of Arts and Philosophy, Ghent University, Ghent 9000, Belgium*

Camille Pironneau − *ArcheOs Laboratory for Biological Anthropology, Faculty of Arts and Philosophy, Ghent University, Ghent 9000, Belgium*

Grégory Abrams − *ArcheOs Laboratory for Biological Anthropology, Faculty of Arts and Philosophy, Ghent University, Ghent 9000, Belgium; Scladina Cave Archaeological Centre, Espace muséal d'Andenne, Andenne 5300, Belgium*

Robbin Bouwmeester − *VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Biomolecular Medicine, Ghent University, Ghent 9052, Belgium;* orcid.org/0000-0001-6807-7029

Peter Van der Plaetsen − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent 9000, Belgium; ArcheOs Laboratory for Biological Anthropology, Faculty of Arts and Philosophy, Department of Biomolecular Medicine, Department of Applied Mathematics, Computer Science and Statistics, and Department of Biochemistry, Ghent University, Ghent 9000, Belgium; VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Evolution & Diversity Dynamics Lab, UR Geology, Université de Liège, Liège 4000, Belgium; Scladina Cave Archaeological Centre, Espace muséal d'Andenne, Andenne 5300, Belgium; Dept. of Anthropology, MSC01 1040, University of New Mexico, Albuquerque, New Mexico 87131-0001, United States; CNRS, UAR 3290 - MSAP - Miniaturisation pour la Synthèse, l'Analyse et la Protéomique, Univ. Lille, Lille F-59000, France; Archéologie préhistorique, Département des sciences historiques, Université de Liège, Liège 4000, Belgium; School of Biological and Environmental Sciences, Research Centre in Evolutionary Anthropology and Palaeoecologys, Liverpool John Moores University, Liverpool L3 3AF, U.K.*

Kévin Di Modica − *Scladina Cave Archaeological Centre, Espace muséal d'Andenne, Andenne 5300, Belgium*

Marcel Otte − *Archéologie préhistorique, Département des sciences historiques, Université de Liège, Liège 4000, Belgium*

Lawrence Guy Straus − *Dept. of Anthropology, MSC01 1040, University of New Mexico, Albuquerque, New Mexico 87131-0001, United States*

Valentin Fischer − *Evolution & Diversity Dynamics Lab, UR Geology, Université de Liège, Liège 4000, Belgium*

Fabrice Bray − *CNRS, UAR 3290 - MSAP - Miniaturisation pour la Synthèse, l'Analyse et la Protéomique, Univ. Lille, Lille F-59000, France;* orcid.org/0000-0002-4723-8206

Bart Mesuere − *VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9052, Belgium; Department of Applied Mathematics, Computer Science and Statistics and Department of Biochemistry, Ghent University, Ghent B-9000, Belgium;* orcid.org/0000-0003-0610-3441

Isabelle De Groote − *ArcheOs Laboratory for Biological Anthropology, Faculty of Arts and Philosophy, Ghent University, Ghent 9000, Belgium; School of Biological and Environmental Sciences, Research Centre in Evolutionary Anthropology and Palaeoecologys, Liverpool John Moores University, Liverpool L3 3AF, U.K.*

Dieter Deforce − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent 9000, Belgium*

Simon Daled − *ProGenTomics, Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent 9000, Belgium;* orcid.org/0000-0001-8129-2217

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jproteome.4c00962

Author Contributions

[§§]I.E., A.B., S.D. and M.D.: contributed equally to this work. I.E., A.B., S.D., M.D.: Conceptualization. P.R.: Sample collection Abri du Pape, Oudenaarde-Donk, and Remouchamps. G.A.: Sample collection Scladina Cave. K.D.M., M.O., V.F., G.A., L.S., P.V.dP.: Permissions to extract original proteomics samples. C.P., G.A., K.D.M.: Morphological reassessment of remains. I.E., A.B.: Protein extraction. A.B., S.D., M.D.: Optimization of LC-MS/MS measurements. I.E., R.B., B.M.: Coding and visualization. I.E.: Data processing. M.D., D.D., I.D.G.: Supervision. I.E.: Optimization and validation of algorithm:. I.E., A.B., S.D., M.D.: Writing—original draft. G.A., P.R., I.D.G., C.P.: Writing—archeological contribution, writing—review and editing. All authors provided feedback on the manuscript.

Funding

This research was funded by a grant from the Special Research Fund (BOF) of Ghent University with funding code BOF.GOA.2022.0002.03.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to give special thanks to Rachel Siân Dennis for laying the groundwork for the protein extraction protocol, and giving valuable insights during the adaptation of said protocol. Also, we would like to express our gratitude to ProGenTomics (www.progentomics.be) for ongoing support in data acquisition and instrumental optimizations. We extend our heartfelt thanks to all reviewers and editors for their insights and suggested improvements to our manuscript.

■ REFERENCES

(1) Le Meillour, L.; Zazzo, A.; Lesur, J.; et al. Identification of degraded bone and tooth splinters from arid environments using palaeoproteomics. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **2018**, *511*, 472−482.

(2) Demarchi, B.; Mackie, M.; Li, Z.; et al. Survival of mineral-bound peptides into the Miocene. *eLife* **2022**, *11*, No. e82849, DOI: 10.7554/eLife.82849.

(3) Demarchi, B.; Hall, S.; Roncal-Herrero, T.; et al. Protein sequences bound to mineral surfaces persist into deep time. *eLife* **2016**, *5*, No. e17092, DOI: 10.7554/eLife.17092.

(4) Hendy, J.; et al. Ancient proteins from ceramic vessels at Çatalhöyük West reveal the hidden cuisine of early farmers. *Nat. Commun.* **2018**, *9*, No. 4064.

(5) Buckley, M.; Wadsworth, C. Proteome degradation in ancient bone: Diagenesis and phylogenetic potential. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **2014**, *416*, 69−79.

(6) Buckley, M.; Lawless, C.; Rybczynski, N. Collagen sequence analysis of fossil camels, Camelops and c.f. Paracamelus, from the Arctic and sub-Arctic of Plio-Pleistocene North America. *J. Proteomics* **2019**, *194*, 218−225.

(7) Paterson, R. S.et al. A 20+ Ma old enamel proteome from Canada's High Arctic reveals diversification of Rhinocerotidae in the middle Eocene-Oligocene *bioRxiv* 2024 DOI: 10.1101/2024.06.07.597871.

(8) Rüther, P. L.; Husic, I. M.; Bangsgaard, P.; et al. SPIN enables high throughput species identification of archaeological bone by proteomics. *Nat. Commun.* **2022**, *13*, No. 2458.

(9) Le Meillour, L.; Sinet-Mathiot, V.; Ásmundsdóttir, R. D.; et al. Increasing sustainability in palaeoproteomics by optimizing digestion times for large-scale archaeological bone analyses. *iScience* **2024**, *27*, No. 109432.

(10) Mylopotamitaki, D.; Harking, F. S.; Taurozzi, A. J.; et al. Comparing extraction method efficiency for high-throughput palaeoproteomic bone species identification. *Sci. Rep.* **2023**, *13*, No. 18345.

(11) Horn, I. R.; Kenens, Y.; Palmblad, N. M.; et al. Palaeoproteomics of bird bones for taxonomic classification. *Zool. J. Linn. Soc.* **2019**, *186*, 650−665.

(12) Shoulders, M. D.; Raines, R. T. Collagen structure and stability. *Annu. Rev. Biochem.* **2009**, *78*, 929−958.

(13) Bray, F.; Fabrizi, I.; Flament, S.; et al. Robust High-Throughput Proteomics Identification and Deamidation Quantitation of Extinct Species up to Pleistocene with Ultrahigh-Resolution MALDI-FTICR Mass Spectrometry. *Anal. Chem.* **2023**, *95*, 7422−7432.

(14) Buckley, M.; Collins, M.; Thomas-Oaies, J.; Wilson, J. C. Species identification by analysis of bone collagen using matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 3843−3854.

(15) Brown, S.; Higham, T.; Slon, V.; et al. Identification of a new hominin bone from Denisova Cave, Siberia using collagen fingerprinting and mitochondrial DNA analysis. *Sci. Rep.* **2016**, *6*, No. 23559.

(16) Ostrom, P. H.; Schall, M.; Gandhi, H.; et al. New strategies for characterizing ancient proteins using matrix-assisted laser desorption ionization mass spectrometry. *Geochim. Cosmochim. Acta* **2000**, *64*, 1043−1050.

(17) Richter, K. K.; Codlin, M. C.; Seabrook, M.; Warinner, C. A primer for ZooMS applications in archaeology. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, No. e2109323119.

(18) Bekker-Jensen, D. B.; Martínez-Val, A.; Steigerwald, S.; et al. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. *Mol. Cell. Proteomics* **2020**, *19*, 716−729.

(19) Dallongeville, S.; Garnier, N.; Rolando, C.; Tokarski, C. Proteins in Art, Archaeology, and Paleontology: From Detection to Identification. *Chem. Rev.* **2016**, *116*, 2−79.

(20) Aebersold, R.; Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **2016**, *537*, 347−355.

(21) Verheggen, K.; Ræder, H.; Berven, F. S.; et al. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev.* **2020**, *39*, 292−306.

(22) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data *Electrophoresis*, *20* 3551 3567 DOI: 10.1002/(SICI)1522-2683(19991201)20:183.0.CO;2-2.

(23) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protoc.* **2016**, *11*, 2301−2319.

(24) Cox, J.; Neuhauser, N.; Michalski, A.; et al. Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794−1805.

(25) Elias, J. E.; Gygi, S. P.Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. In *Methods in Molecular Biology*; Springer, 2010; Vol. *604*, pp 55−71.

(26) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.

(27) Ma, K.; Vitek, O.; Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinf.* **2012**, *13*, No. S1.

(28) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.

(29) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1719−1727.

(30) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **2020**, *17*, 41−44.

(31) Van Puyvelde, B.; et al. Removing the Hidden Data Dependency of DIA with Predicted Spectral Libraries. *Proteomics* **2020**, *20*, No. 1900306.

(32) Buckley, M.; et al. Collagen Sequence Analysis of the Extinct Giant Ground Sloths Lestodon and Megatherium. *PLoS One* **2015**, *10*, No. e0139611.

(33) Buckley, M.; Harvey, V. L.; Orihuela, J.; et al. Collagen Sequence Analysis Reveals Evolutionary History of Extinct West Indies Nesophontes (Island-Shrews). *Mol. Biol. Evol.* **2020**, *37*, 2931−2943.

(34) Behrensmeyer, A. K. Taphonomy. *Encyclopedia of Geology* **2021**, 12−22.

(35) Grupe, G.Taphonomy and Fossilization. In *International Encyclopedia of Biological Anthropology*; Wiley, 2018; pp 1−8.

(36) Mesuere, B.; Devreese, B.; Debyser, G.; et al. Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* **2012**, *11*, 5773−5780.

(37) Schallert, K.; Verschaffelt, P.; Mesuere, B.; et al. Pout2Prot: An Efficient Tool to Create Protein (Sub)groups from Percolator Output Files. *J. Proteome Res.* **2022**, *21*, 1175−1180.

(38) Rahir, E. *L'habitat Tardenoisien Des Grottes de Remouchamps, Chaleux et Montaigle.* (Bruxelles, 1921).

(39) Dewez, M. Remouchamps - Préhistoire. *Bull. Soc. R. Belg. Anthropol. Préhistoire* **1974**, *85*, 42−111.

(40) Crombé, P.; Pironneau, C.; Robert, P.; et al. Human response to the Younger Dryas along the southern North Sea basin, Northwest Europe. *Sci. Rep.* **2024**, *14*, No. 18074.

(41) Léotard, J. M. Occupations préhistoriques à l'Abri du Pape (Roches de Freyr − Dinant). *Notae Praehistoricae* **1989**, *9*, 27−28.

(42) Léotard, J. M.; Straus, L. G.; Otte, M.*L'abri du Pape. Bivouacs, Enterrements et Cachettes sur la Haute Meuse Belge : du Mésolithique au Bas Empire Romain − Bivouacs, Burials and Retreats along the Upper Belgian Meuse : from the Mesolithic to the Low Roman Empire*; ERAUL (Etudes et Recherches Archeologiques de l'Universite de Liege), 1999; Vol. *88*, p 365.

(43) Noiret, P.; et al. Recherches paléolithiques et mésolithiques en Belgique, 1993 : le Trou Magrite, Huccorgne et l'Abri du Pape. *Notae Praehistoricae* **1994**, *1994*, 45−62.

(44) Otte, M.; et al. Fouilles 1994 à l'Abri du Pape et à la grotte du Bois Laiterie (Province de Namur). *Notae Praehistoricae* **1994**, *14*, 45−68.

(45) Parent, J.-P.; Van der Plaetsen, P.; Vanmoerkerke, J. Mesolithische en neolithische sites aan de Donk te Oudenaarde. *Archeol. Belg.* **1986**, *2*, 15−18.

(46) Parent, J.-P.; Van Plaetsen, P.; Vanmoerkerke, J. *Les Fouilles de Sauvetage d'Oudenaarde-Donk. Les Cahiers de Préhistoire Du Nord*. 1989; Vol. *6*.

(47) Parent, J.-P.; Van Der Plaetsen, P.; Vanmoerkerke, J. Prehistorische jagers en veetelers aan de donk te Oudenaarde. *VOBOV-Inf.* **1986**, 24−25.

(48) Ameels, V.; et al. Recent steentijdonderzoek in de regio Oudenaarde (Oost-Vlaanderen, België). *Notae Praehistoricae* **2003**, *23*, 61−65.

(49) Otte, M.Recherches Aux Grottes de Sclayn; Etudes et Recherches Archéologiques de l'Université de Liège, 1992; Vol. *1*, p 182.

(50) Abrams, G.; Devièse, T.; Pirson, S.; et al. Investigating the co-occurrence of Neanderthals and modern humans in Belgium through direct radiocarbon dating of bone implements. *J. Hum. Evol.* **2024**, *186*, No. 103471.

(51) Pirson, S.et al.The Palaeoenvironmental Context and Chronostratigraphic Framework of the Scladina Cave Sedimentary Sequence (units 5 to 3-SUP). In *Scladina I-4A Juvenile Neandertal*; Etudes et Recherches Archéologiques de l'Université de Liège: Andenne, Belgium, 2014; pp 69−92.

(52) Pirson, S. The Stratigraphic Sequence of Scladina Cave. In *The Scladina I-4A Juvenile Neandertal*; Toussaint, M.; Bonjean, D., Eds.; Etudes et Recherches Archéologiques de l'Université de Liège: Andenne, Belgium, 2014; pp 49−68.

(53) Charters, D.; Brown, R. P.; Abrams, G.; et al. Morphological evolution of the cave bear (Ursus spelaeus) mandibular molars: coordinated size and shape changes through the Scladina Cave chronostratigraphy. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **2022**, *587*, No. 110787.

(54) Charters, D.; Brown, R. P.; Abrams, G.; et al. Mandibular ecomorphology in the genus ursus (Ursidae, Carnivora): relevance for the palaeoecological adaptations of cave bears (U. spelaeus) from Scladina cave. *Hist. Biol.* **2024**, 1−15.

(55) Abrams, G.; Bello, S. M.; Di Modica, K.; Pirson, S.; Bonjean, D. When Neanderthals used cave bear (Ursus spelaeus) remains: Bone retouchers from unit 5 of Scladina Cave (Belgium). *Quat. Int.* **2014**, *326-327*, 274−287.

(56) Abrams, G. Palaeolithic Bone Retouchers from Belgium: A Preliminary Overview of the Recent Research through Historic and Modern Bone Collections. In *Origins of Bone Tool Technologies*; Hutson, J. M., Ed.; Roman-Germanic Central Museum - Leibniz Research Institute for Archaeology, 2018; pp 197−214.

(57) Pirson, S. *Contribution à L′′etude des d́epôts d'Entŕee de Grotte en Belgique au Pĺeistocène SuṕerieurFacult́e des Sciences*Université de Liège: Liège, 2007; p 435.

(58) Chambers, M. C.; Maclean, B.; Burke, R.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918−920.

(59) Frankenfield, A. M.; Ni, J.; Ahmed, M.; Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *J. Proteome Res.* **2022**, *21*, 2104−2113.

(60) Bray, F.; Flament, S.; Abrams, G.; et al. Extinct species identification from late middle Pleistocene and earlier Upper Pleistocene bone fragments and tools not recognizable from their osteomorphological study by an enhanced proteomics protocol. *Archaeometry* **2023**, *65*, 196−212.

(61) Gilbert, C.; Krupicka, V.; Galluzzi, F.; et al. Species identification of ivory and bone museum objects using minimally invasive proteomics. *Sci. Adv.* **2024**, *10*, No. eadi9028, DOI: 10.1126/sciadv.adi9028.

(62) The, M.; Edfors, F.; Perez-Riverol, Y.; et al. A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms. *J. Proteome Res.* **2018**, *17*, 1879−1886.

(63) Degroeve, S.; Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **2013**, *29*, 3199−3203.

(64) Bray, J. R.; Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **1957**, *27*, 325−349.

(65) Cappellini, E.; Jensen, L. J.; Szklarczyk, D.; et al. Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *J. Proteome Res.* **2012**, *11*, 917−926.

(66) Feng, X. dong.; Li, L. w.; Zhang, J. h.; et al. Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics* **2017**, *18*, No. 143.

(67) Haghighi, Z.; Mackie, M.; Apalnes Ørnhøi, A.; et al. Palaeoproteomic identification of the original binder and modern contaminants in distemper paints from Uvdal stave church, Norway. *Sci. Rep.* **2024**, *14*, No. 12858.

(68) Ge, R.; Zhou, L.; Zhang, Y.; Liu, J.; Yang, L. A scientific study of a Han ancient adhesive: First discovery of the use of cattle bone powder in pottery bonding. *J. Cultural Heritage* **2024**, *67*, 277−283.

(69) Fremout, W.; Dhaenens, M.; Saverwyns, S.; et al. Tryptic peptide analysis of protein binders in works of art by liquid chromatography−tandem mass spectrometry. *Anal. Chim. Acta* **2010**, *658*, 156−162.

(70) Scibè, C.; Eng-Wilmot, K.; Lam, T.; et al. Palaeoproteomics and microanalysis reveal techniques of production of animal-based metal threads in medieval textiles. *Sci. Rep.* **2024**, *14*, No. 5320.

(71) Hendy, J. Ancient protein analysis in archaeology. *Sci. Adv.* **2021**, *7*, 9314−9329.