The De-Modifier: a tool for screening modification-induced alternate peptide taxonomy in palaeoproteomics

Miranda Evans[1,2*]

*** Corresponding Author.**
ORCID: 0000-0002-9284-9268
Email address: mae52@cam.ac.uk
1: Department of Archaeology, University of Cambridge, Cambridge, UK
2: BioArCh, Department of Archaeology, University of York, York, UK

Key words: Palaeoproteomics, Taxonomy, Authentication, Bioinformatics

Abstract

In palaeoproteomic research, the accuracy of taxonomic matches is crucial, as research questions frequently hinge on which species were utilised by ancient people. However, protein modifications including deamidation of glutamine and asparagine, and conversion of glutamine or glutamic acid to pyroglutamic acid, can change the sequence of peptides, leading to spurious taxonomic detections and potentially inaccurate archaeological interpretations. While a handful of examples of this phenomenon have been reported in the literature, the issue is potentially much wider reaching than currently realised. In reality, any time a peptide containing a deamidated glutamine or asparagine residue, an unmodified glutamic acid or aspartic acid residue, or a pyroglutamic acid modification is detected by proteomic search software, the sequence, and therefore potentially its taxonomy, may be incorrect, which could potentially lead to unsound archaeological interpretations. `The De-Modifier` is a fast, open-access tool which solves this issue by screening for modification-induced alternate peptide taxonomy, enabling archaeologists to make informed interpretations of the taxonomies of proteins detected in ancient samples. To assess its utility, `the De-Modifier` is tested against an archaeological dataset containing all unique peptides reported in palaeoproteomic studies of dental calculus and vessels. The results reveal that modification-induced alternate peptide taxonomies are severely under-reported, occurring 15% more frequently than previously understood. Modifications were found to produce three different types of inaccurate taxonomic matches: those which yielded

completely different taxonomic lowest common ancestors to the input peptide, those which were more taxonomically specific than the input peptide, and those which were less taxonomically specific than the input peptide. `The De-Modifier` therefore enables the detection of potentially inaccurate peptide taxonomies, avoiding spurious archaeological interpretations in future studies.

Introduction and Background

In palaeoproteomics, the reported taxonomy of ancient peptides is assessed by analysing the lowest common ancestor (LCA) of all the taxa to which a given peptide matches. It is crucial that these taxonomic assignments are secure, because they frequently form the basis of archaeological interpretations, e.g. the detection of newly introduced plant taxa (Scott et al., 2021), multitaxa dairying (Evans et al., 2023), or mixed ingredient processing (Hendy et al., 2018a). Deamidation is a post-translational modification (PTM) in which asparagine (N) is converted to aspartic acid (D), and glutamine (Q) to glutamic acid (E). It is caused by the loss of an amide group which causes a mass shift of +0.98 Da (Figure 1) which is widely reported in both MS1 (ZooMS) and MS2 analysis. Asparagine has been found to deamidate at a higher rate than glutamine, which deamidates considerably more slowly at natural pH via the formation of a cyclic intermediate (Robinson et al., 2004). Given that deamidation occurs at a relatively regular rate under constant conditions, it has been proposed as a molecular clock (Robinson et al., 2004), a marker of preservational quality (Schroeter and Cleland, 2016; van Doorn et al., 2012), or even an indicator of processing (Nair et al., 2023). Deamidation rates have been used as a tool for establishing the authenticity of palaeontological, archaeological or heritage proteins, initially using bulk deamidation rates (Hill et al., 2015; Orlando et al., 2013; van Doorn et al., 2012; Welker et al., 2017, 2015), and later using site-specific measures (Mackie et al., 2018; Ramsøe et al., 2020). Deamidation rates are also impacted by protein structure (Kossiakoff, 1988; Robinson et al., 2004; Xie and Schowen, 1999), temperature (Stratton et al., 2001), and pH (Hao et al., 2011; Wilson et al., 2012).

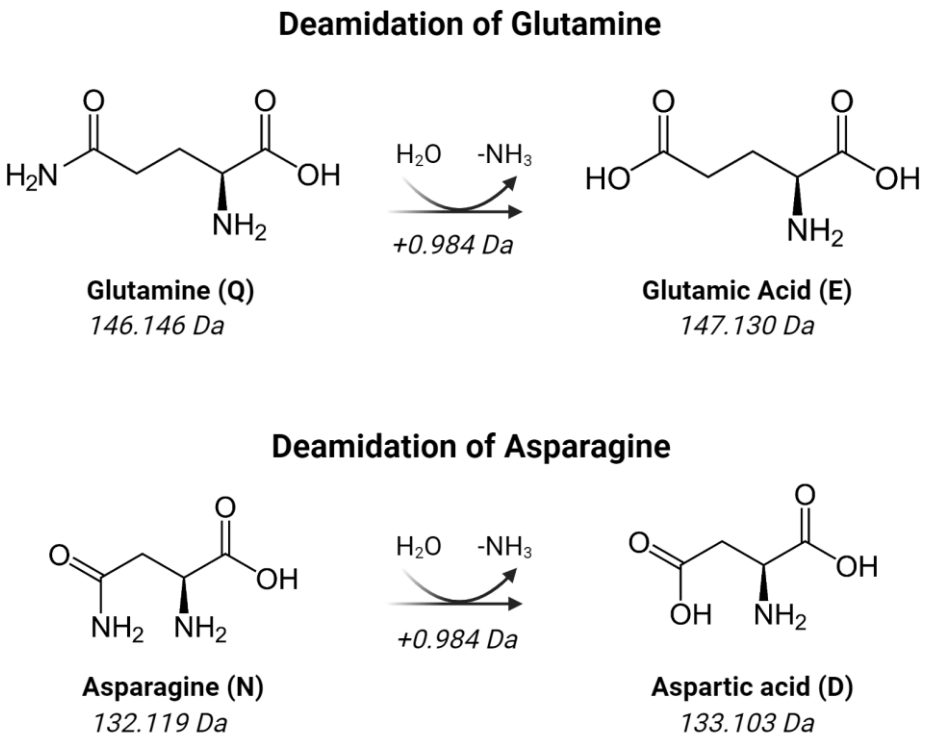

Figure 1: top: Simplified illustration of deamidation of Glutamine to Glutamic Acid, middle: deamidation of Asparagine to Aspartic acid. Figure created in Biorender.com

Recognising that the mass of unmodified aspartic acid (D) is indistinguishable from deamidated asparagine (N), Warinner et al. (2014) observed that such modification induced sequence permutations (henceforth MISPs) could result in potentially incorrect LCA matches. In particular, deamidation of an asparagine residue in the Bovinae whey protein β-lactoglobulin (βLG) peptide TPEVD**D**EALEK caused it to be indistinguishable from the *Ovis* aries βLG peptide TPEVD**N**EALEK, meaning that the detection of unmodified TPEVD**D**EALEK by proteomic software could equally be derived from Bovinae or *Ovis* milk in ancient samples. Crucially, detection of inaccurate dairy species could lead to spurious interpretations of dairy exploitation and species introduction. Given that the peptide in question is more frequently detected than any other in ancient dental calculus and ceramic residues (Hendy et al., 2018a, 2018b; Jeong et al., 2018; Wilkin et al., 2020), this discovery has gained widespread application (e.g. Charlton et al., 2019; Evans et al., 2023; Hendy et al., 2018a, 2018c). Several similar sequence shifts have been noted in the interim, especially for other βLG peptides. Jersie-Christensen et al. (2018) reported that the βLG peptide

PTPEG**D/N**LEILLQK which they detected in medieval dental calculus samples could match to either Bovinae or *Ovis* respectively. Analysing neonate dog bones from 430–960 CE Japan, Tsutaya et al. (2019) reported that the peptide TME**D/N**LDLQK could belong to either Canidae, the most parsimonious interpretation, or alternatively to *Callorhinus ursinus* (northern fur seal), *Odobenus rosmarus divergens* (Pacific walrus), and *Leptonychotes weddellii* (Weddell seal). Lastly, in their analysis of a well-preserved Early Bronze Age wooden container from Switzerland, Colonese et al. (2017) reported an MISP for the peptide TLFGS**D/N**PPIPTPVLTK, which could render its initial LCA match to *Hordeum vulgare* insecure. To date, no E/Q MISPs have been reported in the ancient protein literature, which is perhaps unsurprisingly given the lower rate of E/Q deamidation compared to N/D, however, deamidation of glutamine could also potentially result in incorrect taxonomic detection.

Another PTM which causes amino acid mass shifts that could potentially lead to sequence permutations and thus, inaccurate LCAs is the conversion of N-terminus glutamine or glutamic acid to pyroglutamic acid (Gln →pyro-Glu or Glu →pyro-Glu). Pyroglutamic acid forms via two different pathways (Figure 2): the loss of an amide group (i.e. deamidation) converts glutamine to pyroglutamic acid, or alternatively, the loss of a water molecule (i.e. dehydration) converts glutamic acid to pyroglutamic acid. The conversion of glutamic acid to pyroglutamic acid occurs when it is heated over 180°C, while glutamine to pyroglutamic acid is exacerbated by enzymatic activity in vitro (for a review: Kumar and Bachhawat, 2012). Due to the identical masses of glutamine minus NH_3, pyro-Glu, and glutamic acid minus H_2O, any time a peptide with the Gln→pyroGlu (n-term Q) or Glu → pyroGlu (n-term E) modification is detected by search software, the peptide n-terminus amino acid could equally be a glutamic acid (E) or glutamine (Q). So far, no instances of this modification causing MISPs have been reported in the archaeological literature. However, its implications could potentially be far reaching.

Conversion of Glutamine and Glutamic Acid to Pyro-Glutamic Acid

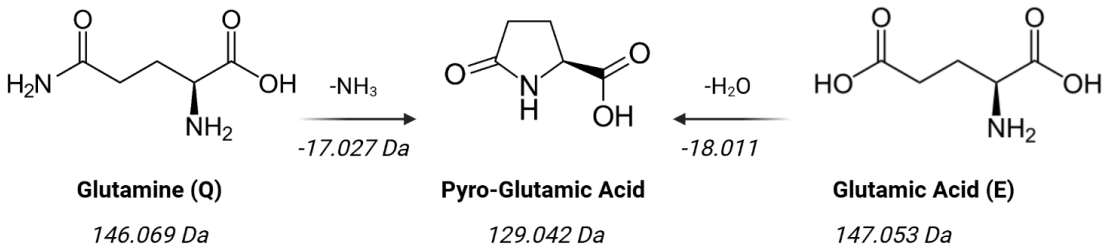

Figure 2: Conversion of Glutamine and Glutamic Acid to Pyro-Glutamic Acid. Figure created in Biorender.com

Given the lack of tools for MISP detection, researchers seeking to screen for MISPs are left to manually calculate deamidation permutations and their taxonomies, which is a time-consuming task, and is yet to be standard practice. It is thus perhaps unsurprising that none of the reported examples involve multiple modification sites, despite the equal mass of multi-deamidation site MISPs and those with single modifications, and the input sequence. Peptides with numerous potential deamidation sites drastically increase the number of possible alternate sequences (permutations) and potentially their taxonomies, making their detection time-consuming without programmatic tools.
Therefore, despite the handful of reported examples, the issue of modification-induced shifts to taxonomic detection is likely much more widespread than previously realised. In reality, any time a peptide containing a deamidated Q or N, an unmodified E or D or a Gln/Glu→ pyroGlu modification is detected by search software, the sequence and potentially its taxonomy may be incorrect.

To enable researchers to combat this issue, I present `The De-Modifier`. This tool screens for all possible modification-induced sequence permutations to a given list of peptides and assigns taxonomy using the Unipept API (Mesuere et al., 2016). I then use `The De-Modifier` to screen published ancient dietary peptides for alternative, equally possible taxonomic detections. The results reveal many modification-induced sequence alterations resulting in equifinal taxonomic detections, which have yet to be recognised in previous palaeoproteomic studies. As more ancient metaproteomic analyses are published, further modification-induced sequence permutations will likely be identified.

Materials and Methods

The De-Modifier

`The De-Modifier` screens for potential modification-induced sequence changes, which yield alternate taxonomies to the input peptide. It consists of a Python3 script (https://github.com/miranda-e/Demodifier) which, given a csv containing one column of peptides and one column of their corresponding modifications in either Mascot or Maxquant format, compiles all possible deamidation-induced permutations of that input sequence for

Glutamine and Asparagine residues and all possible pyro-Glu induced permutations. It then calls the Unipept API (Mesuere et al., 2016) using the pept2lca function to gather the lowest LCAs for each peptide permutation and compiles these results for the researcher's perusal.

`The De-Modifier` simulates four potential substitutions (Figure 3). First, `The De-Modifier` simulates *undoing* deamidation modifications: i.e. it "re-amidates" all permutations of glutamic acid and aspartic acid up to the number of deamidations detected for a given input peptide by search software, "returning" them to glutamine and asparagine respectively. Only permutations with an identical mass to the input peptide are created: i.e. those with the same number or fewer deamidated residues than the input peptide, given that only modified glutamic acid and aspartic acid residues have an identical mass as glutamine and asparagine, respectively. For each resultant *re-amidation* permutation, all permutations of *deamidation* of glutamine and asparagine are then simulated (Figure 3) by replacing these residues with glutamic acid and aspartic acid, respectively, with no maximum number of substitutions (given that unmodified glutamine and asparagine residues have the same mass as modified glutamic acid and aspartic acid residues respectively). Finally, for each resultant permutation, `The De-Modifier` then simulates Q to E substitutions for peptides bearing N-terminus Q with the Gln → pyroGlu modification detection, and E to Q substitutions for peptides bearing N-terminus E with Glu → pyroGlu modification detection (Figure 3). Any duplicate peptide permutations for a given input peptide are removed before all remaining peptide permutations are assigned LCA using the Unipept pept2LCA API (Mesuere et al., 2016). The resultant data are saved in two output csv files, the first of which contains input peptides annotated with their input LCA, their total number of MISPs, the number of MISPs that yielded LCAs, the specific MISPs that yielded LCA results, their resulting LCAs, and whether the resulting LCAs are identical. The second output csv contains all MISPs generated and their LCAs, where they were available.

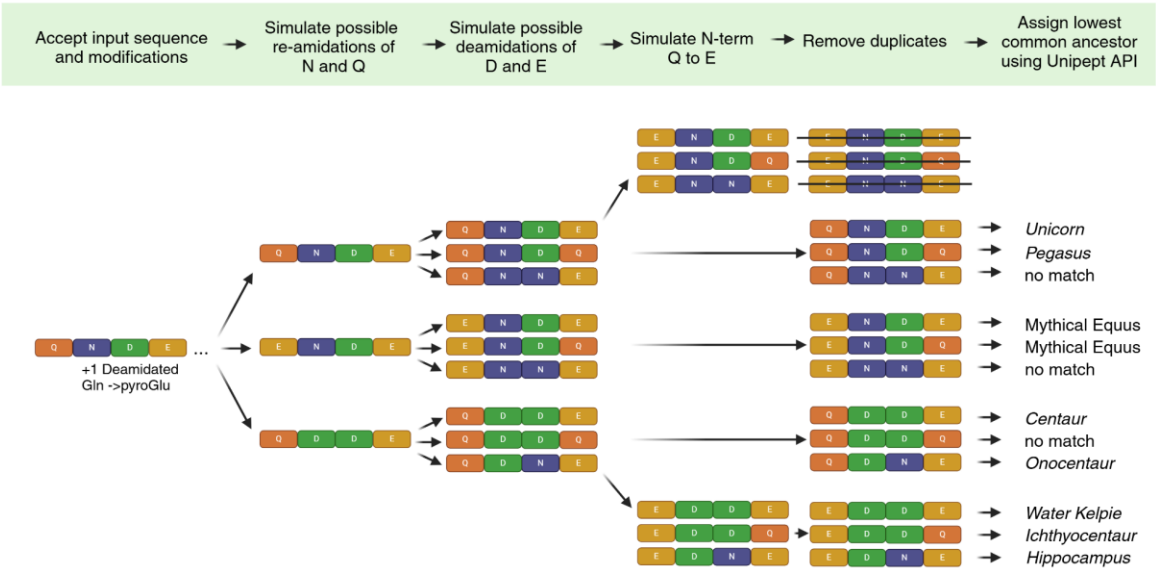

Figure 3: The De-modifier's process of making sequence permutations for an imaginary peptide. First, re-amidations are simulated, followed by deamidations, and then Q to E or E to Q substitutions (due to pyroGlu formation). Peptide permutation duplicates are removed before LCA is assigned for each peptide permutation using Unipept pept2LCA. Figure created in Biorender.com

Test dataset

A test dataset of peptides and modifications previously detected in archaeological samples was assembled to explore the `De-Modifier's` ability to detect modification-induced sequence variations and their resulting taxonomies. The test dataset consists of all unique peptides reported in ancient protein studies of dental calculus and vessels which are readily accessible (see supplementary Information 1), and which include corresponding peptide modifications on a per-peptide basis. The dataset was filtered to exclude identical peptide and modification combinations (with PTMs other than deamidation and pyroglutamic acid formation excluded), and sequence and modification terminology was converted to Mascot format (Supplementary Information 2). The resulting dataset consisted of 1076 unique peptide and modification combinations drawn from 18 published studies (Bleasdale et al., 2021; Charlton et al., 2019; Chowdhury et al., 2021; Evans et al., 2023; Geber et al., 2019; Hendy et al., 2018a, 2018b; Jeong et al., 2018; Lyu et al., 2024; Scott et al., 2021; Siano et al., 2022; Solazzo et al., 2008; Tan et al., 2024; Ventresca Miller et al., 2023; Warinner et al., 2014; Wilkin et al., 2024, 2023, 2021) and my PhD thesis (Evans 2024).

Results and Discussion

How common were modification induced sequence permutations?

Modification-induced sequence permutations were identified for 80.9% of input peptides, with 870 of the 1076 input peptides generating at least one MISP, meaning the remaining 206 input peptides contained no unmodified D/E, deamidated N/Q or pyro-Glu residues (Figure 3). A total of 14675 MISPs were generated from the 1076 input peptides. The median number of MISPs generated for an input peptide was 4, with a mode of 2 and mean of 13.5, although the mean was substantially increased by one input beta casein peptide, FQSEEQQQTEDELQDKIHPFAQTQ + 3 deamidated (NQ), which yielded 4096 MISPs (only one of which matched to a real protein sequence).

How frequently did modification induced sequence permutations result in multiple taxonomic matches?

The results revealed that 16.6% of peptides (179 of 1076) generated at least one MISP which resulted in a taxonomic identification different from that of the input peptide (Figure 4A). Only 18 of these have been reported previously, meaning that modification induced alternate peptide taxonomy is 15% more common than currently reported. While some peptides had multiple MISP taxonomies in addition to their input peptide taxonomy, most only had one additional permutation (Figure 4B). The vast majority of MISPs yielding LCAs resulted from deamidation (220 of 231), while a small minority was caused by a pyroglutamic acid modification (11 of 231). This could partly be because few studies in the test dataset included the formation of N-terminus pyroglutamic acid as a modification in their search parameters.

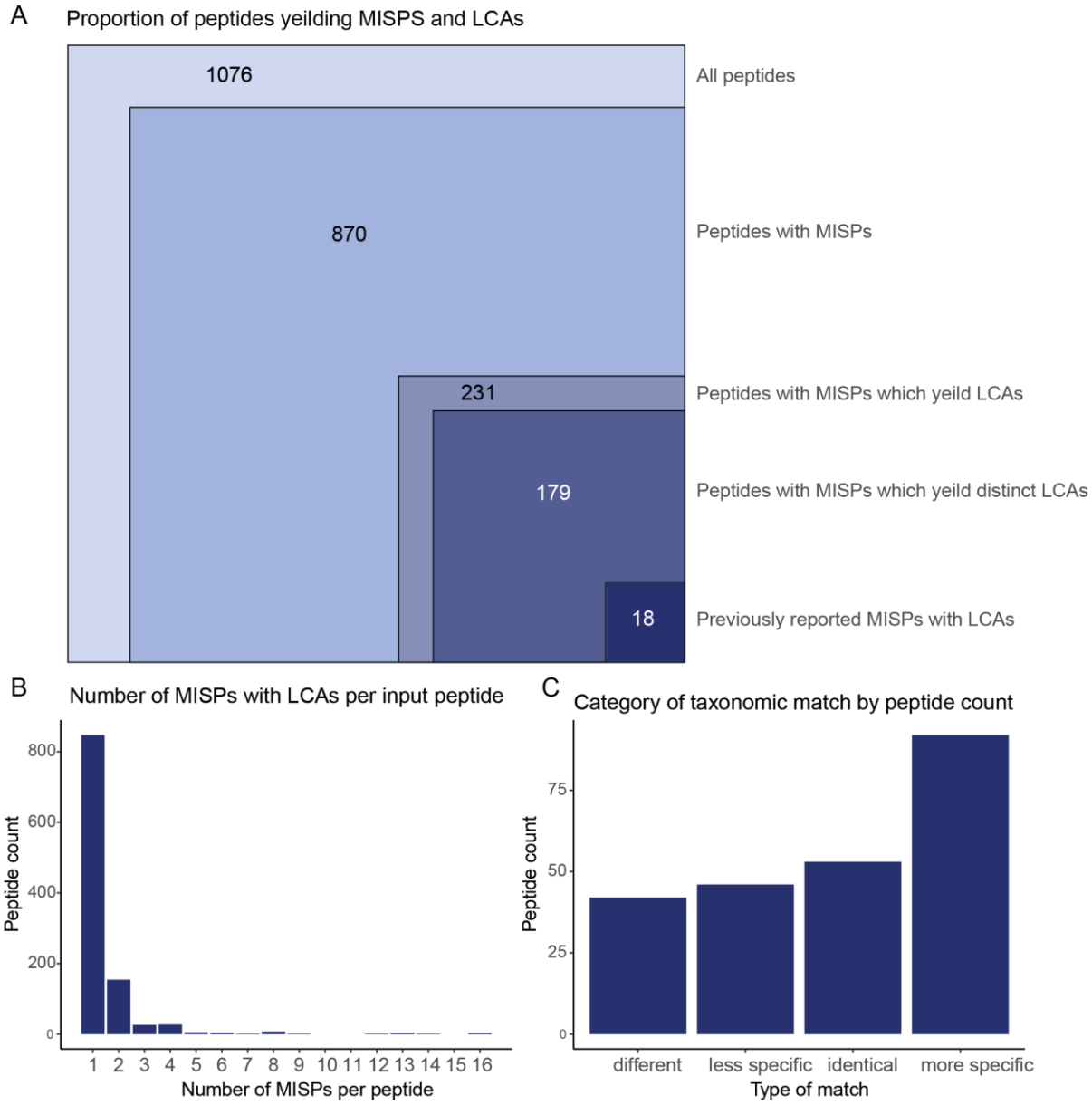

Figure 4: A: Proportional figure showing the total number of peptides in the test dataset, number of peptides with MISPs, number of peptides with MISPs which match to real protein sequences, number of peptides with MISPs that match to real protein sequences, those which additionally match to multiple distinct sequences, and those which have previously been reported, B: Number of Modification-induced sequence permutations with LCAs by peptide count (excluding input peptides which generated no additional LCAs), C: MISP LCA categories by input peptide count including MISP LCAs which were different to the input peptide, identical to the input peptide, less taxonomically specific than the input peptide, or more taxonomically specific than the input peptide.

Archaeological implications of modification-induced alternative peptide taxonomy

When scrutinised, it was apparent that some, but not all, of the 231 input peptides with MISPs which yielded LCAs would lead to inaccurate archaeological interpretations. Excluding the 52 MISPs which rendered identical LCAs to the input sequence (Supplementary Information 3). The remaining 179 peptides with MISPs could be classified into three categories (Figure 4C, Supplementary Information 3):

1) Peptides that could equally be derived from two or more totally different taxonomies
2) Peptides that could be more taxonomically specific than originally thought
3) Peptides that could be less taxonomically specific than originally thought

Of the input peptides with multiple LCA matches, if they had at least one “less specific” match, they were categorised into the third category, even if they also had “more specific” matches, given the more serious implications of less specific taxonomy on an archaeological ingredient interpretation.

The first category included peptides which could belong to two completely different taxonomic lineages, as is the case for the well-known peptide TPEVD**D/N**EALEK, which could equally be derived from *Ovis* or Bovinae βLG. This class consisted of 42 peptides derived from 19 proteins (Supplementary Information 3). Sometimes, such peptides would be easily detectable because they initially appear to belong to an archaeologically unlikely taxonomy, such as the haemoglobin peptide VDQVGAEALGR (deamidated) detected by Evans (2024), which initially matched to *Triturus cristatus* (Northern crested newt), but has a MISP which matches to Laurasiatheria (a superorder of placental mammals which includes all true insectivores, bats, carnivorans, pangolins, even-toed ungulates and odd-toed ungulates). However, other examples could easily be missed because their input peptide LCAs seem perfectly reasonable. For example, the β-lactoglobulin peptide IPAVFKIDALNEN + 2 Deamidated (NQ) initially appears to match Pecora (an infra-order of even-toed hoofed ruminants), but also has a MISP which matches to *Sus scrofa*. This class of peptides can result in potentially spurious taxonomic detections, which could impact archaeological interpretation. Ancient samples often yield only a few peptides of a particular protein, of which fewer still are likely to be taxonomically specific. In such cases, unscreened MISPs could alter archaeological interpretations by leading the researcher to believe that a

potentially spurious taxon was utilised on their site, causing them to miss evidence for a taxon which was actually present.

The second class of MISPs includes peptides that could potentially be more taxonomically specific than indicated by the input peptide LCA (Supplementary Information 3). This class comprised 91 peptides derived from 56 proteins. For instance, the α-lactalbumin peptide NNGKTEYGLFQINNK + 3 Deamidated (NQ) initially appears taxonomically conserved, matching to Laurasiatheria, however `The De-Modifier` revealed a MISP belonging specifically to *Equus caballus* (Horse), a species for which the archaeological sample bore other peptide matches (Ventresca Miller et al., 2023). Another example is the βLG region VYVEELKPTPE (identified in seven of the test dataset studies), which initially has an LCA of Pecora, but in fact generates several MISPs matching to more specific taxonomies (*bos taurus, Moschus moschiferus* and *Odocoileus virginianus texanus).* Peptides from plant proteins were also identified in this category, including the 12S seed storage globulin 1 peptide ALPVDVLANAYR (unmodified) detected by Hendy et al. (2018b), which initially matches to Magnoliopsida (a large class of flowering plants) but has a MISP matching specifically to *Avena sativa* (Oat) (although at the time of their study, both MISPs would have matched to *Avena* due to subsequent database additions). Another example is the Gamma-hordein-1 peptide VMQQQCCLQLAQIPEQYK + 2 Deamidated (NQ), detected by Hendy et al. (2018a), which initially matches to Triticeae (a tribe of grassess which includes many domesticated grains including wheat, barley and rye) but has a MISP matching specifically to *Hordeum vulgare subsp. vulgare* (domesticated barley)- a taxon for which there was additional proteomic evidence in this sample.

This class of MISPs may not change archaeological interpretation as the researcher can't exclude the possibility that the peptide only bears evidence for the higher taxonomy. However such findings may offer additional insight when viewed in the context of other peptide detections. For example, the alpha S1 casein peptide region IGVNQELAYFYP**E**LFR (unmodified), detected by Tanasi et al. (2021) and Siano et al. (2022) in Maltese and Italian contexts respectively, initially appears to belong to Bovinae but has a MISP (IGVNQELAYFYP**Q**LFR) which belongs specifically to *Bubalus Bubalis* (water buffalo). Similarly, the frequently detected βLG peptide VLVLDTDYKK (unmodified), which was reported in 13 of the test-dataset studies, initially appears to belong to Pecora, yet also has MISP belonging to *Bubalus Bubalis*. Water buffalo are a major dairy animal, meaning this finding would be important in archaeological contexts where they are present, especially considering that water buffalo βLG only has one species-specific tryptic peptide, and it is not in a sequence region that has frequently been detected.

Conversely, the third class of MISP taxonomies are those that are less taxonomically specific than the input peptide taxonomy. In the test dataset, this class consisted of 46 peptides derived from 31 different proteins (Supplementary Information 3). These examples have the potential to change archaeological interpretations by reducing the taxonomic specificity of detected proteins, meaning that archaeological evidence for more specific (but potentially inaccurate) taxa is removed (although I note that in the test data set, the samples often luckily had additional taxonomically-specific peptides, meaning the protein taxonomy would not have changed). For example, while it initially appears that the α-amylase inhibitor peptide SGPWSWCNPATGYK + Deamidated (NQ) detected by Tanasi et al. (2021) matches *Triticum aestivum* (Common wheat), `The De-Modifier` revealed a MISP with an LCA of Triticeae (a tribe of grasses which includes many domesticated grains such as wheat, barley and rye). Similarly, the β casein peptide GVAGEPGRNGLPGGPGLR + Deamidated (NQ) detected by Siano et al. (2022) initially appears to match *Bos taurus* (cattle) but in fact has a MISP with the LCA of Artiodactyla (an order of even-toed ungulates which contains most domestic and several wild species). Another example is the 11S globulin isoform 4 peptide QEEEPYYGR (unmodified) detected by Scott et al. (2021), which initially matches *Sesamum indicum* (Sesame) but has a MISP belonging to Eukaryota. While in each of these studies the presence of additional species-specific peptides in the sample meant that the protein taxonomic assessment would not have been altered, if such peptides had been the lynchpin in a protein's taxonomic assessment, this could have impacted the archaeological interpretation.

As with the first category of matches, occasionally, matches of this class may be readily detectable due to archaeologically unlikely taxonomic matches yielded by the input peptides. For instance, Evans (2024) detected the Haemoglobin subunit beta peptide QEVGGEALGR + Gln→pyro-Glu in a Romano-British vessel that initially appears to belong to *Hippopotamus amphibius* (Hippopotamus), which is highly improbable in this context. Upon further inspection this peptide generated a MISP matching to Sarcopterygii (a clade of vertebrates which contains all tetrapods). As can be seen, this class of MISPs could have serious implications on archaeological interpretation of proteomic results, as they may reveal the initial peptide taxonomy to be inaccurately specific.

This analysis revealed that `The De-modifier` was effective at screening for MISPS in the test dataset, which were shown to be more common than previously reported. The results revealed that modification-induced alternate peptide taxonomies can be pivotal in archaeological analyses, as they can completely change taxonomic detections or provide

more or less specific taxonomic detections than initially thought. I note that the risk of incorrect taxonomic matches due to MISPs was alleviated in many cases in the test dataset due to the presence of other taxonomically specific peptides in the ancient samples in question. However, this will not always be the case, and since taxonomic specificity is one of the significant advantages of metaproteomics over other analytic methods, recognising these alternate taxonomies in archaeological interpretations is crucial in avoiding spurious or overly specific taxonomic detections.

Challenges and Future Directions

The biggest challenge to `The De-Modifier` is that automated LCA assignment using Unipept pept2LCA can only sometimes capture the nuances of manual BLAST searching. As such, I stress that `The De-Modifier` was designed to screen for MISPs that yielded alternate taxonomies but not to replace manual BLAST-based LCA analysis. When a peptide matches to select (but not closely related) taxonomies, the LCA can appear needlessly broad, particularly where peptide matches to archaeologically impossible or unlikely species are present. Therefore, once peptides bearing MISPs have been detected by `The De-Modifier`, users should further assess these taxonomic matches using tools that enable scrutiny of each sequence match (e.g. NCBI protein-protein BLAST GUI). By scrutinising matched protein sequences manually, the user is able to investigate the quality of sequences and their relevance to the context. For example several bacterial sequences in NCBI NR appear to have become contaminated with common food proteins such as grain sequences (Scott et al., 2021), or those used in cell culture including milk proteins (Wilkin et al., 2021) and ovalbumin (Geber et al., 2019). This means that automatic LCAs may be assigned at a much higher taxonomy than is actually accurate. Similarly, sometimes manually scrutinising the BLAST matched sequences allows the user to flag potentially low quality matches, for example, Wilkin et al. (2023) note that certain Bovinae and Caprinae blood peptides additionally match to white tailed deer and elk, but only in predicted protein sequences. Lastly, when addressing certain archaeological questions, the taxonomy to which a sequence does *not* match can sometimes be just as meaningful as the taxonomy to which a sequence does match. For example, in their analysis of dietary proteins present in human dental calculus from British Neolithic individuals, Charlton et al. (2019) discovered that one sample harboured evidence for βLG derived from *Capra sp.*, as well as evidence for Bovidae excluding *capra*, meaning that the individual had consumed milk from multiple distinct taxa. As a result of these challenges, while automating LCA in `The De-Modifier` was necessary to detect permutations with real sequence matches, manually BLAST searching and interrogating taxonomy trees in the web version on NCBI BLAST can yield

valuable information, and despite its time-consuming nature should be seen as the gold standard of LCA attribution in ancient protein studies.

The decision to use Unipept rather than a BLAST-based tool was made due both to recent advances in Unipept, and difficulties and limitations of other tools. Previously Unipept was only capable of generating LCA for tryptic peptides (or those with missed cleavages), meaning that it could not assign LCA for many of the degraded peptides identified in palaeoproteomic studies. However, a recent update allows semi-tryptic and non-tryptic peptides to be matched (Vande Moortele et al., 2024). While many any popular BLAST-based LCA-assignment tools were investigated for use in `The De-Modifier`, none fulfilled the criteria of generating LCA for peptides (rather than whole proteins), on the command line (rather than through a GUI), and which do not require the end user to download the entire NCBI/TrembL database or use paid cloud servers (neither of which are accessible to many users).This meant that it was not possible to integrate these BLAST-based tools. Nevertheless, throughout this study it became apparent that Peptide LCA generated using Unipept does not always match the LCA generated by NCBI BLAST searching. This is partly a function of the different databases the two tools use (NCBI protein-protein BLAST uses the NR database by default, while Unipept uses UniprotKB (Swissprot + TrEMBLE) by default. As such, cross checking is advisable, and a larger conversation about taxonomic assignment in palaeoproteomics is warranted. I also note that the taxonomic results of the case study are not intended to be compared directly with author-attributed LCAs from the original studies. This would be an inaccurate comparison as, with the addition of new protein sequences to the databases over time and the different tools used in taxonomic assessment, LCAs may have changed since studies were published.

`The De-Modifier` currently screens for deamidation and pyroglutamic acid-formation induced sequence alterations, however it is probable that other PTMs could cause similar issues. For example, Engels et al. (2024) very recently created a tool for collagen species classification, which screens for isobaric alanine-to-serine substitutions in addition to deamidation. Future research could expand `The De-Modifier` to detect this and other additional PTMs.

Lastly, the current version of `The De-Modifier` treats all modification sites as equally probable, despite the fact that certain modifications and modification sites may be more likely than others. For instance, deamidation of Q is less common given that it occurs over a much slower timescale than deamidation of N (Robinson et al., 2004), and residues in

flexible regions have been found to be more likely sites of asparagine deamidation (Kossiakoff, 1988). In the case of pyroglutamic acid conversion, the two different formation pathways occur under different conditions, which may render a particular starting amino acid more likely than the other (e.g. if samples have never been heated and are of young thermal age, perhaps dehydration of Glu(E) is less likely than deamidation of Gln(Q)). Therefore, future iterations could introduce some form of probabilistic scoring. However, while some modification scenarios may be more likely than others, given the unknown thermal and enzymatic history of many archaeological samples, I decided to employ a conservative approach, allowing the `De-Modifier` to output all MISPs regardless of probability, the rationale being that if there is any possibility that modification could occur, it should be screened for.

Conclusion

In this paper I present `The De-Modifier`: an open access tool to screen for PTM-induced alternate peptide taxonomies. The results of this study reveal that modification-induced alternate peptide taxonomies are much more common than previously recognised in ancient protein studies. Often these multimorphic peptide sequences are only discovered when unexpected taxonomic hits are unearthed, while “normal” seeming taxonomic BLAST results do not always garner scrutiny. Screening of all sequences which could have MISPs is therefore necessary, regardless of initial input peptide taxonomy. The results revealed that modification-induced alternate peptide taxonomy is 15% more common than reported in ancient protein studies, and can result in peptide permutations with completely different, more specified or less specific taxonomy compared to the input peptide. It is likely that future proteomic analyses will reveal additional peptides with alternate taxonomies, meaning screening for MISPs should be standard practice in protein studies. `The De-Modifier` provides a simple and efficient open access tool for detecting such sequences, thus avoiding spurious or overly specific taxonomic detections which could yield inaccurate archaeological interpretations.

Acknowledgements

I would like to thank Matthew Collins for providing feedback on a version of this manuscript.

Bibliography

Bleasdale, M., Richter, K.K., Janzen, A., Brown, S., Scott, A., Zech, J., Wilkin, S., Wang, K., Schiffels, S., Desideri, J., Besse, M., Reinold, J., Saad, M., Babiker, H., Power, R.C., Ndiema, E., Ogola, C., Manthi, F.K., Zahir, M., Petraglia, M., Trachsel, C., Nanni, P., Grossmann, J., Hendy, J., Crowther, A., Roberts, P., Goldstein, S.T., Boivin, N., 2021. Ancient proteins provide evidence of dairy consumption in eastern Africa. Nat. Commun. 12, 1–11. https://doi.org/10.1038/s41467-020-20682-3

Charlton, S., Ramsøe, A., Collins, M., Craig, O.E., Fischer, R., Alexander, M., Speller, C.F., 2019. New insights into Neolithic milk consumption through proteomic analysis of dental calculus. Archaeol. Anthropol. Sci. 11, 6183–6196. https://doi.org/10.1007/s12520-019-00911-7

Chowdhury, M.P., Campbell, S., Buckley, M., 2021. Proteomic analysis of archaeological ceramics from Tell Khaiber, southern Iraq. J. Archaeol. Sci. 132, 105414. https://doi.org/10.1016/j.jas.2021.105414

Colonese, A.C., Hendy, J., Lucquin, A., Speller, C.F., Collins, M.J., Carrer, F., Gubler, R., Kühn, M., Fischer, R., Craig, O.E., 2017. New criteria for the molecular identification of cereal grains associated with archaeological artefacts. Sci. Rep. 7, 6633. https://doi.org/10.1038/s41598-017-06390-x

Engels, I., Burnett, A.C., Dhaenens, M., Daled, S., Deforce, D., De Groote, I., Robert, P., Pironneau, C., Bouwmeester, R., Fischer, V., Abrams, G., Otte, M., Straus, L.G., Di Modica, K., Bray, F., Van der Plaatsen, P., 2024. ClassiCOL: LC-MS/MS analysis for ancient species Classification via Collagen peptide ambiguation. bioRxiv. https://doi.org/10.1101/2024.10.01.616034

Evans, M., Lundy, J., Lucquin, A., Hagan, R., Kowalski, Ł., Wilczyńki, J., Bickle, P., Adamczak, K., Craig, O.E., Robson, H.K., Hendy, J., 2023. Detection of dairy products from multiple taxa in Late Neolithic pottery from Poland: an integrated biomolecular approach. R Soc Open Sci 10, 230124. https://doi.org/10.1098/rsos.230124

Evans, M (2024, June 28). The proteomics of pottery and its residues: applications to ancient foodways (PhD). University of Cambridge

Geber, J., Tromp, M., Scott, A., Bouwman, A., Nanni, P., Grossmann, J., Hendy, J., Warinner, C., 2019. Relief food subsistence revealed by microparticle and proteomic analyses of dental calculus from victims of the Great Irish Famine. Proc. Natl. Acad. Sci. U. S. A. 116, 19380–19385. https://doi.org/10.1073/pnas.1908839116

Hao, P., Ren, Y., Alpert, A.J., Sze, S.K., 2011. Detection, evaluation and minimization of nonenzymatic deamidation in proteomic sample preparation. Mol. Cell. Proteomics 10,

O111.009381. https://doi.org/10.1074/mcp.O111.009381

Hendy, J., Colonese, A.C., Franz, I., Fernandes, R., Fischer, R., Orton, D., Lucquin, A., Spindler, L., Anvari, J., Stroud, E., Biehl, P.F., Speller, C., Boivin, N., Mackie, M., Jersie-Christensen, R.R., Olsen, J.V., Collins, M.J., Craig, O.E., Rosenstock, E., 2018a. Ancient proteins from ceramic vessels at Çatalhöyük West reveal the hidden cuisine of early farmers. Nat. Commun. 9, 1–10. https://doi.org/10.1038/s41467-018-06335-6

Hendy, J., Warinner, C., Bouwman, A., Collins, M.J., Fiddyment, S., Fischer, R., Hagan, R., Hofman, C.A., Holst, M., Chaves, E., Klaus, L., Larson, G., Mackie, M., McGrath, K., Mundorff, A.Z., Radini, A., Rao, H., Trachsel, C., Velsko, I.M., Speller, C.F., 2018b. Proteomic evidence of dietary sources in ancient dental calculus. Proc. Biol. Sci. 285. https://doi.org/10.1098/rspb.2018.0977

Hendy, J., Welker, F., Demarchi, B., Speller, C., Warinner, C., Collins, M.J., 2018c. A guide to ancient protein studies. Nat Ecol Evol 2, 791–799. https://doi.org/10.1038/s41559-018-0510-x

Hill, R.C., Wither, M.J., Nemkov, T., Barrett, A., D'Alessandro, A., Dzieciatkowska, M., Hansen, K.C., 2015. Preserved proteins from extinct bison latifrons identified by tandem mass spectrometry; Hydroxylysine glycosides are a common feature of ancient collagen. Mol. Cell. Proteomics 14, 1946–1958. https://doi.org/10.1074/mcp.m114.047787

Jeong, C., Wilkin, S., Amgalantugs, T., Bouwman, A.S., Taylor, W.T.T., Hagan, R.W., Bromage, S., Tsolmon, S., Trachsel, C., Grossmann, J., Littleton, J., Makarewicz, C.A., Krigbaum, J., Burri, M., Scott, A., Davaasambuu, G., Wright, J., Irmer, F., Myagmar, E., Boivin, N., Robbeets, M., Rühli, F.J., Krause, J., Frohlich, B., Hendy, J., Warinner, C., 2018. Bronze Age population dynamics and the rise of dairy pastoralism on the eastern Eurasian steppe. Proceedings of the National Academy of Sciences 115, E11248–E11255. https://doi.org/10.1073/pnas.1813608115

Jersie-Christensen, R.R., Lanigan, L.T., Lyon, D., Mackie, M., Belstrøm, D., Kelstrup, C.D., Fotakis, A.K., Willerslev, E., Lynnerup, N., Jensen, L.J., Cappellini, E., Olsen, J.V., 2018. Quantitative metaproteomics of medieval dental calculus reveals individual oral health status. Nat. Commun. 9, 4744. https://doi.org/10.1038/s41467-018-07148-3

Kossiakoff, A.A., 1988. Tertiary structure is a principal determinant to protein deamidation. Science 240, 191–194. https://doi.org/10.1126/science.3353715

Kumar, A., Bachhawat, A.K., 2012. Pyroglutamic acid: throwing light on a lightly studied metabolite. Curr. Sci. 102, 288–297.

Lyu, N., Yan, L., Wang, T., Lin, L., Rao, H., Yang, Y., 2024. Characterization of pottery foodcrusts through lipid and proteomic analyses: A case study from the Xiawan site in Yixing city, East China. J. Archaeol. Sci. 161, 105902.

https://doi.org/10.1016/j.jas.2023.105902
Mackie, M., Rüther, P., Samodova, D., Di Gianvincenzo, F., Granzotto, C., Lyon, D., Peggie, D.A., Howard, H., Harrison, L., Jensen, L.J., Olsen, J.V., Cappellini, E., 2018. Palaeoproteomic Profiling of Conservation Layers on a 14th Century Italian Wall Painting. Angew. Chem. Int. Ed Engl. 57, 7369–7374. https://doi.org/10.1002/anie.201713020
Mesuere, B., Willems, T., Van der Jeugt, F., Devreese, B., Vandamme, P., Dawyndt, P., 2016. Unipept web services for metaproteomics analysis. Bioinformatics 32, 1746–1748. https://doi.org/10.1093/bioinformatics/btw039
Nair, B., Palomo, I.R., Markussen, B., Wiuf, C., Fiddyment, S., Collins, M.J., 2023. Parchment Glutamine Index (PQI): A novel method to estimate glutamine deamidation levels in parchment collagen obtained from low-quality MALDI-TOF data. Peer Community J. 3. https://doi.org/10.24072/pcjournal.230
Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P.L.F., Fumagalli, M., Vilstrup, J.T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C.D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A.M.V., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G.D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J.F., Weinstock, J., Gregersen, K., Røed, K.H., Eisenmann, V., Rubin, C.J., Miller, D.C., Antczak, D.F., Bertelsen, M.F., Brunak, S., Al-Rasheid, K.A.S., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M.T.P., Kjær, K., Sicheritz-Ponten, T., Jensen, L.J., Olsen, J.V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., Willerslev, E., 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499, 74–78. https://doi.org/10.1038/nature12323
Ramsøe, A., van Heekeren, V., Ponce, P., Fischer, R., Barnes, I., Speller, C., Collins, M.J., 2020. DeamiDATE 1.0: Site-specific deamidation as a tool to assess authenticity of members of ancient proteomes. J. Archaeol. Sci. 115, 105080. https://doi.org/10.1016/j.jas.2020.105080
Robinson, N.E., Robinson, Z.W., Robinson, B.R., Robinson, A.L., Robinson, J.A., Robinson, M.L., Robinson, A.B., 2004. Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. J. Pept. Res. 63, 426–436. https://doi.org/10.1111/j.1399-3011.2004.00151.x
Schroeter, E.R., Cleland, T.P., 2016. Glutamine deamidation: an indicator of antiquity, or preservational quality? Rapid Commun. Mass Spectrom. 30, 251–255. https://doi.org/10.1002/rcm.7445
Scott, A., Power, R.C., Altmann-Wendling, V., Artzy, M., Martin, M.A.S., Eisenmann, S., Hagan, R., Salazar-García, D.C., Salmon, Y., Yegorov, D., Milevski, I., Finkelstein, I.,

Stockhammer, P.W., Warinner, C., 2021. Exotic foods reveal contact between South Asia and the Near East during the second millennium BCE. Proc. Natl. Acad. Sci. U. S. A. 118. https://doi.org/10.1073/pnas.2014956117

Siano, F., Picariello, G., Caruso, T., Esposito, S., Rescigno, C., Addeo, F., Vasca, E., 2022. Proteomics and Integrated Techniques to Characterize Organic Residues in Funerary Findings from Italic Populations of the First Millennium BC. J. Proteome Res. 21, 1330–1339. https://doi.org/10.1021/acs.jproteome.2c00093

Solazzo, C., Fitzhugh, W.W., Rolando, C., Tokarski, C., 2008. Identification of Protein Remains in Archaeological Potsherds by Proteomics. Anal. Chem. 80, 4590–4597. https://doi.org/10.1021/ac800515v

Stratton, L.P., Kelly, R.M., Rowe, J., Shively, J.E., Smith, D.D., Carpenter, J.F., Manning, M.C., 2001. Controlling deamidation rates in a model peptide: effects of temperature, peptide concentration, and additives. J. Pharm. Sci. 90, 2141–2148. https://doi.org/10.1002/jps.1165

Tanasi, D., Cucina, A., Cunsolo, V., Saletti, R., Di Francesco, A., Greco, E., Foti, S., 2021. Paleoproteomic profiling of organic residues on prehistoric pottery from Malta. Amino Acids 53, 295–312. https://doi.org/10.1007/s00726-021-02946-4

Tan, L., Fan, X., Rao, H., Zhang, H., Yang, Y., 2024. The scientific analysis of the bronze mous excavated from Wushan, Chongqing, China: new perspectives from alloy composition analysis and proteomic method. Heritage Science 12, 88. https://doi.org/10.1186/s40494-024-01203-7

Tsutaya, T., Mackie, M., Koenig, C., Sato, T., Weber, A.W., Kato, H., Olsen, J.V., Cappellini, E., 2019. Palaeoproteomic identification of breast milk protein residues from the archaeological skeletal remains of a neonatal dog. Sci. Rep. 9, 12841. https://doi.org/10.1038/s41598-019-49183-0

Vande Moortele, T., Devlaminck, B., Van de Vyver, S., Van Den Bossche, T., Martens, L., Dawyndt, P., Mesuere, B., Verschaffelt, P., 2024. Unipept 6.0: Expanding metaproteomics analysis with support for missed cleavages, semi-tryptic and non-tryptic peptides. bioRxiv. https://doi.org/10.1101/2024.09.26.615136

van Doorn, N.L., Wilson, J., Hollund, H., Soressi, M., Collins, M.J., 2012. Site-specific deamidation of glutamine: a new marker of bone collagen deterioration: Site-specific deamidation of glutamine. Rapid Commun. Mass Spectrom. 26, 2319–2327. https://doi.org/10.1002/rcm.6351

Ventresca Miller, A.R., Wilkin, S., Bayarsaikhan, J., Ramsøe, A., Clark, J., Byambadorj, B., Vanderwarf, S., Vanwezer, N., Haruda, A., Fernandes, R., Miller, B., Boivin, N., 2023. Permafrost preservation reveals proteomic evidence for yak milk consumption in the 13th century. Commun Biol 6, 351. https://doi.org/10.1038/s42003-023-04723-3

Warinner, C., Hendy, J., Speller, C., Cappellini, E., Fischer, R., Trachsel, C., Arneborg, J., Lynnerup, N., Craig, O.E., Swallow, D.M., Fotakis, A., Christensen, R.J., Olsen, J.V., Liebert, A., Montalva, N., Fiddyment, S., Charlton, S., Mackie, M., Canci, A., Bouwman, A., Rühli, F., Gilbert, M.T.P., Collins, M.J., 2014. Direct evidence of milk consumption from ancient human dental calculus. Sci. Rep. 4, 1–6. https://doi.org/10.1038/srep07104

Welker, F., Collins, M.J., Thomas, J.A., Wadsley, M., Brace, S., Cappellini, E., Turvey, S.T., Reguero, M., Gelfo, J.N., Kramarz, A., Burger, J., Thomas-Oates, J., Ashford, D.A., Ashton, P.D., Rowsell, K., Porter, D.M., Kessler, B., Fischer, R., Baessmann, C., Kaspar, S., Olsen, J.V., Kiley, P., Elliott, J.A., Kelstrup, C.D., Mullin, V., Hofreiter, M., Willerslev, E., Hublin, J.-J., Orlando, L., Barnes, I., MacPhee, R.D.E., 2015. Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. Nature 522, 81–84. https://doi.org/10.1038/nature14249

Welker, F., Soressi, M.A., Roussel, M., van Riemsdijk, I., Hublin, J.-J., Collins, M.J., 2017. Variations in glutamine deamidation for a Châtelperronian bone assemblage as measured by peptide mass fingerprinting of collagen. Sci. Technol. Archaeol. Res. 3, 15–27. https://doi.org/10.1080/20548923.2016.1258825

Wilkin, S., Bayarsaikhan, J., Ganbold, A., Batsuuri, A., Ishtseren, L., Nakamura, D., Eregzen, G., Ventresca-Miller, A., Miller, B.K., 2024. Cauldrons of Bronze Age nomads reveals 2700 year old yak milk and the deep antiquity of food preparation techniques. Sci. Rep. 14, 11625. https://doi.org/10.1038/s41598-024-60607-4

Wilkin, S., Hommel, P., Ventresca Miller, A., Boivin, N., Pedergnana, A., Shishlina, N., Trifonov, V., 2023. Curated cauldrons: Preserved proteins from early copper-alloy vessels illuminate feasting practices in the Caucasian steppe. iScience 26, 107482. https://doi.org/10.1016/j.isci.2023.107482

Wilkin, S., Ventresca Miller, A., Fernandes, R., Spengler, R., Taylor, W.T.-T., Brown, D.R., Reich, D., Kennett, D.J., Culleton, B.J., Kunz, L., Fortes, C., Kitova, A., Kuznetsov, P., Epimakhov, A., Zaibert, V.F., Outram, A.K., Kitov, E., Khokhlov, A., Anthony, D., Boivin, N., 2021. Dairying enabled Early Bronze Age Yamnaya steppe expansions. Nature 598, 629–633. https://doi.org/10.1038/s41586-021-03798-4

Wilkin, S., Ventresca Miller, A., Taylor, W.T.T., Miller, B.K., Hagan, R.W., Bleasdale, M., Scott, A., Gankhuyg, S., Ramsøe, A., Uliziibayar, S., Trachsel, C., Nanni, P., Grossmann, J., Orlando, L., Horton, M., Stockhammer, P.W., Myagmar, E., Boivin, N., Warinner, C., Hendy, J., 2020. Dairy pastoralism sustained eastern Eurasian steppe populations for 5,000 years. Nat Ecol Evol 4, 346–355. https://doi.org/10.1038/s41559-020-1120-y

Wilson, J., van Doorn, N.L., Collins, M.J., 2012. Assessing the extent of bone degradation using glutamine deamidation in collagen. Anal. Chem. 84, 9041–9048.

https://doi.org/10.1021/ac301333t

Xie, M., Schowen, R.L., 1999. Secondary structure and protein deamidation. J. Pharm. Sci. 88, 8–13. https://doi.org/10.1021/js9802493