Unraveling diversity by isolating peptide sequences specific to distinct taxonomic groups

Eleftherios Bochalis[1,2], Michail Patsakis[1,3], Nikol Chantzi[1,3], Ioannis Mouratidis[1,3], Dionysios Chartoumpekis[2], Ilias Georgakopoulos-Soares[1,3,*]

[1] Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA, USA.
[2] Department of Internal Medicine, Division of Endocrinology, Medical School, University of Patras, Patras, Greece.
[3] Huck Institute of the Life Sciences, Pennsylvania State University, University Park, PA, USA
* corresponding authors: izg5139@psu.edu

Abstract
The identification of succinct, universal fingerprints that enable the characterization of individual taxonomies can reveal insights into trait development and can have widespread applications in pathogen diagnostics, human healthcare, ecology and the characterization of biomes. Here, we investigated the existence of peptide k-mer sequences that are exclusively present in a specific taxonomy and absent in every other taxonomic level, termed taxonomic quasi-primes. By analyzing proteomes across 24,073 species, we identified quasi-prime peptides specific to superkingdoms, kingdoms, and phyla, uncovering their taxonomic distributions and functional relevance. These peptides exhibit remarkable sequence uniqueness at six- and seven-amino-acid lengths, offering insights into evolutionary divergence and lineage-specific adaptations. Moreover, we show that human quasi-prime loci are more prone to harboring pathogenic variants, underscoring their functional significance. This study introduces taxonomic quasi-primes and offers insights into their contributions to proteomic diversity, evolutionary pathways, and functional adaptations across the tree of life, while emphasizing their potential impact on human health and disease.

Introduction
The number of available reference proteomes has rapidly increased in recent years, a trend that is expected to continue in the foreseeable future (UniProt Consortium 2023). The availability of a large and diverse set of proteomes of different organisms provides an opportunity to increase our understanding of protein sequence and functional diversity in nature across taxonomic groups. Such research could reveal insights in trait development, through findings pertaining to sequence conservation and divergence mechanisms, the emergence of proteins with new functional roles and can have applications in biomarker discovery, pathogen surveillance and human health among others (Al-Amrani et al. 2021; Lacerda and Reardon 2009). Such advances can be facilitated if the availability of ever-expanding proteomic information is coupled with novel and insightful algorithms to process this abundance of biological information.

Peptide k-mers are defined as oligopeptide sequences of length *k* and are often used in proteomics analyses (Moeckel et al. 2024). The number of possible peptide k-mers exponentiates with k-mer length, leading to oligopeptide sequence uniqueness even at low k-mer lengths (Mouratidis et al. 2024; Georgakopoulos-Soares et al. 2021a). Because of their ease of identification, peptide k-mers have been implemented in a number of applications including in

mass-spectrometry-based proteomics (Chapman 2013), motif search and evolutionary studies (Wen et al. 2014), for taxonomic classification, antimicrobial resistance and pathogen detection (ValizadehAslani et al. 2020) and for the identification of therapeutic targets (Wu et al. 2019; Hajisharifi et al. 2014) among others.

Nullpeptides are the shortest oligopeptide sequences absent from a proteome (Hampikian and Andersen 2007). We and others have previously provided evidence that there are selection constraints against certain nullpeptides (Georgakopoulos-Soares et al. 2021a; Koulouras and Frith 2021; Navon et al. 2016; Poznański et al. 2018). Nullpeptides have been previously used as cancer killing compounds (Alileche et al. 2012; Alileche and Hampikian 2017; Ali et al. 2024), indicating that they can be used as potential drugs, and a subset of them has been shown to recurrently emerge during cancer development, while neoantigens with nullpeptides have been shown to be more immunogenic (Tsiatsianis et al. 2024). Additionally, peptide primes are the subset of nullpeptide sequences that are absent from every proteome (Hampikian and Andersen 2007). Previous research has provided evidence that peptide primes are immunomodulators and can enhance antigen specific immune responses in vaccine adjuvants (Patel et al. 2012). These studies exemplify the utility of different sets of k-mer peptides across biological problems.

We recently described the concept of quasi-prime peptides, which are the shortest peptides that are unique to a species's proteome and are absent from every other proteome known (Mouratidis et al. 2023). We demonstrated that quasi-prime peptides can be identified across taxonomic groups at six and seven amino acid kmer lengths and characterize the set of proteins that harbor them. Here, we have extended the concepts of quasi-prime peptides to incorporate taxonomic groups, identifying k-mer peptides that are present in one or multiple species of a taxonomic group, but absent from every known proteome outside that group. Using 24,073 reference proteomes, we provide proof of the existence of k-mer peptides with this property at the superkingdom, kingdom and phylum levels **(Figure 1)**. We demonstrate the role of quasi-prime peptides in evolutionary divergence and taxonomic adaptations and provide evidence that human quasi-prime loci are more likely to harbor pathogenic variants, emphasizing their functional importance. Our study introduces taxonomic quasi-primes and provides evidence for their roles in shaping proteomic diversity, evolutionary trajectories, and functional adaptations across the tree of life, while underscoring their potential relevance to human health and disease.

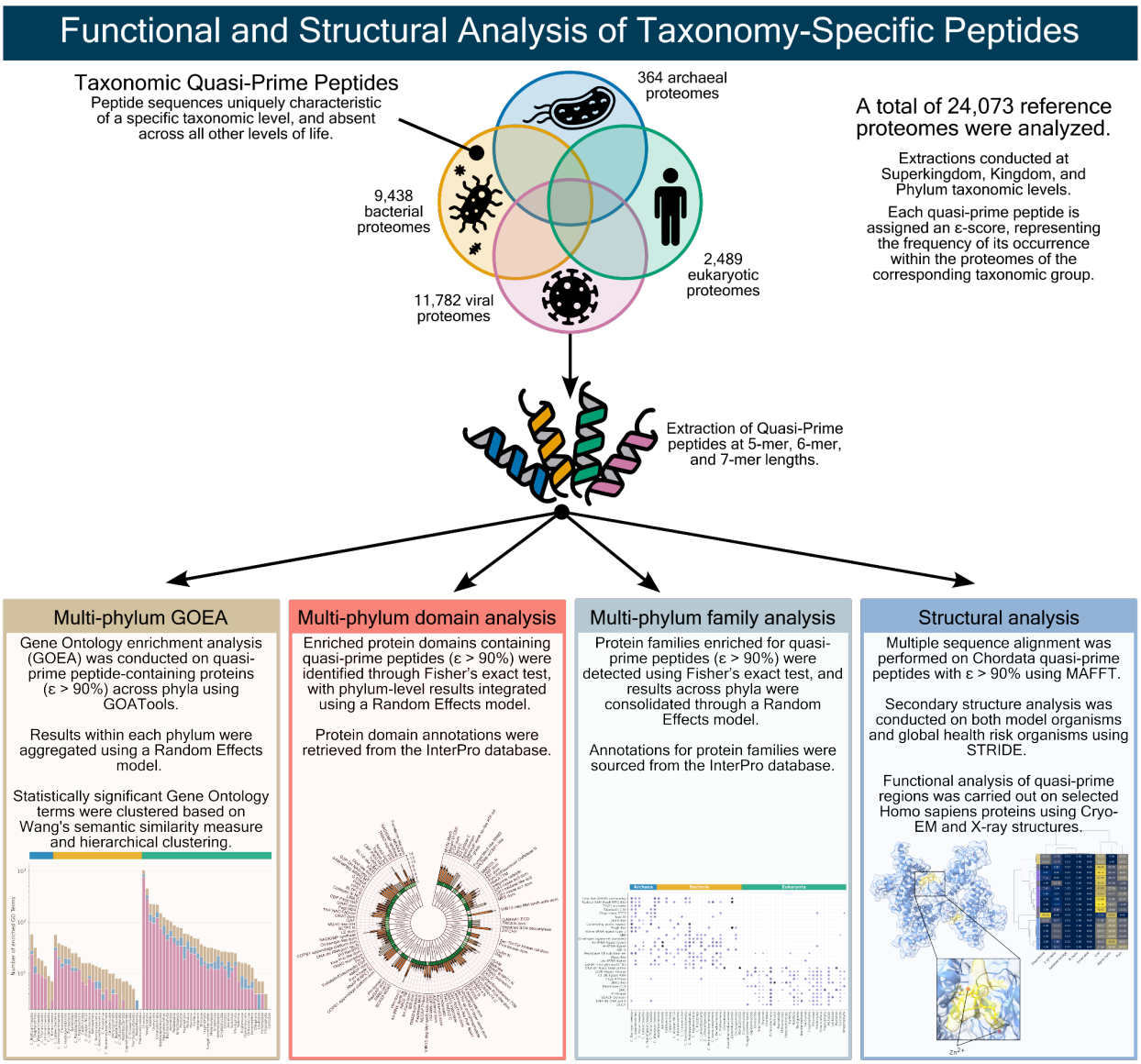


Figure 1: Overview of taxonomic quasi-prime peptide identification and analysis pipeline. Identification of taxonomic quasi-prime peptides was performed for 24,073 reference proteomes at the superkingdom, kingdom and phylum levels. To understand the roles of taxonomic quasi-prime peptides a thorough characterization of was performed, including gene ontology enrichment analyses, protein domain and family analyses and structural examinations.

Materials and methods

Proteomic datasets

Reference proteomes were obtained from UniProt (Release 2024_01), comprising a total of 24,073 species, including 364 archaeal, 9,438 bacterial, 2,489 eukaryotic, and 11,782 viral proteomes.

Peptide k-mer extraction was performed as previously described in (Mouratidis et al. 2023), for k-mer lengths of five to seven amino acids. We defined T as the superset of all considered taxonomies, K as a given k-mer and P as a proteome.

Definition of k-mer. We say that a k-mer K belongs to a taxonomic group T_i if and only if there exists at least one species S in T_i such that K is found in S, that is:

$$K \in T_i \text{ if and only if } \exists S \in T_i : K \in S$$

The set of all taxonomic quasi-primes of group T_i is then defined as:

$$Q(T_i) = \{\, K | K \in T_i \wedge \forall j \neq i : \; K \notin T_j \}$$

Definition of Taxonomy. For species classification, a taxonomy organizes species into hierarchical categories such as superkingdoms, kingdoms, and phyla.

Definition of ε-score:
The ε-score for a k-mer K in a taxonomic group T_i is a measure of the frequency with which K appears in the proteomes of taxonomic group T_i.

$$\varepsilon(K, T_i) = \frac{|\{P \in T_i : K \in P\}|}{T_i}$$

An ε-score equal to zero indicates complete absence of the k-mer K across all species of the taxonomic group T_i, while an ε-score of one hundred indicates universal presence of k-mer K across all member species of the taxonomic group T_i.

Definition of taxonomic quasi-primes peptides
Taxonomic quasi-prime peptides were defined as peptide sequences present in species of a taxonomic group and absent from all species outside that taxonomic group.

$$Q(T_i) = \{\, K | K \in T_i \wedge \forall j \neq i : \; K \notin T_j \}$$

Quasi-prime peptide protein matching
The Peptide Match command line tool (Chen et al. 2013) was used to map quasi-prime peptide sequences to the protein sequences containing them. The Lucene index needed for Peptide Match was created using UniProt Reference Proteome sequences containing SwissProt and TrEMBL entries. After the mapping process, a file containing the quasi-prime peptides of interest, the corresponding UniProt Accession ID, and the range containing the peptide was obtained.

Species clustering based on Phylum ε-scores
Uniform Manifold Approximation and Projection (UMAP; McInnes, Healy, and Melville 2018) was employed using the Python library umap-learn (version 0.5.7) to analyze phylum quasi-prime 7-mers corresponding to the top 50th percentile of ε-scores within each phylum. This approach aimed to visualize how species clustered in two-dimensional space based on their taxonomic

quasi-prime composition. A semi-supervised, density-based UMAP was implemented, incorporating a target weight of 0.25 for clustering based on labels. The algorithm parameters were configured with 30 neighbors and a minimum distance of 0.1 to optimize cluster resolution.

Multi-species Gene Ontology Enrichment Analysis

Multi-species Gene Ontology Enrichment Analysis (MGOEA) was performed on quasi-prime peptide-containing proteins using GOATools v1.4.12 (Klopfenstein et al. 2018) at the Phylum level. A single Gene Ontology Enrichment Analysis (Ashburner et al. 2000) was performed for each Phylum species and results were combined using appropriate statistical methods (See Random effects model). The study population for each analysis consisted of the proteins with quasi-prime peptides, whereas the background population was represented by all the proteins expressed from the corresponding species. The Open Biological and Biomedical Ontologies (OBO) 1.4 file (.obo), containing ontology information and needed for the GOATools package, was obtained from the Gene Ontology Resource (Gene Ontology Consortium et al. 2023) (https://geneontology.org/, release 2024-09-08) and included a total of 40,939 GO terms and 7,894,411 annotations for 5,426 species. Finally, the Gene Ontology Annotation file (GAF) representing the relationship between UniProt Accession and Gene Ontology terms was downloaded from the Gene Ontology Annotation (GOA) Database version 222, released on 05 August, 2024 (https://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz). Pre-MGOEA species were filtered to keep only those that possessed a protein count greater than 10, while also ensuring that protein counts were not propagated during the analysis. Post analysis only Gene Ontology terms with a p-value less than 0.05 were selected.

Gene Ontology result clustering

Post-MGOEA enriched terms with an adjusted p-value less than 0.05 were clustered into broader representative GO terms to minimize redundancy and noise and highlight broader functional themes present across the phyla. Each GO term class (Biological Process, Molecular Function and Cellular Component) within each phylum was handled separately and Wang's semantic similarity measure (Wang et al. 2007) was used to calculate the pairwise similarity of terms. These similarity values were converted to a distance matrix for agglomerative hierarchical clustering with the average linkage method. The optimal clustering threshold (Bettembourg, Diot, and Dameron 2015) specifically for each GO term class was set to 0.54 for biological processes, 0.535 for molecular functions and 0.52 for cellular components. Post-clustering a representative term for each cluster was selected based on the adjusted p-value and ties between terms were resolved using the presence percentage across phylum-specific species and the combined LOR, where terms with the highest value were characterized as representative. For the representative terms a weighted average LOR was calculated using the standard error as weight followed by winsorization at the 95th percentile to minimize the effect of extreme values. GO terms that were not assigned to any cluster were also retained, if the term's adjusted p-value was less than 0.05.

Multi-species Protein Entry Enrichment Analysis

Quasi-prime peptide-containing proteins were subjected to multi-species Protein Entry Enrichment Analysis (MPEA), where the presence of quasi-prime peptides within functional protein domains as well as the protein family composition of these proteins was analyzed. Protein

domain and protein family data were obtained from the InterPro database (version 101.0 updated on 25th July 2024) (Paysan-Lafosse et al. 2023), which contains 14,950 domains and 26,089 family entries. For the enrichment analysis, Fisher's exact test was implemented, from which Odds Ratio (OR) and a p-value was calculated for each entry, which were downstream combined to a final value that represents the effect size across Phylum species. Haldane-Anscombe correction (Agresti 1999) was applied to all cells of the 2x2 contingency table used in Fisher's exact test to account for 0 values, which may lead to infinite estimates. Only entries with a p-value less than 0.05 OR greater than 1 and with presence across more than 1 species were subjected to effect combination, whereas entries present only in one species were retained, if they passed the p-value and OR cut-off.

Enrichment combination across species using Random Effects Meta-Analysis method

A meta-analysis technique employing a random effects model was performed to identify gene ontology terms and functional entries (domains or families) that are present across multiple species of the same Phylum and evaluate the combined enrichment of each of them. Common items were filtered based on the criteria described earlier, and the natural logarithm of the OR value (LOR) was calculated. An original fixed-effect weight was calculated for each study (species), so that studies with more precise estimates (smaller standard error) are given a larger weight in the analysis, as follows:

$$w_i = \frac{1}{{SE_{LOR,i}}^2}$$

where, w_i represents the weight for the i_{th} study, $SE_{LOR,i}$ is the standard error of the LOR value for the ith study and ${SE_{LOR,i}}^2$ is the variance of the LOR estimate for the i_{th} study. The initial combined enrichment was obtained through the computation of the fixed effects weighted mean, which will be later used to assess the heterogeneity between-studies. The calculation of the fixed effects weighted mean goes as follows:

$$\bar{\theta} = \frac{\sum_{i=1}^{k} w_i \cdot LOR_i}{\sum_{i=1}^{k} w_i}$$

where $\bar{\theta}$ represents the weighted mean LOR using the fixed-effects model and k represents the total number of studies. The use of Cochran's Q statistic was implemented to measure the total variability in enrichment values across studies and will be used to to estimate variance due to heterogeneity. This formula was used to calculate the Q-statistic:

$$Q = \sum_{i=1}^{k} w_i \cdot \left(LOR_i - \bar{\theta}\right)^2$$

where Q is Cochran's Q statistic and the term $\left(LOR_i - \bar{\theta}\right)^2$ represents the squared deviation of each study's effect size from the weighted mean. A constant was computed to adjust the variance of the calculated weights and also for the estimation of the between-study variance. The calculation goes as follows:

$$C = \sum_{i=1}^{k} w_i - \left(\frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right)$$

where C is the constant, $\sum_{i=1}^{k} w_i$ is the sum of weights obtained from all studies and $\sum_{i=1}^{k} w_i^2$ is the sum of the squared weights. A fraction adjustment is performed, because it accounts for the variability of the weights. As implied earlier, between study-variance was estimated using the Tau-squared statistic, due to its ability to calculate the amount of variance in enrichment values due to real differences between studies rather than chance and it can ensure that the variance estimate is non-negative. The t^2 statistic is calculated using this formula:

$$t^2 = \frac{max(0,(Q-df))}{C}$$

where t^2 is the heterogeneity variance and df represents the degrees of freedom (k-1) used to assess the statistical significance of the Q statistic against the chi-squared distribution. The final weights used for the combined enrichment values were obtained using the following random effects model:

$$w_i^* = \frac{1}{{SE_{LOR,i}}^2 + t^2}$$

w_i^* is the final adjusted weight for the ith study using the random effects model and ${SE_{LOR,i}}^2$ is the within-study variance of the ith study. The computation of the combined overall enrichment value for each item using the random effects model is represented below:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i^* \cdot LOR_i}{\sum_{i=1}^{k} w_i^*}$$

$\hat{\theta}$ is the combined enrichment value (LOR) using the random effects model, the numerator represents the sum of adjusted LOR values while the denominator represents the sum of the adjusted weights. In all these calculations, we have assumed that taking the natural logarithm of the odds ratio provides a consistent measure of enrichment effect size. This assumption enables us to extend the enrichment analysis across species and calculate the final combined enrichment value for each item. The final step to this random-effects model was to calculate the statistical significance of the combined enrichment value against the null-hypothesis that the combined enrichment value is zero. To achieve this, first we calculated the standard error of the combined enrichment value, then we computed a modified Z-score, and with this Z-score a final p-value was obtained. The formula used for modified Z-score is the following:

$$Z = \frac{\hat{\theta}}{SE_{\hat{\theta}}}$$

where a larger absolute value of Z indicates a more significant deviation from the null-hypothesis and the formula for the p-value calculation:

$$p = 2 \cdot (1 - T(|Z|, df = n - 1))$$

where p represents the calculated p-value and $T(|Z|)$ denotes the cumulative distribution function of the t-distribution with n-1 degrees of freedom. The choice to calculate p-values based on the t-distribution was made to address the small sample size of certain underrepresented phyla. Items were filtered to keep only those with a meta-analysis adjusted p-value less than 0.05. Multiple testing correction was applied using the Benjamini-Hochberg procedure.

Secondary and tertiary structure analysis

Quasi-prime peptides were subjected to structural analysis. Original PDB files of the proteins containing quasi-prime peptides were obtained from the AlphaFold Protein Structure Database (Jumper et al. 2021; Varadi et al. 2024) updated as of September 2024. Quasi-prime peptide regions from each protein were extracted and their secondary structure was identified using the STRIDE (STRuctural IDEntification) algorithm (Frishman and Argos 1995) as provided by the ssbio v0.9.9 tool package. Quasi-prime peptides that presented with no hydrogen-bonds were characterized to have a Disordered conformation. PDB files presenting the interaction between protein and ligands were downloaded from the RCSB Protein Data Bank (Berman et al. 2000) (updated as of 2024, download timestamp: November 21st 2024), containing a total of 225,681 available structures. Only conformations obtained through Cryo-EM or X-Ray were selected. Protein visualization and hydrogen-bond detection between ligand and protein was performed with the use of the UCSF ChimeraX version 1.8 software (Meng et al. 2023).

Protein ortholog mapping and Multiple Sequence Alignment

Protein orthologs for Homo Sapiens proteins across the Chordata phylum were obtained through the EggNOG v5.0.0 (Huerta-Cepas et al. 2019) database, which contains 4.4 million orthologous groups and data for 5,090 Organisms and 2,052 Viruses. The orthologous groups of choice were later subjected to Multiple Sequence Alignment (MSA) using the MAFFT version 7 command line tool (Katoh and Standley 2013) with the options: --localpair --maxiterate 1000 --amino --thread 5. The resulting alignments were trimmed using ClipKIT v.2.3.0 (Steenwyk et al. 2020) with the options -smart-gap to remove poorly aligned protein regions and improve the phylogenetic signal by focusing on well-conserved segments. The trimmed results were visualized using Jalview v2.11.4.1 (Waterhouse et al. 2009). Sequences were ordered based on pairwise similarity, quasi-prime-containing regions were highlighted whereas distant areas that did not contain quasi-prime peptides were hidden from the visualization.

Pathogenicity prediction in taxonomic quasi-prime protein regions.

To predict the pathogenicity of single-nucleotide missense variants, we employed AlphaMissense (J. Cheng et al. 2023), (version v3, updated as of September 19, 2023, https://zenodo.org/records/10813168). The analysis included all possible missense variants (approximately 71 million) derived from 19,000 canonical protein-coding transcripts in the human genome (hg38 build). Our study specifically focused on human proteins containing taxonomic

quasi-prime 7-mers with ε-scores exceeding 90%. We conducted a comparative analysis of the pathogenicity associated with missense mutations located within these taxonomic quasi-prime loci and those occurring outside these loci. Mutations with AlphaMissense scores below 0.1 were classified as likely benign, while those with scores above 0.9 were designated as highly pathogenic.

Results

Identification of sequences that are unique to individual taxa

It remains unknown if peptide k-mer sequences which are unique to a particular taxonomic group play a role in the emergence of novel traits within individual taxa. Here, we investigated the potential presence of taxonomic quasi-prime peptide sequences, which are k-mer peptide sequences that are found in proteomes of a single taxon, and otherwise absent from all other taxa. We performed this investigation at the superkingdom, kingdom and phylum levels for k-mer lengths of five, six and seven amino acids in all organisms with an available reference proteome, totalling 24,073 reference proteomes spanning the tree of life. K-mer lengths below five amino acids were not considered, as tetrapeptides are highly prevalent and in the human proteome all possible tetrapeptides are observed (Georgakopoulos-Soares et al. 2021b; Chantzi et al. 2024). Octapeptides and longer oligopeptides were not considered as the possible proteome space becomes extremely large (20^8), limiting the set of k-mers that are shared between multiple species in a taxonomy.

Derivation of superkingdom- and kingdom- specific quasi-prime peptides

First, we define taxonomic quasi-primes, sequences specific to a taxonomy and define the ε-score, representing the percentage of species of the taxonomy in which a peptide k-mer is found in. (see Methods). Taxonomic quasi-prime five-mers are identifiable at the superkingdom level and are exclusively found within the Eukaryota. We identified 12 distinct five-mers with ε-scores ranging from 0.52% to 3.74%, and a median ε-score (ε_M) of 1.67% **(Supplementary Figure 1)**. The sequences of these 12 peptides are detailed in **Supplementary Table 1**. Regarding taxonomic quasi-prime six-mers, we find that the median percentage of species containing these six amino acid peptides is 0.01% in viruses, 0.02% in bacteria, 0.24% in eukaryotes, and 0.27% in archaea (**Figure 2a**.) Similar patterns are observed for seven-amino acid taxonomic quasi-prime peptides (**Figure 2a**). The larger number of taxonomic quasi-primes identified in Eukaryotes stems from their larger proteome sizes (Spearman's correlation coefficient ρ=0.963, p-value<0.001) **(Supplementary Figure 2)**. Additionally, due to the large k-mer space these findings translate to a considerable number of superkingdom-specific peptides. Specifically, at six and seven amino acids peptide k-mer length, we observe thousands of peptides that are only found in species of a single superkingdom or only in Viruses (**Figure 2a**). For instance, we observe 85,373 and 86,483,511 peptides found uniquely in bacterial proteomes at six and seven amino acids k-mer lengths respectively (**Figure 2a**).

We also find that the six amino acid taxonomic quasi-prime peptides found in the largest number of species (ε_{max}) in each of the three superkingdoms and in viruses were in 0.59%, 10.10%, 17.31% and 92.33% of the species in viruses, bacteria, archaea and eukaryotes respectively. This means that we can identify one six-mer peptide found in 92.33% of eukaryotes and otherwise

absent from all bacteria, archaea and viruses with reference proteomes available. Interestingly, for peptide lengths of seven amino acids we are able to identify ε_{max} of 98.11% for eukaryotic species for the peptide sequence SAPNYCY, which maps to proteins belonging to the serine/threonine phosphatase family, and are highly conserved in eukaryotes (Ohama 2019) (**Figure 2a,c**). The observed patterns underscore the significance of taxonomic quasi-prime peptides in distinguishing between superkingdoms, offering valuable insights into proteomic diversity across different domains of life and holding potential as molecular markers for superkingdom-level classification.

Next, we analyzed the distribution of taxonomic quasi-primes across various organismal and viral kingdoms, observing significant variability within individual groups. This variability was particularly pronounced in viral kingdoms, where ε_M values ranged from 0.02% to 10.00% at the six-mer level and from 0.02% to 5.00% at the seven-mer level in Heunggongvirae and Zilligvirae, respectively (**Figure 2b**). Among eukaryotic kingdoms, the lowest ε_M value was found in Fungi (0.09%), while the highest was observed in Protista (0.55%) **(Figure 2b)**. Further analysis of ε_{max} at the six-mer level within these kingdoms revealed values ranging from 17.03% in Protista to 87.18% in Metazoa. For seven-mers, ε_{max} values reached up to 37.91% in Protista and 98.78% in Metazoa **(Supplementary Table 1)**. We identified the taxonomic quasi-prime seven-mer, CKGFFKR, which represents the ε_{max} within the Metazoan kingdom. This seven-mer was mapped to proteins associated with the nuclear hormone receptor family, whose genes show strong sequence conservation and little evidence for positive selection in Metazoans (Krasowski et al. 2005) (**Figure 2b,d**).

We conclude that we can identify highly superkindom-specific and kingdom-specific peptides shared in a significant proportion of the species of those taxonomic groups. We note, however, that the number of taxonomic quasi-prime peptides we can detect is influenced by the proteome size and the number of species available within each superkingdom and kingdom. As more reference proteomes become available, this representation will become increasingly complete.

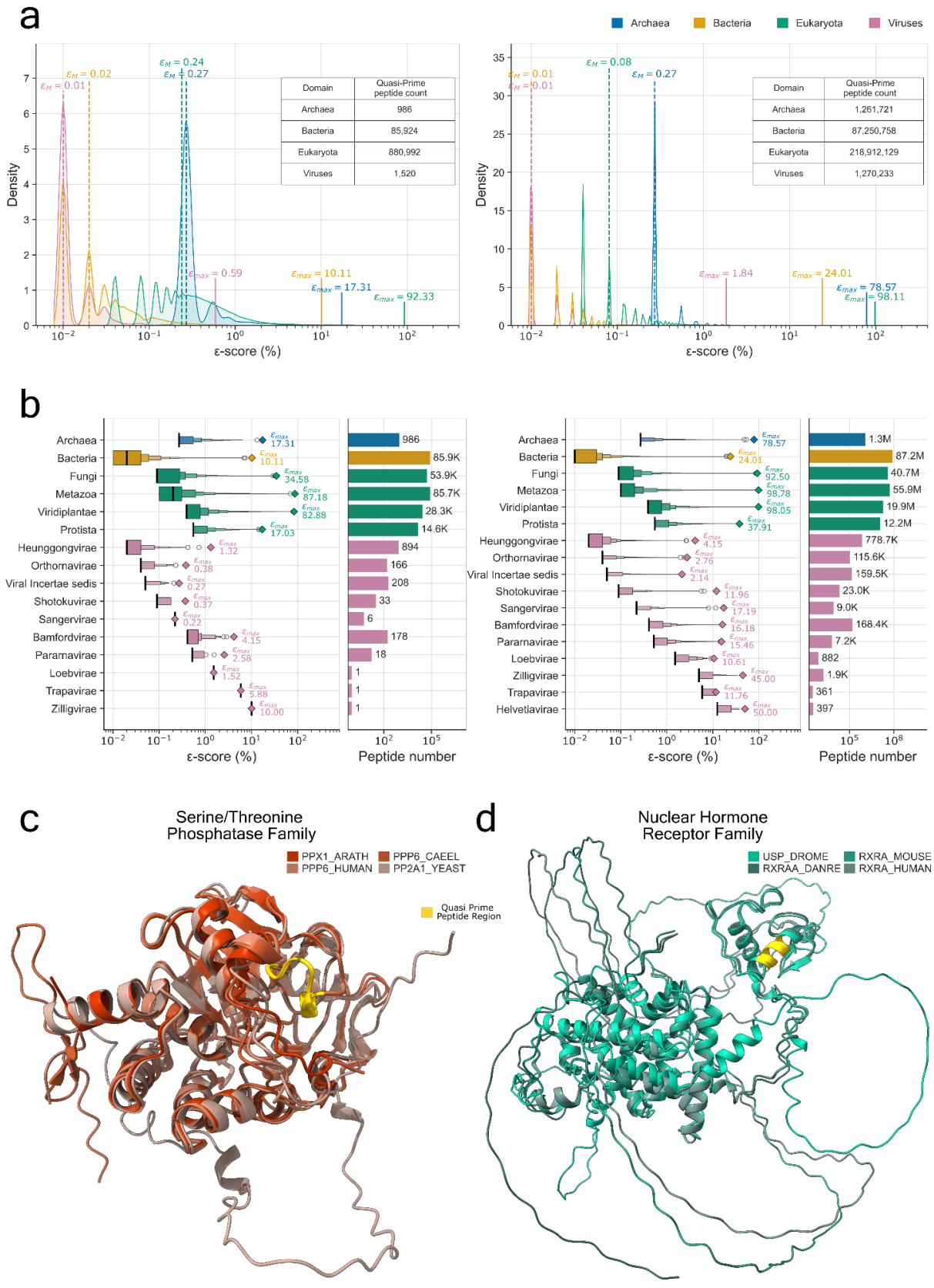
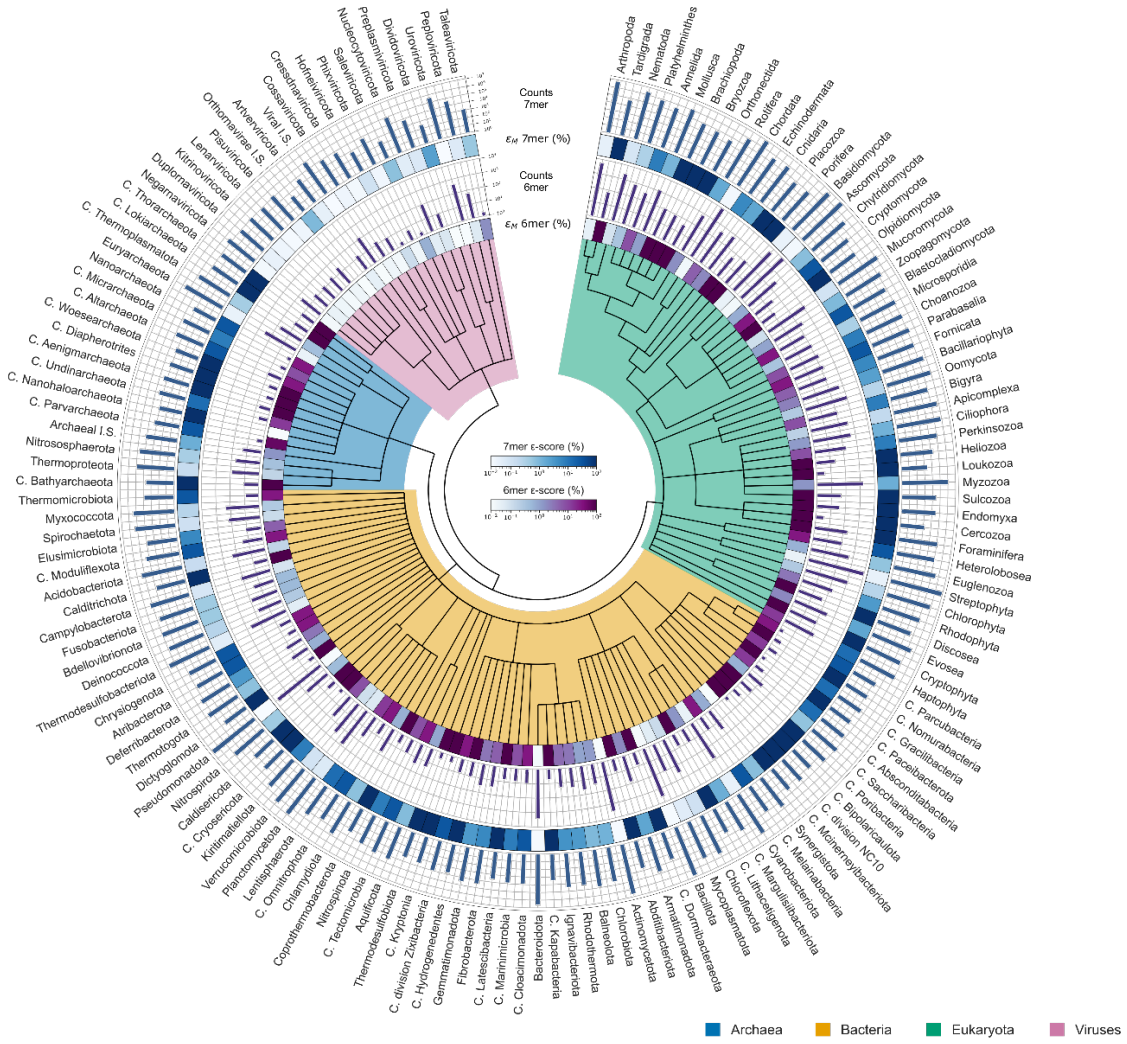

a
Archaea
Bacteria
Eukaryota
Viruses
Density
ε-score (%)
Domain | Quasi-Prime peptide count
Archaea | 986
Bacteria | 85,924
Eukaryota | 880,992
Viruses | 1,520
Domain | Quasi-Prime peptide count
Archaea | 1,261,721
Bacteria | 87,250,758
Eukaryota | 218,912,129
Viruses | 1,270,233
b
Archaea
Bacteria
Fungi
Metazoa
Viridiplantae
Protista
Heunggongvirae
Orthornavirae
Viral Incertae sedis
Shotokuvirae
Sangervirae
Bamfordvirae
Pararnavirae
Loebvirae
Trapavirae
Zilligvirae
Helvetiavirae
ε-score (%)
Peptide number
c
Serine/Threonine Phosphatase Family
PPX1_ARATH
PPP6_CAEEL
PPP6_HUMAN
PP2A1_YEAST
Quasi Prime Peptide Region
d
Nuclear Hormone Receptor Family
USP_DROME
RXRA_MOUSE
RXRAA_DANRE
RXRA_HUMAN

Figure 2: Quasi-prime peptide count and ε-score distribution. a Kernel Density Estimate plots showing quasi-prime peptide ε-score distribution at Superkingdom level. Dotted lines represent the median ε-score (ε_M) and solid lines represent the maximum ε-score (ε_{max}) value. Tables inside each plot display the number of unique quasi-prime peptide counts. **b** Quasi-prime peptide ε-score distribution and peptide count at Kingdom level. Left to right: Letter-value plot of ε-score distribution, where the ε_M value is depicted as a solid black line and the ε_{max} value as a rhombus. Barplot of the unique quasi-prime peptide counts. (Data for quasi-prime peptide 6mers are displayed to the left and quasi-prime peptide 7mers are displayed to the right). **c** Superposition of proteins representing the Serine/Threonine Phosphatase family at the eukaryotic superkingdom. Results are shown for *Arabidopsis thaliana, Caenorhabditis elegans, Homo sapiens*, and *Saccharomyces cerevisiae* orthologs. **d** Superposition of proteins representing the Nuclear Hormone Receptor family at the metazoan kingdom. Results are shown for *Drosophila melanogaster*, *Mus musculus*, *Danio rerio*, and *Homo sapiens* orthologs. Taxonomic quasi-prime peptides are marked in yellow.

Derivation of phylum-specific quasi-prime peptides

We investigated the presence and distribution of taxonomic quasi-primes across various phyla. At the six-amino-acid sequence length, ε_M values displayed substantial variation. In eukaryotic phyla, values ranged from 0.14% in Ascomycota species to 100.00% in Foraminifera and sixteen other eukaryotic phyla. For bacterial phyla, ε_M values spanned from 0.03% in Pseudomonadota to 100.00% in Abditibacteriota and 17 additional bacterial phyla. Archaeal phyla exhibited values between 0.43% in Euryarchaeota and 100.00% in seven candidate archaeal phyla. In viral phyla, ε_M values ranged from 0.02% in Uroviricota to 10.00% in Taleaviricota (**Figure 3-4; Supplementary Figure 3**). A similar distribution pattern was observed for the seven-amino-acid sequence length. In eukaryotic phyla, Ascomycota again exhibited the lowest ε_M value at 0.14%, while Porifera reached the maximum value of 100.00%, shared with 16 other eukaryotic phyla. Among bacterial phyla, ε_M values ranged from 0.03% for Pseudomonadota to 100.00% for Abditibacteriota and 18 additional bacterial phyla. For archaeal phyla, Euryarchaeota showed the lowest ε_M value of 0.43%, while Candidatus Bathyarchaeota and seven other candidate archaeal phyla attained ε_M of 100.00%. Viral phyla exhibited values ranging from 0.02% in Uroviricota to 12.50% in Dividoviricota (**Figure 3-4; Supplementary Figure 3**). We also observe that superkingdoms cluster by taxonomic quasi-primes (**Figure 3b**). The complete distribution of ε_M values alongside ε_{max} for each phylum is available at the **Supplementary Table 3**.

a


Counts 7mer
εM 7mer (%)
Counts 6mer
εM 6mer (%)
7mer ε-score (%)
6mer ε-score (%)
Arthropoda
Tardigrada
Nematoda
Platyhelminthes
Annelida
Mollusca
Brachiopoda
Bryozoa
Orthonectida
Rotifera
Chordata
Echinodermata
Cnidaria
Placozoa
Porifera
Basidiomycota
Ascomycota
Chytridiomycota
Cryptomycota
Olpidiomycota
Mucoromycota
Zoopagomycota
Blastocladiomycota
Microsporidia
Choanozoa
Parabasalia
Fornicata
Bacillariophyta
Oomycota
Bigyra
Apicomplexa
Ciliophora
Perkinsozoa
Heliozoa
Loukozoa
Myzozoa
Sulcozoa
Endomyxa
Cercozoa
Foraminifera
Heterolobosea
Euglenozoa
Streptophyta
Chlorophyta
Rhodophyta
Discosea
Evosea
Cryptophyta
Haptophyta
C. Parcubacteria
C. Nomurabacteria
C. Gracilibacteria
C. Paceibacterota
C. Absconditabacteria
C. Saccharibacteria
C. Poribacteria
C. Bipolaricaulota
C. division NC10
C. Micrarchaeota
C. Synergistota
C. Melainabacteria
C. Cyanobacteriota
C. Margulisbacteriota
C. Lithacetigenota
Chloroflexota
Mycoplasmatota
Bacillota
C. Dormibacteraeota
Armatimonadota
Abditibacteriota
Actinomycetota
Chlorobiota
Balneolota
Rhodothermota
Ignavibacteriota
C. Kapabacteria
Bacteroidota
C. Cloacimonadota
C. Marinimicrobia
C. Latescibacteria
Fibrobacterota
Gemmatimonadota
C. Hydrogenedentes
C. Krystonia
C. division Zixibacteria
Aquificota
Thermodesulfobiota
C. Tectomicrobia
Nitrospinota
Coprothermobacteriota
Chlamydiota
C. Omnitrophota
Lentisphaerota
Planctomycetota
Verrucomicrobiota
Kiritimatiellota
C. Cryosericota
Caldisericota
Nitrospirota
Pseudomonadota
Dictyoglomota
Thermotogota
Deferribacterota
Aquificota
Chrysiogenota
Thermodesulfobacteriota
Deinococcota
Bdellovibrionota
Fusobacteriota
Campylobacterota
Calditrichota
Acidobacteriota
C. Moduliflexota
Elusimicrobiota
Spirochaetota
Myxococcota
Thermomicrobiota
C. Bathyarchaeota
Thermoproteota
Nitrososphaerota
Archaeal I.S.
C. Parvarchaeota
C. Nanohaloarchaeota
C. Undinarchaeota
C. Aenigmarchaeota
C. Diapherotrites
C. Woesearchaeota
C. Altarchaeota
C. Micrarchaeota
Nanoarchaeota
Euryarchaeota
C. Thermoplasmatota
C. Lokiarchaeota
C. Thorarchaeota
Negamaviricota
Duplornaviricota
Kitrinoviricota
Lenarviricota
Pisuviricota
Orthornavirae I.S.
Artverviricota
Viral I.S.
Cressdnaviricota
Hofneiviricota
Phixviricota
Saleviricota
Nucleocytoviricota
Preplasmiviricota
Dividoviricota
Uroviricota
Peploviricota
Takaviricota
Archaea
Bacteria
Eukaryota
Viruses

b

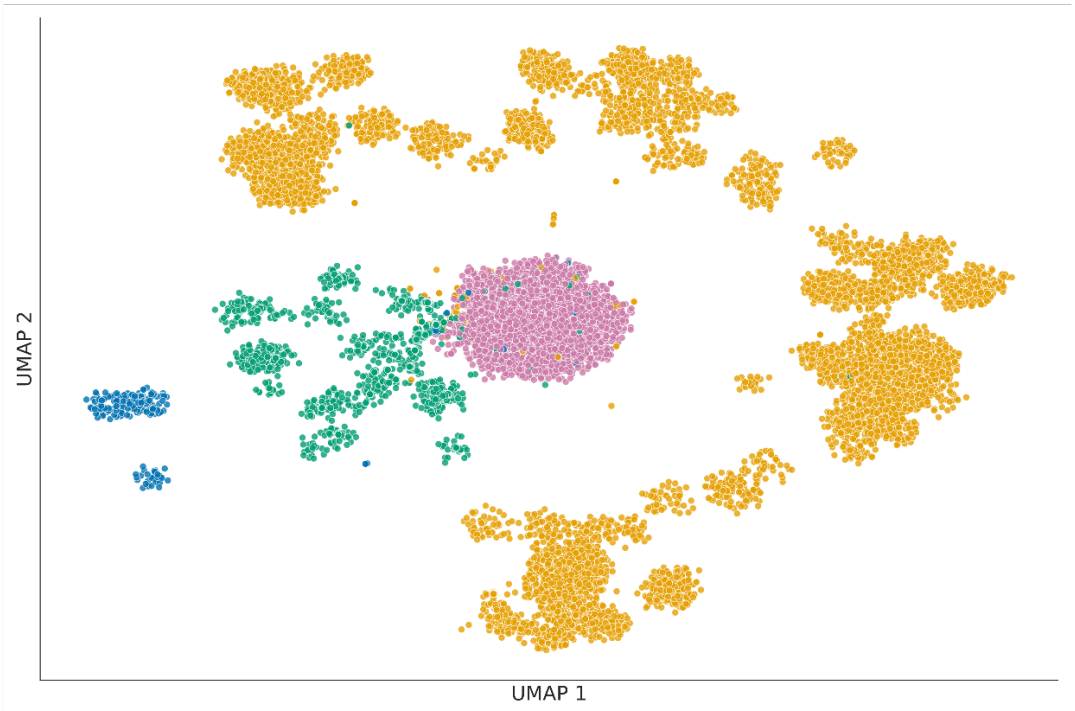

UMAP 2
UMAP 1

Figure 3: Quasi-prime peptide statistics at Phylum level. a Circos plot depicting the phylogenetic organization of Superkingdoms, Kingdoms and Phyla at the Archaeal, Bacterial, Eukaryotic and Viral levels (center). The number of unique quasi-prime peptides, alongside the median ε-score value (ε_M) at 6mer and 7mer length is also shown. Inner to outer: ε_M for 6mer quasi-prime peptides; total 6mer quasi-prime peptide counts; ε_M for 7mer quasi-prime peptides; total 7mer quasi-prime peptide counts. **b** UMAP plot depicting the clustering of reference proteomes based on their taxonomic quasi-prime 7mers. Only the top 50th percentile of quasi-primes based on their ε-score was used for the clustering.

Across all three cellular superkingdoms, several phyla exhibited an ε_{max} of 100.00%, whereas for viral phyla the ε_{max} was capped at 50%. This is likely the result of viruses' rapid evolution in response to host immune pressures resulting in viral phyla encompassing greater genetic variation than cellular phyla. For both Streptophyta and Nematoda, we analyzed the proteins from which these highly phylum-specific peptide seven-mers originated. In Streptophyta, three taxonomic quasi-primes were identified with an ε_{max} of 99.11%. These sequences, TPWPGNN, EHFCIHA, and THHEYIQ, were found in *Arabidopsis thaliana* within the proteins Cellulose Synthase A Catalytic Subunit 4 (UDP-forming), 3-Ketoacyl-CoA Synthase 1, and Callose Synthase 5, respectively (**Figure 4c-e**). In Nematoda, a single peptide, ICPKYCA, was identified with an ε_{max} of 97.65%. This peptide is located in the cuticle collagen DPY-13 protein of *C. elegans* (**Figure 4f**), which is crucial for cuticle formation, serving both as an exoskeleton and a protective barrier against environmental challenges.

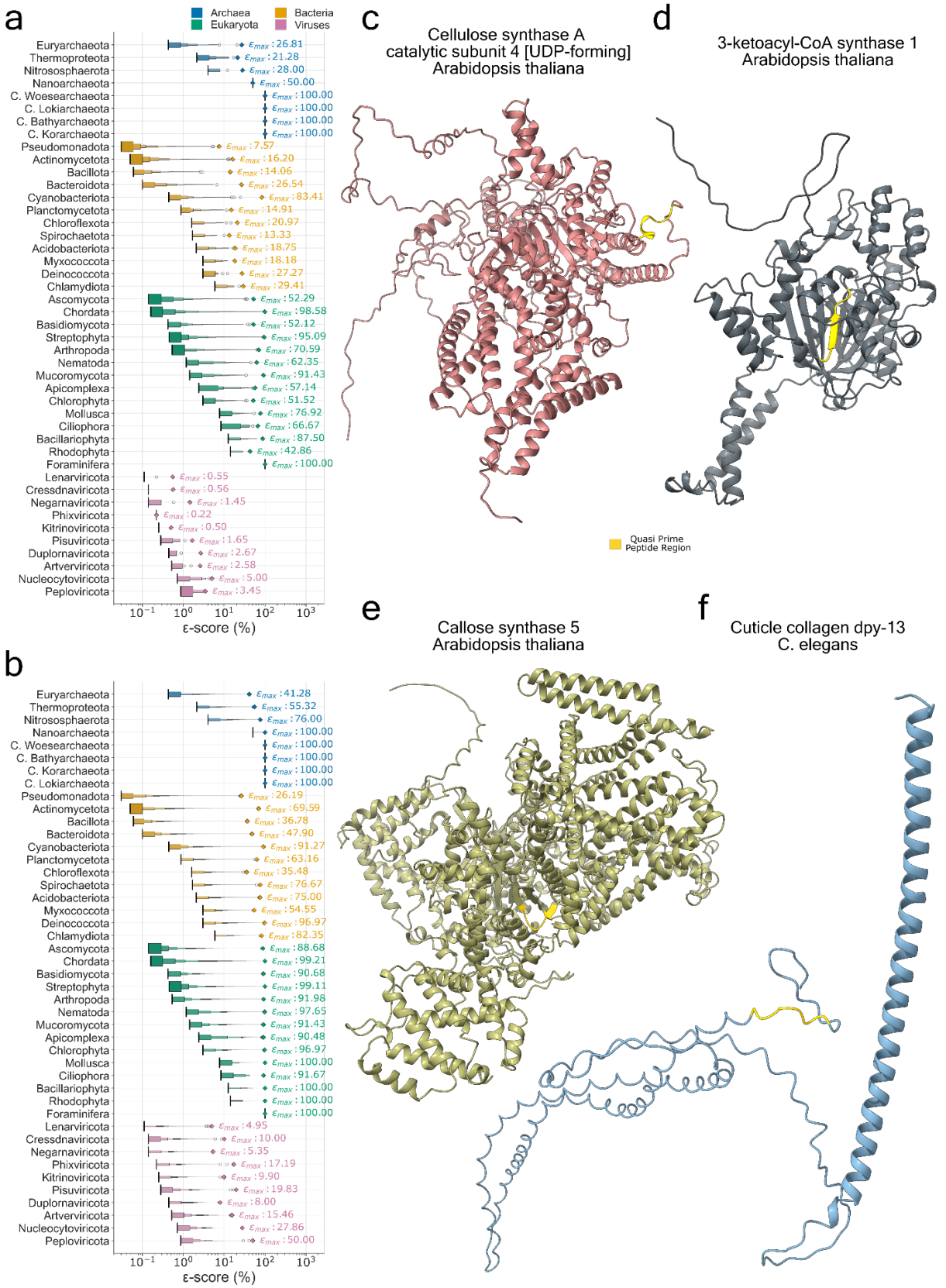

a
Archaea
Eukaryota
Bacteria
Viruses
Euryarchaeota
Thermoproteota
Nitrososphaerota
Nanoarchaeota
C. Woesearchaeota
C. Lokiarchaeota
C. Bathyarchaeota
C. Korarchaeota
Pseudomonadota
Actinomycetota
Bacillota
Bacteroidota
Cyanobacteriota
Planctomycetota
Chloroflexota
Spirochaetota
Acidobacteriota
Myxococcota
Deinococcota
Chlamydiota
Ascomycota
Chordata
Basidiomycota
Streptophyta
Arthropoda
Nematoda
Mucoromycota
Apicomplexa
Chlorophyta
Mollusca
Ciliophora
Bacillariophyta
Rhodophyta
Foraminifera
Lenarviricota
Cressdnaviricota
Negarnaviricota
Phixviricota
Kitrinoviricota
Pisuviricota
Duplornaviricota
Artverviricota
Nucleocytoviricota
Peploviricota
ε-score (%)
b
c
Cellulose synthase A
catalytic subunit 4 [UDP-forming]
Arabidopsis thaliana
d
3-ketoacyl-CoA synthase 1
Arabidopsis thaliana
Quasi Prime
Peptide Region
e
Callose synthase 5
Arabidopsis thaliana
f
Cuticle collagen dpy-13
C. elegans

Figure 4: ε-score distribution of quasi-prime peptides across representative phyla at the superkingdom level. Letter-value plots illustrate the ε-score distributions for representative phyla at the superkingdom level. The ε_M value is depicted as a solid black line and the ε_{max} value as a rhombus. Distributions are shown separately for taxonomic quasi-prime peptides of different lengths: **a** Taxonomic quasi-prime six-mers **b** Taxonomic quasi-prime seven-mers. The color represents the corresponding superkingdom. **c-f** Protein structure of: **c** the Cellulose Synthase A Catalytic Subunit 4, **d** 3-Ketoacyl-CoA Synthase 1, **e** Callose synthase 5, in *Arabidopsis thaliana* and cuticle collagen DPY-13 protein of *C. elegans.* Taxonomic quasi-prime peptides are marked in yellow.

Taxonomic quasi-primes enable the detection of loci divergence and functional adaptations across taxa.

To explore the functional roles of taxonomic quasi-prime peptides, we conducted a GO enrichment analysis tailored to individual taxa. We developed an phylum-wide enrichment estimation via a Random Effects Meta-Analysis model (see Methods). The analysis focused on seven-mer quasi-primes with ε-scores above 90% for each phylum in Archaea, Bacteria, and Eukaryotes, emphasizing representative phylum-wide functional enrichments.

We identified statistically significant (adjusted p-value < 0.05) enriched GO terms across Biological Processes (BP), Molecular Functions (MF) and Cellular Components (CC) **(Figure 5a)**. BPs were the most enriched GO class, followed by MFs, with CCs being less prevalent. Eukaryota, led by Chordata and Streptophyta, exhibited the highest number of enriched GO terms, particularly in BPs, while Archaea and Bacteria showed comparable numbers, reflecting their complexity, proteome size, and horizontal gene transfer **(Figure 5a)**. Within Archaea, Candidatus Bathyarchaeota, one of the most prevalent microorganisms on Earth (Feng et al. 2019), exhibited a high number of enriched BPs **(Figure 5a)**, likely due to its extensive metabolic versatility, linked to protein degradation, glycolysis, and the Wood–Ljungdahl pathway (Feng et al. 2019).

We examined the prevalence of specific GO terms across multiple phyla within each superkingdom. Transmembrane transport was enriched across all three superkingdoms **(Figure 5b),** highlighting lineage-specific conservation. Eukaryotic-exclusive terms, such as RNA polymerase II transcription regulation, were absent in prokaryotes. Ion transport was enriched in eukaryotes with nervous systems, such as Chordata and Mollusca **(Figure 5b)**. Key MFs, like metal ion binding and glycosyltransferase activity, possessed unique quasi-primes across superkingdoms, reflecting their role in fundamental cellular roles (**Figure 5c**). Examples of metal ion binding include oxidoreductase activity requiring iron ions in prokaryotic respiration and energy transduction (Barquera 2014), zinc ion binding for eukaryotic transcription factors (Kamaliyan and Clarke 2024), and glycosyltransferase activity for carbohydrate modifications using manganese ions (Breton et al. 2006) (**Figure 5c**). Enriched 4Fe-4S cluster binding in Archaea and Bacteria highlighted their role in anaerobic respiration and redox regulation (Beinert, Holm, and Münck 1997; Ibrahim et al. 2020). Membrane CC was enriched across all cellular superkingdoms showing showcasing adaptations in membrane protein structures through quasi-primes **(Figure 5d).**

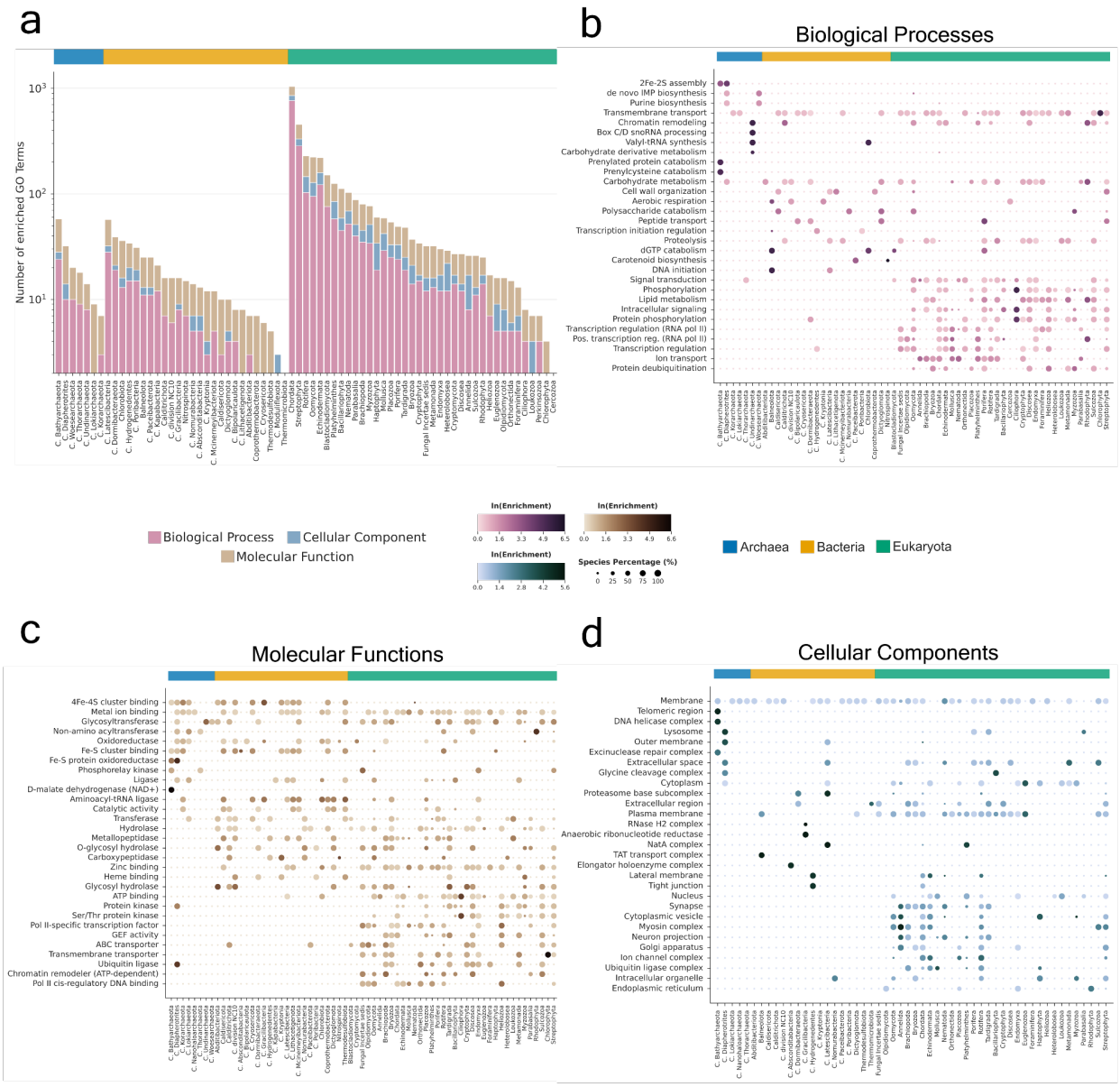


Figure 5: Multi-phylum Gene Ontology (GO) enrichment analysis at superkingdom level. a Stacked bar chart illustrating the total number of enriched GO terms, categorized by GO class, at cellular organism superkingdom level. **b-d** Heatmaps highlighting the top ten GO class specific terms most broadly enriched across phyla within each of the three cellular organism superkingdoms. Dot size indicates the prevalence of each GO term among species within a given phylum, while dot color denotes the combined natural log (ln) of enrichment values, reflecting the strength of enrichment. Heatmaps are organized as follows: **b** Biological Processes **c** Molecular Functions **d** Cellular Components.

To uncover shared functional themes, we analyzed top enriched GO terms for representative Phyla, which were clustered into broader categories using Wang's semantic similarity measure

and hierarchical clustering **(Figure 6a-c)**. In Chordata, receptor tyrosine kinase signaling and ephrin receptor binding **(Figure 6a-b)** were enriched, crucial for nervous system development through axon guidance (Kullander et al. 2001; Kao and Kania 2011) and angiogenesis (N. Cheng, Brantley, and Chen 2002). Clathrin binding, alongside synaptic vesicles and postsynapse, were enriched, highlighting the role of clathrin-mediated endocytosis in Chordata (McMahon and Boucrot 2011) **(Figure 6b-c)**.

Mollusca showed enrichment of estrogen signaling, since they rely on environmental uptake of estrogen hormones (Balbi, Ciacci, and Canesi 2019). Poribacteria, a symbiotic phylum associated with Porifera, showed unexpected enrichment of GO terms related to RNA polymerase II transcription initiation BP and dopamine monooxygenase MF **(Figure 6a-b)**. Poribacteria are known to mimic Porifera species and possess eukaryotic-like protein domains (Kamke et al. 2014), a property that appears to extend to taxonomic quasi-primes as well. In Nematoda, enriched GO terms included G-protein coupled receptor signaling (PLC-activating GPCR), linked to calcium release and muscle contraction, and postsynapse CCs (Liu et al. 2021) **(Figure 6b-c)**.

Cytoskeletal GO term clusters were enriched with taxonomic quasi-primes across multiple phyla **(Figure 6c)**. For instance, spectrin was enriched in Nematoda for stabilizing sensory neurons (Krieg, Dunn, and Goodman 2014), and the F-actin capping complex CC was critical for cytoskeletal remodeling in Chordata (Cooper and Sept 2008). Cytoskeletal components were enriched in other phyla, such as Euglenozoa (microtubules) and Cryptomycota (axonemal microtubules), essential for flagellar motility and host interactions (Jones et al. 2011).

These findings show the unique functional roles of taxonomic quasi-prime peptides with a high ε-score across diverse taxa, with enriched GO terms reflecting adaptations to specific cellular and environmental demands within each superkingdom.

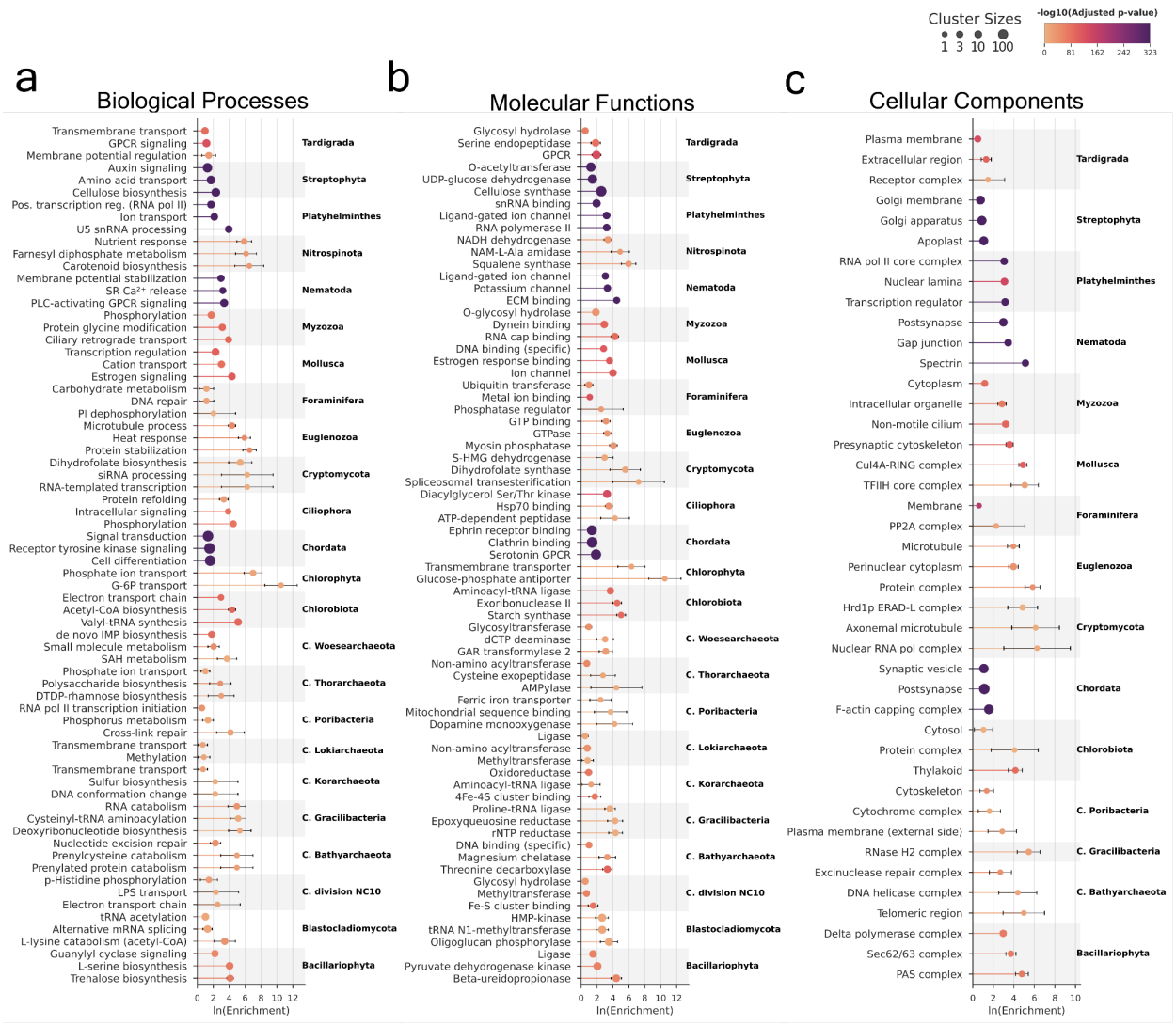


Figure 6: Gene Ontology (GO) Term Enrichment Analysis across superkingdom representative phyla, with taxonomic quasi-prime peptides. Lollipop plots display the mean enrichment of the top three most enriched clustered GO terms for representative phyla of the cellular organism superkingdoms, categorized into: **a** Biological Processes, **b** Molecular Functions, **c** Cellular Components. Only GO terms with an adjusted p-value less than 0.05 and a species representation greater than 5% are included. Dot size represents cluster size, lollipop color represents the -log10 adjusted p-value and error bars show the 95% confidence interval of the calculations. GO terms have been grouped into broader clusters using Wang's semantic similarity measure combined with hierarchical clustering.

Taxonomic quasi-primes detect superkingdom- and phylum-level protein adaptations. To investigate taxon-specific functional changes, we analyzed the distribution of taxonomic quasi-primes within protein domains and families, collectively termed as entries. Across all three superkingdoms, the major facilitator superfamily domain (MFS_dom), critical for transmembrane transport (Pao, Paulsen, and Saier 1998), emerged as a highly enriched domain **(Figure 7a)**,

underscoring its pivotal role in transport processes essential for cellular survival and adaptation (Complete protein domain names can be found at: **Supplementary Table 3**). This aligns with previously found enriched transmembrane transport BP across all three superkingdoms

Archaea and bacteria displayed similar enrichment patterns in entries, which were largely absent in eukaryota, and vice versa **(Figure 7a, Figure 8a)**. This indicates evolutionary conservation of entries across these two superkingdoms, due to their shared environmental challenges, horizontal gene transfer events and adaptive responses. Enriched domains included those associated with oxidoreductase activity, transferase activity, catalytic activity, and metal ion binding. Particularly, the radical S-adenosyl-L-methionine (rSAM) domain and its associated families (e.g., Elp3/MiaA/NifB-like, PqqE-like) **(Figure 7a, Figure 8a)** were enriched, reflecting their roles in essential enzymatic functions. These families belong to the rSAM enzyme superfamily (Frey, Hegeman, and Ruzicka 2008) and involve 4Fe-4S cluster binding (4Fe4S_Fe-S-bd), critical for catalysis.

Aminoacyl-tRNA synthetase domains were enriched in Archaea and Bacteria (aa-tRNA-synth Ia and aa-tRNA-synth II), with bacterial-specific enrichment of the class II G/P/S/T subtype **(Figure 7a)**. Three class Ia tRNA ligase protein families (Valine-, Leucine and Isoleucine-tRNA ligases) are found enriched predominantly in most archaeal and bacterial phyla **(Figure 8a)**, supporting protein synthesis under extreme conditions. These adaptations compensate for the limited diversity of post-transcriptional modifications in prokaryotes.

In Eukaryotes, enriched entries included the ion transport domain, **(Figure 7a)**, found in sodium, potassium and calcium ion channels, the TRIAD supradomain present in the E3 ubiquitin ligase RBR family and the myosin head motor domain alongside the dynein heavy chain family **(Figure 8a)**. E3 ubiquitin ligases are highly conserved across eukaryotes, since they are a part of the ubiquitin-proteasome system, involved in protein degradation and the maintenance of cellular homeostasis (Yang et al. 2021). Enrichment of motor protein entries highlights the necessity of quasi-prime adaptations for intracellular transport and cellular motility.

We also identified the top three enriched entries within specific phyla. In Cyanobacteria entries related to the circadian clock oscillator protein family (e.g. KaiC) **(Figure 7b, Figure 8b)** are enriched suggesting the high evolutionary conservation of quasi-primes in the regulation of day-night cycles (Markson and O'Shea 2009). In Candidatus Paceibacterota the UV-induced DNA damage repair photolyase family was enriched, warranting further study **(Figure 8b)**. In Chordata, the Heparan sulphate-N-deacetylase protein domain, critical for the biosynthesis of heparan sulfate (Sarrazin, Lamanna, and Esko 2011), is highlighted **(Figure 7c)**. Also, the sodium channel A8 and A1 subunit families, alongside the gamma-aminobutyric-acid A receptor, alpha 1 subunit are present across most chordata species further facilitating the involvement of quasi-primes in the function of the nervous system. Leucine-rich repeat families were prominent in Arthropoda **(Figure 8c)** underscoring roles in immune defense and development, with emphasis on Toll-like receptors (Dey et al. 2022) and extracellular matrix integrity (Matsushima, Miyashita, and Kretsinger 2021). Little is known about how these families function in Arthropods, and the existence of highly specific taxonomic quasi-primes emphasizes the need for further study.

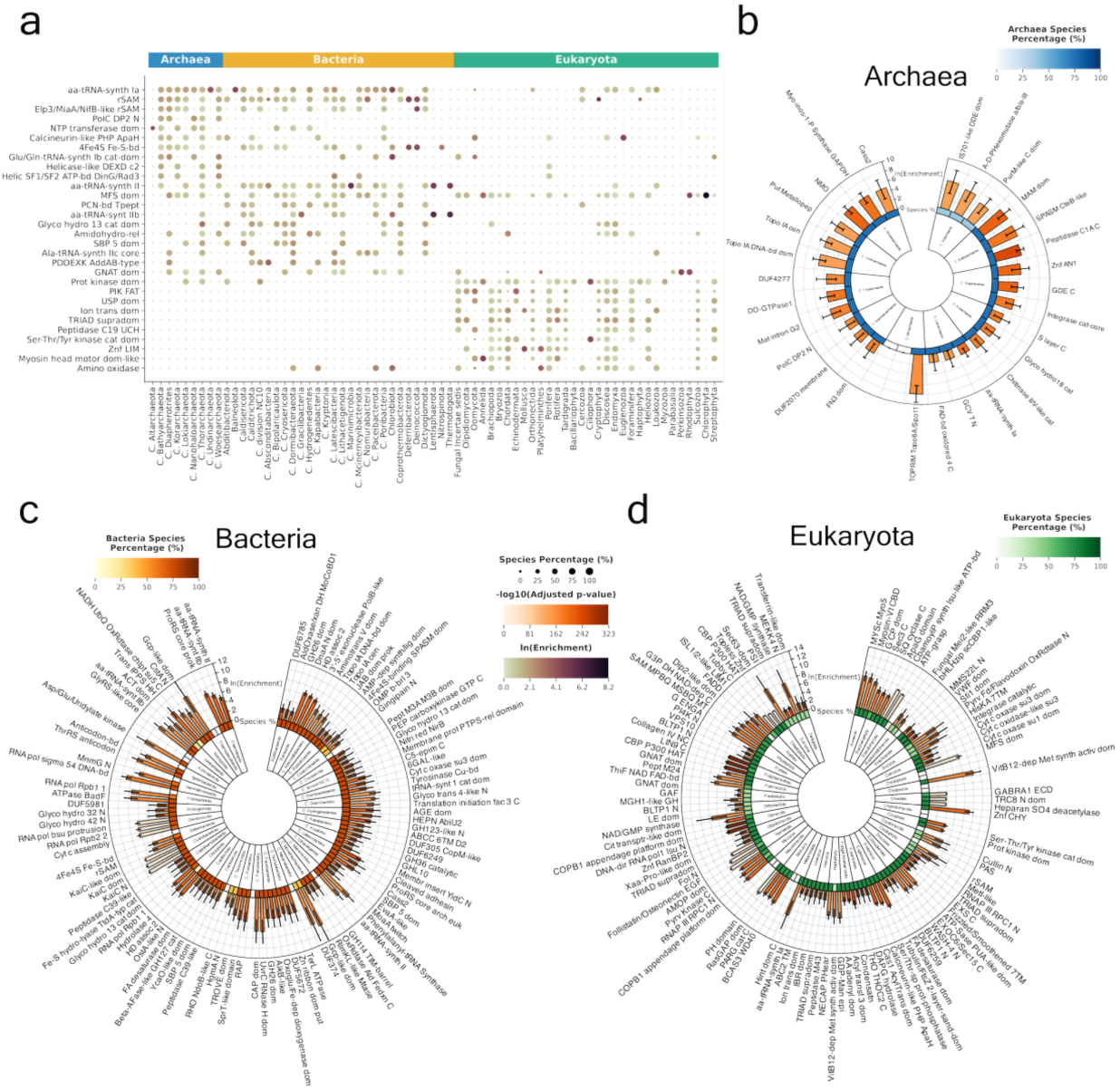


Figure 7: Preferential presence of taxonomic quasi-primes in specific protein domains. **a** Size- and color-coded heatmap displaying the top ten protein domains with taxonomic quasi-primes enriched across most phyla within each of the three cellular organism superkingdoms. Dot size indicates the prevalence of each protein domain within species of the respective phylum, while dot color corresponds to the combined ln(Enrichment) value for each domain. **b-d** Circos plots of the top three enriched protein domains per phylum with an adjusted p-value less than 0.05 and a species prevalence greater than 5% for each cellular organism superkingdom. **b** Archaea, **c** Bacteria, and **d** Eukaryotes. Inner to outer: Phylum name; Heatmap depicting the percentage of the species within the respective phylum that have the specific protein domain enriched; Barplot showing the ln(Enrichment) value with error bars representing the 95% Confidence Interval of the calculation.

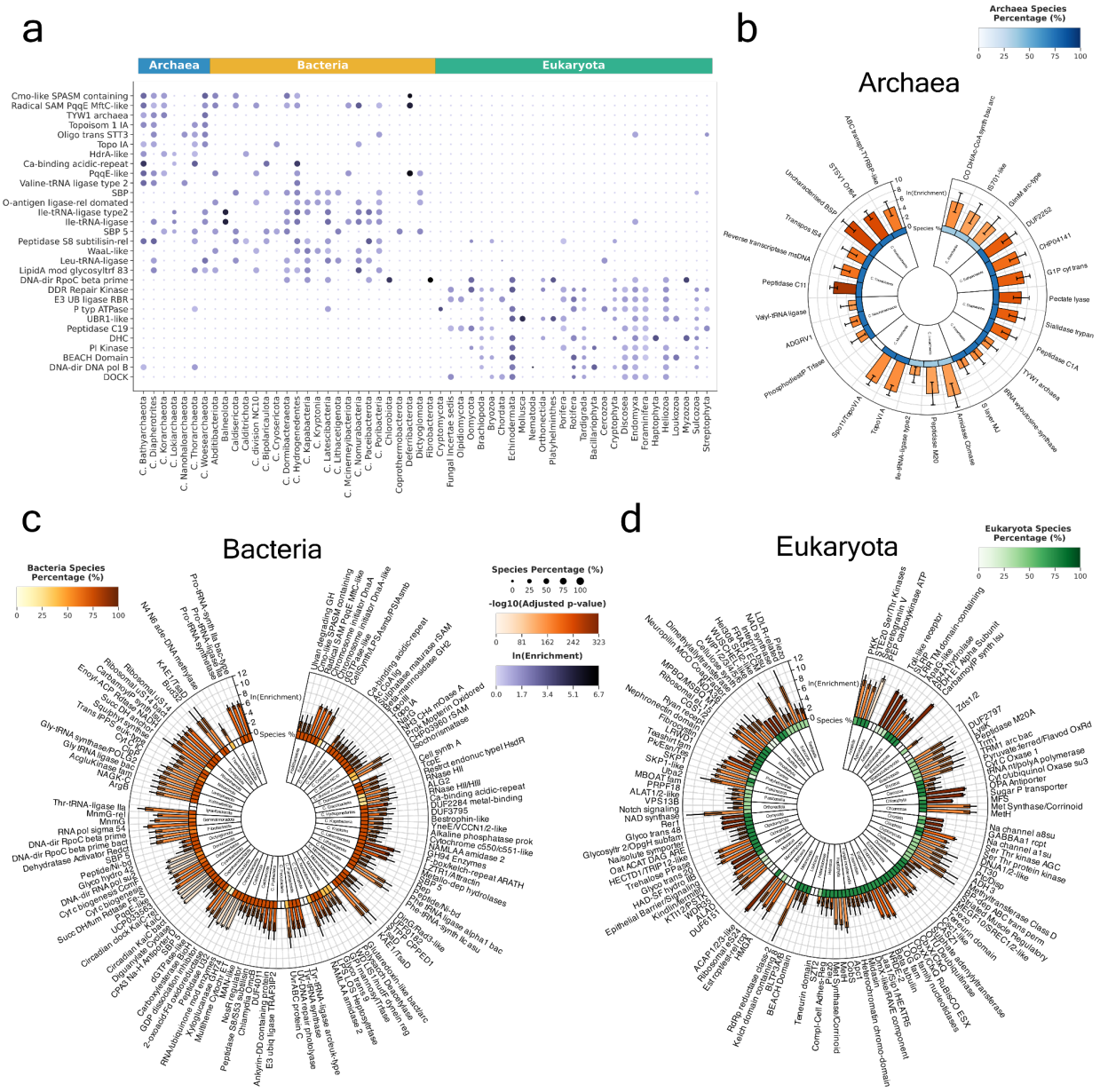


Figure 8: Analysis of the predominant taxonomic quasi-primes in superkingdom specific protein families. a Size- and color-coded heatmap identifying the top ten enriched protein families widely present across phyla within each of the three cellular organism superkingdoms. Dot size indicates the prevalence of each protein family within species of the respective phylum, while dot color corresponds to the combined ln(Enrichment) value for each family. **b-d** Circos plots showcasing the top three most enriched protein families per phylum with an adjusted p-value less than 0.05 and a species prevalence greater than 5% in: **b** Archaea, **c** Bacteria, and **d** Eukaryotes. Inner to outer: Phylum name; Heatmap depicting the percentage of the species within the respective phylum that have the specific protein family enriched; Barplot showing the ln(Enrichment) value with error bars representing the 95% Confidence Interval of the calculation.

Secondary Structure Analysis of Taxonomic Quasi-Primes in Diverse Organisms

We next analyzed the secondary structures of protein loci containing taxonomic quasi-prime seven-mers with an ε-score greater than 90% across thirteen model organisms and seventeen pathogens of global health concern. Our findings revealed a distinct enrichment of taxonomic quasi-primes in protein coils, followed by alpha helices and turns **(Figure 9a-b)**. We evaluated the statistical significance of this structural preference using Kruskal-Wallis tests. For coiled-like conformations (encompassing alpha helices, coils, and turns), the analysis yielded an H-statistic of 213,788.82 with a p-value < 0.001 in model organisms, and an H-statistic of 1,352.19 with a p-value < 0.001 in pathogen proteomes. Nevertheless, we report significant differences depending on the organism studied and the taxonomy; for instance *Schizosaccharomyces pombe* shows a preference for taxonomic quasi-primes in Alpha helices and *Plasmodium falciparum* in Disordered secondary structures (**Figure 9a-b**). These findings showcase the structural diversity and potential functional significance of taxonomic quasi-prime peptides.

Multiple Sequence Alignment and Structural Insights into Chordata Proteins

Building on the aforementioned findings, we selected specific human proteins to investigate the functions of their corresponding quasi-primes in greater detail. We focused our study only on quasi-prime seven-mers with an ε-score greater than 90%, ensuring high confidence in functional relevance (**Supplementary Table 4**). To maintain structural accuracy, we analyzed only protein tertiary structures that were experimentally validated through high-resolution methods such as Cryo-EM or X-ray crystallography. Multiple Sequence Alignment (MSA) was conducted on corresponding protein orthologs across the Chordata species, allowing us to visualize the conservation of quasi-prime regions and to understand evolutionary patterns and functional conservation.

To examine the ion transport function, which is predominantly enriched in Eukaryota, we focused on Sodium Channel 8A (*SCN8A*) **(Figure 9c)**, as the sodium channel α8 family exhibited notable enrichment in Chordata. For processes involving metal ion binding, we selected the E3 ubiquitin-protein ligase ARIH2 (*ARI2*) **(Figure 9d)**, leveraging its association with the enriched TRIAD supradomain and its membership to the E3 ubiquitin ligase RBR family found enriched in the Chordata species, and the heparan sulfate N-deacetylase-N-sulfotransferase 1 (*NDST1*) **(Figure 9e)**, chosen due to the enrichment of the heparan SO_4 deacetylase domain. Lastly, to explore the role of quasi-primes in myosin-related functions and domains, we analyzed Myosin 7 (*MYH7*) **(Figure 9f)**, providing further insight into the functional adaptations of these key proteins.

Twelve taxonomic quasi-primes, spread across various regions of the SCN8A protein, were identified **(Figure 9c)**, suggesting a predominantly structural role rather than a functional one, as these regions do not actively interact with lipid substrates. In the ARI2 protein, three quasi-primes were found to interact directly with zinc ions, through coordinate covalent bonds at the Cys257, Cys260 and His265 sites, which are essential for the structural stability and catalytic activity of the enzyme **(Figure 9d)**, since they are present in the IBR linker domain (Duda et al. 2013). Eleven taxonomic quasi-primes are identified in the NDST1 protein at regions related to calcium and adenosine-3'-5'-diphosphate (A3P). Particularly, the residues His389 and His393 both interact with the calcium ion through coordination covalent bonds, while Phe816 interacts with

A3P through π-stacking interactions **(Figure 9c)**. Quasi-primes present in both the sulfotransferase active site and the deacetylase active site, where they stabilize the structure enabling the correct placement of the N-acetyl-heparan sulfate (Mycroft-West et al., 2024). Additionally, five quasi-primes were identified in the MYH7 protein, interacting with the myosin regulatory light chain 2, ventricular/cardiac muscle isoform (MLC2v) **(Figure 9d)** through one hydrogen bond at the Trp827 site and electrostatic interactions at the positively charged Lys831. This interaction regulates motor function, stabilizes the myosin complex, and enhances calcium sensitivity, ensuring efficient cardiac muscle contraction (Rayment et al. 1993). These functions are critical for proper force generation and adaptability of the heart, with dysregulation linked to cardiomyopathies. These findings highlight the significance of quasi-prime detection in understanding protein interactions and their implications for healthcare.

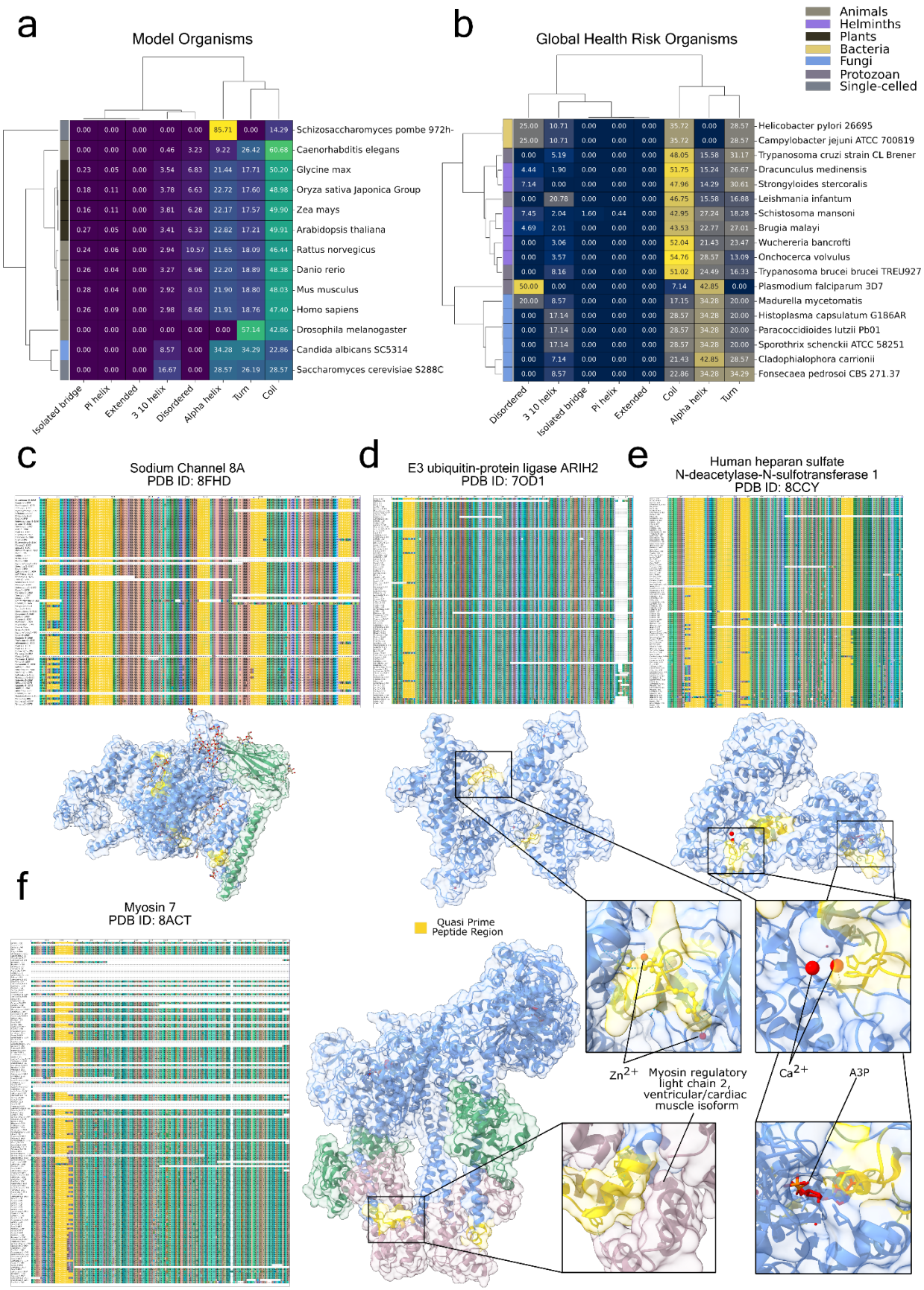

a
Model Organisms
Schizosaccharomyces pombe 972h-
Caenorhabditis elegans
Glycine max
Oryza sativa Japonica Group
Zea mays
Arabidopsis thaliana
Rattus norvegicus
Danio rerio
Mus musculus
Homo sapiens
Drosophila melanogaster
Candida albicans SC5314
Saccharomyces cerevisiae S288C
Isolated bridge
Pi helix
Extended
3 10 helix
Disordered
Alpha helix
Turn
Coil
b
Global Health Risk Organisms
Animals
Helminths
Plants
Bacteria
Fungi
Protozoan
Single-celled
Helicobacter pylori 26695
Campylobacter jejuni ATCC 700819
Trypanosoma cruzi strain CL Brener
Dracunculus medinensis
Strongyloides stercoralis
Leishmania infantum
Schistosoma mansoni
Brugia malayi
Wuchereria bancrofti
Onchocerca volvulus
Trypanosoma brucei brucei TREU927
Plasmodium falciparum 3D7
Madurella mycetomatis
Histoplasma capsulatum G186AR
Paracoccidioides lutzii Pb01
Sporothrix schenckii ATCC 58251
Cladophialophora carrionii
Fonsecaea pedrosoi CBS 271.37
Disordered
3 10 helix
Isolated bridge
Pi helix
Extended
Coil
Alpha helix
Turn
c
Sodium Channel 8A
PDB ID: 8FHD
d
E3 ubiquitin-protein ligase ARIH2
PDB ID: 7OD1
e
Human heparan sulfate
N-deacetylase-N-sulfotransferase 1
PDB ID: 8CCY
f
Myosin 7
PDB ID: 8ACT
Quasi Prime
Peptide Region
Zn2+
Myosin regulatory
light chain 2,
ventricular/cardiac
muscle isoform
Ca2+
A3P

Figure 9: Structural profiling of -ε quasi-primes across evolutionary and pathogenic contexts. a-b Secondary structure profiling of taxonomic quasi-primes in model and global health risk organisms. **a** Clustered heatmap illustrating the secondary structure composition of taxonomic quasi-primes with ε exceeding 90% across selected model organisms. **b** Clustered heatmap of the secondary structure composition of taxonomic quasi-primes with ε values over 90% in selected global health risk organisms, encompassing bacteria, fungi, helminths, protozoans, and other single-celled pathogens. Rows (organisms) and columns (secondary structure types) have been hierarchically clustered based on Euclidean distance, employing Ward's method to reveal patterns and structural similarities among species. Each heatmap cell displays the exact percentage of a specific secondary structure type within each organism. Multiple sequence alignment and structural representations of proteins containing taxonomic quasi-primes with ε-score greater than 90% in: **c** Human Sodium channel 8A (*SCN8A*), **d** Human E3 ubiquitin-protein ligase ARIH2 (*ARIH2*), **e** Human heparan sulfate N-deacetylase-N-sulfotransferase 1 (*NDST1*)**, f** Human Myosin 7 (*MYH7*). Quasi-prime peptide region is displayed in yellow. Inset panels provide a detailed view of these regions, illustrating their specific locations within proteins, where they play a functional role. A3P: Adenosine-3'-5'-Diphosphate

Pathogenic variants are enriched at human taxonomic quasi-prime loci

Finally, we examined if taxonomic quasi-prime loci are more likely to harbor pathogenic variants than surrounding sequences. We used human proteome-wide missense variant effect prediction maps (J. Cheng et al. 2023) and investigated if the subset of missense variants that are pathogenic are more likely to be found at taxonomic quasi-prime loci in humans. We find that the distribution of variants overlapping taxonomic quasi-primes is significantly sifted towards pathogenic effects (Kolmogorov-Smirnov test p-value<0.001, Cliff's delta = 0.448) with taxonomic quasi-prime loci being 2.08-fold more likely to overlap pathogenic variants than expected (**Figure 10**). The increased pathogenicity at taxonomic quasi-prime loci, reflects their significance in traits and characteristics that are taxon-specific.

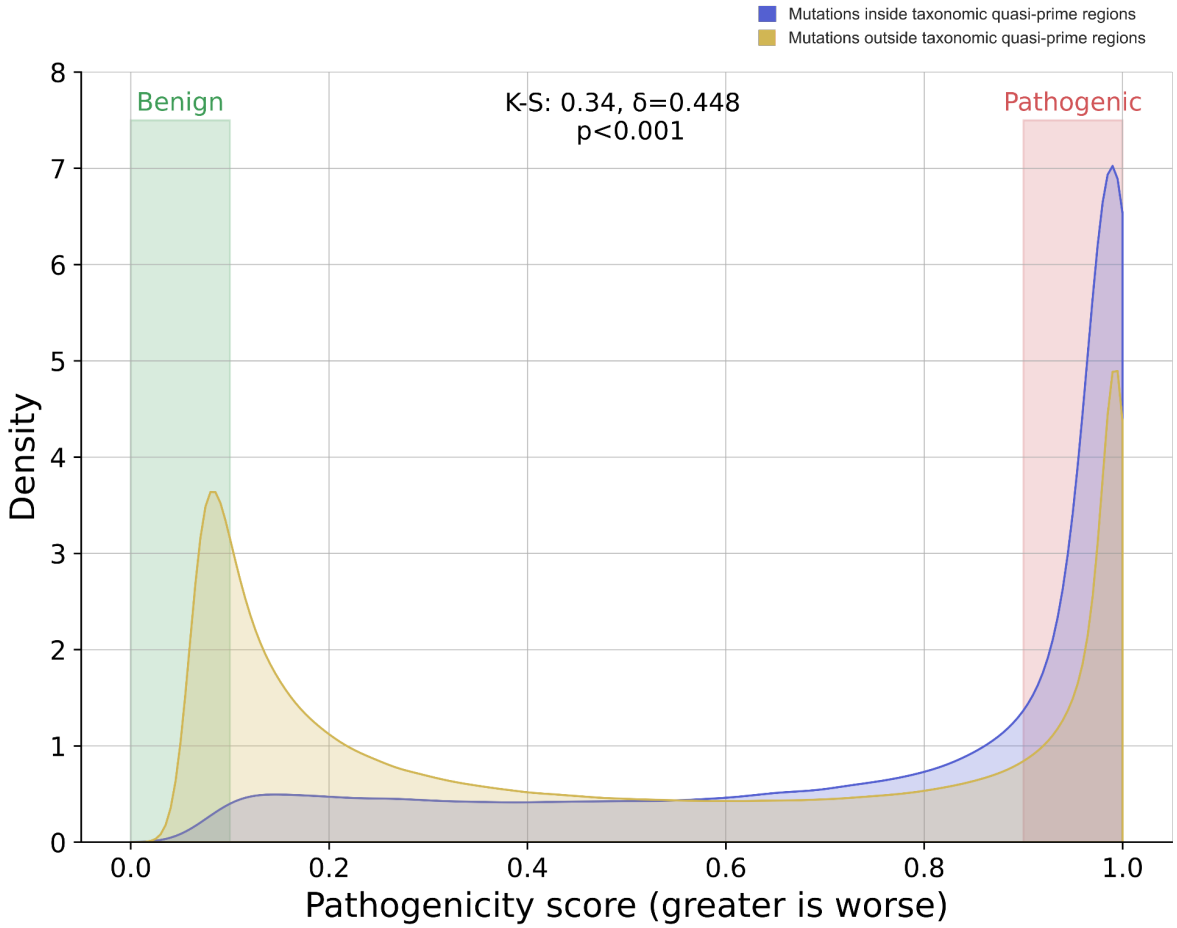


Figure 10: Missense mutations in highly conserved quasi-prime regions are more likely to be pathogenic. The Kernel Density Estimate (KDE) plot compares the distribution of AlphaMissense-predicted pathogenicity scores for missense mutations within human quasi-prime loci (for Chordata quasi-prime 7-mers with ε-scores above 90%) to those outside these loci within the same proteins. Highlighted ranges indicate highly benign mutations (pathogenicity scores < 0.1) and highly pathogenic mutations (pathogenicity scores > 0.9). Statistical metrics: Kolmogorov-Smirnov (K-S) statistic, Cliff's delta (δ), and p-value (p).

Discussion

The presence of short peptide k-mers that are highly specific to individual taxonomies has not been studied to date. Here, we provide evidence that based on the tens of thousands of available reference proteomes, such sequences can be systematically discovered across taxonomic levels. We identify peptides that exhibit remarkable taxonomic uniqueness at six- and seven-amino-acid lengths, offering insights into evolutionary divergence and lineage-specific adaptations. This study highlights taxonomic quasi-prime peptides as a novel approach for advancing our understanding of evolutionary processes and could be useful in developing innovative solutions in healthcare, agriculture, and environmental sustainability.

We observe large variations in the number and frequency of taxonomic quasi-primes when comparing different taxonomies. Protista display substantially lower ε_{max} scores when compared to other eukaryotic kingdoms, indicating that their characteristic quasi-prime peptides appear only in a small subset of the corresponding species. This can be attributed to the highly diverse polyphyletic nature of Protista, which encompasses both multicellular and unicellular organisms. Expanding on this, Protista exhibit a wide variety of environments, from marine and terrestrial ecosystems to parasitic niches, increasing the need to adopt different mechanisms for survivability (Burki, Sandin, and Jamy 2021). In bacteria and archaea, we find that taxonomic quasi-prime peptides are enriched in proteins associated with redox reactions. Eukaryotic phyla, such as Chordata and Streptophyta, displayed the largest number of enriched BPs which is in accordance with their larger proteome sizes and diverse developmental pathways.

The identification of taxonomic quasi-primes offers a novel way to study sequence divergence, speciation, and trait development. By identifying k-mer peptides unique to specific taxa, these sequences can serve as molecular markers of evolutionary processes and lineage-specific adaptations. Analyzing the proteins harboring quasi-prime peptides may reveal how unique traits emerge, such as specialized metabolic pathways or complex physiological processes. Future research could integrate phylogenetic analyses with quasi-prime peptide distributions to trace evolutionary histories and uncover the genetic basis of speciation events, shedding light on how sequence-level variations drive biological diversity and the evolution of novel traits.

Taxonomic quasi-prime peptides could have a wide range of applications in biological research, agriculture, and healthcare. Immunological methods based on antibodies can face limitations in sensitivity and specificity, particularly due to cross-reactivity (Wild 2013). By prioritizing antigens that harbor taxonomic quasi-prime peptides, the sensitivity and specificity of designed antibodies can be increased. Given their short length and unique taxonomic specificity, taxonomic quasi-prime peptides are well-suited for integration into mass-spectrometry workflows (Dupree et al. 2020; Birhanu 2023). Their inherent specificity can enhance the precision of proteomic analyses, enabling more efficient identification of proteins in complex biological samples (Li et al. 2023). In agriculture, taxonomic quasi-prime peptides can serve as highly-specific biomarkers for pathogen detection. Examination of quasi-prime peptides specific to microbial communities or ecological niches could also provide insights into microbiome dynamics and their role in ecosystem function. In healthcare, taxonomic quasi-prime peptides could facilitate the identification of pathogenic organisms in clinical samples, contributing to faster and more precise diagnostics for infectious diseases.

The number of available proteomes represents only a small subset of the species known. Therefore, future work is required to examine how these findings change as more reference proteomes of different organisms become available. Furthermore, the incorporation of protein isoforms and population variants could influence our conclusions and further exploration towards these directions is needed, particularly in eukaryotic proteomes, when such data become available across multiple organisms.

Code availability

The code for taxonomic quasi-prime and taxonomic nullomer extraction and subsequent analysis can be found at:
https://github.com/Georgakopoulos-Soares-lab/taxonomic_quasi_primes

Raw taxonomic quasi-prime extraction data are available at:
https://zenodo.org/records/14385095

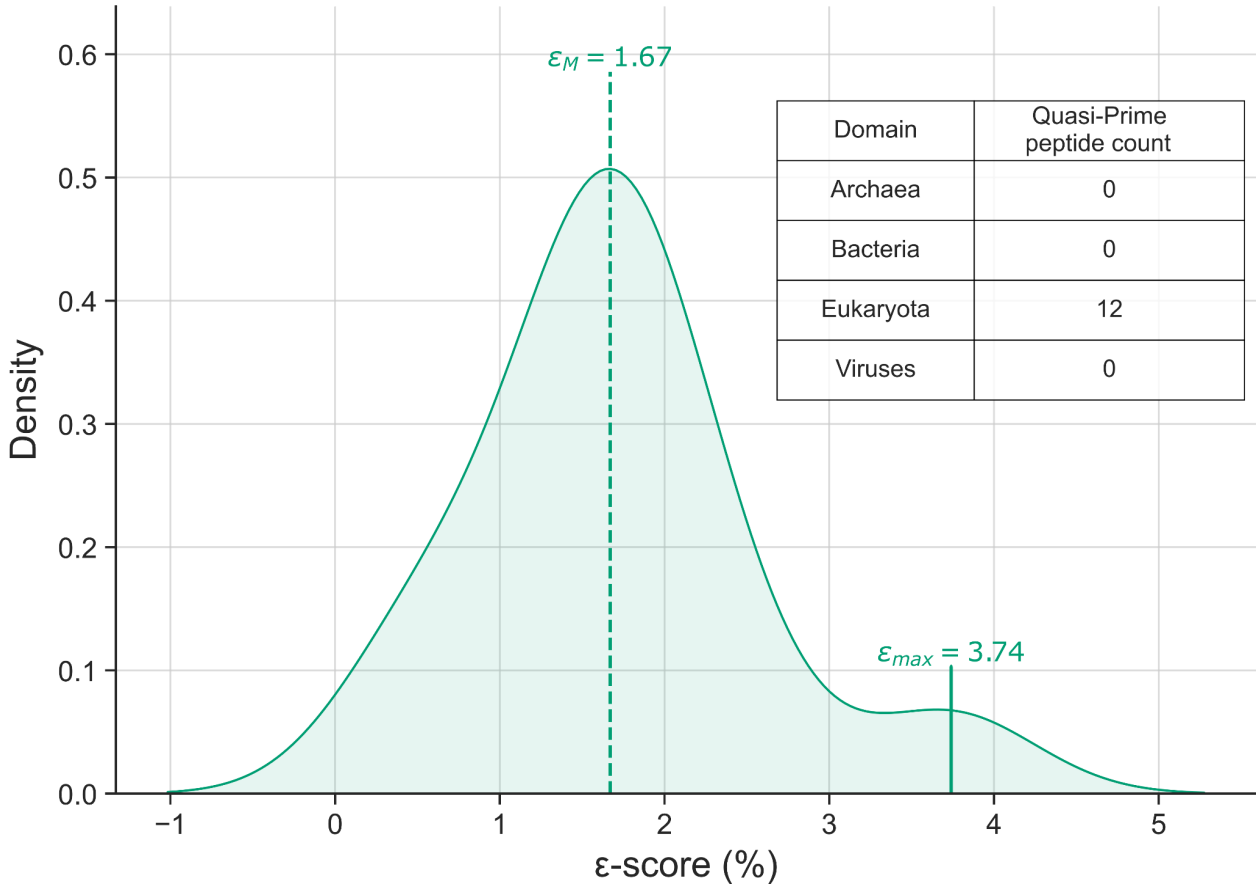
Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM155468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of interests

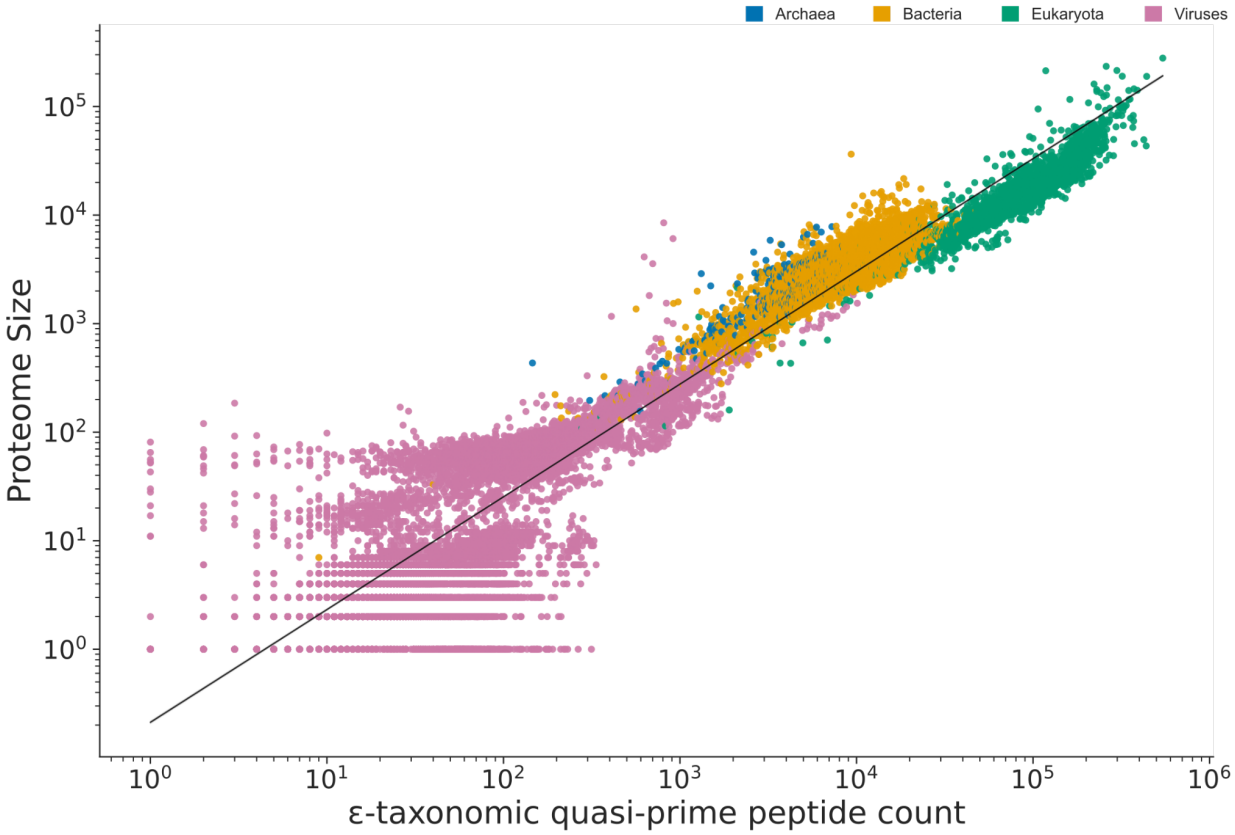
I.M. and I.G.S. have filed patent applications covering embodiments and concepts disclosed in the manuscript (US Patent App. 18/558,992, 2024).

Supplementary Material

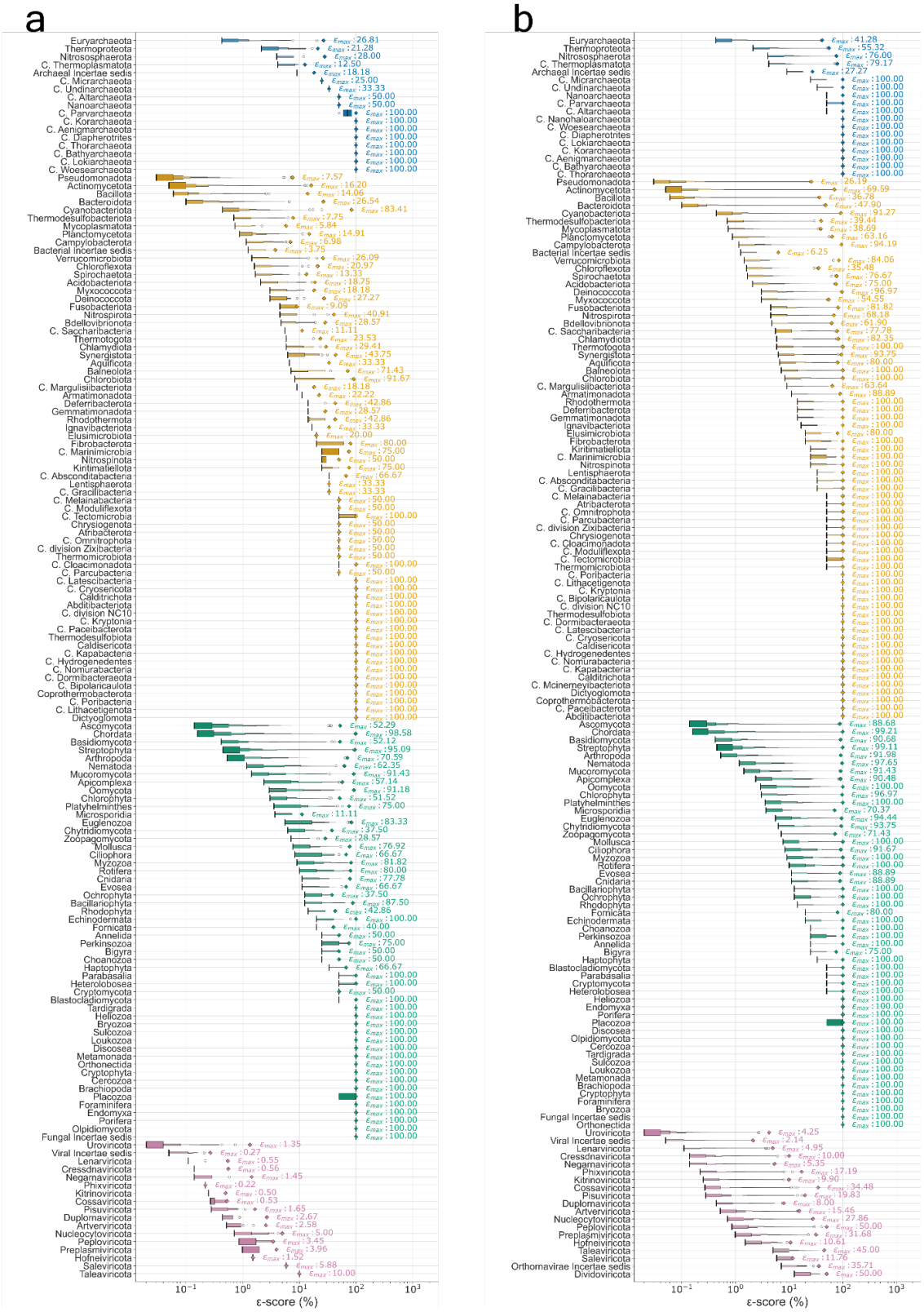


Domain	Quasi-Prime peptide count
Archaea	0
Bacteria	0
Eukaryota	12
Viruses	0

Supplementary Figure 1: Kernel Density Estimate (KDE) plot illustrating the ε-score distribution of superkingdom quasi-prime 5-mers. The smoothing of the curve extending into negative values and exceeding εmax is likely due to the limited sample size of peptides, which affects the Gaussian kernel's estimation fidelity.



Supplementary Figure 2: Scatterplot representing the number of unique taxonomic quasi-prime peptides (6mers and 7mers) observed in each reference proteome as a function of their proteome size.



Supplementary Figure 3: ε-score distribution of quasi-prime peptides across all phyla at the superkingdom level.

Supplementary Table 1: List of Eukaryota quasi-prime 5-mer sequences identified at the superkingdom level, along with their corresponding ε-scores.

Peptide	ε-score (%)
HMCWY	0.52
WCMMW	0.56
WCHCM	1.29
WCCMW	1.41
HYWCC	1.53
HCHWW	1.65
CCHWW	1.69
CHCWM	1.77
KWHMC	1.85
CHCMW	2.05
QYCWC	2.41
WCMCE	3.74

Supplementary Table 2: Quasi-prime peptides (six-mers and seven-mers) exhibiting the highest ε-score across Eukaryotic Kingdoms.

Eukaryotic quasi-prime six-mers			Eukaryotic quasi-prime seven-mers		
Kingdom	Peptide	ε_{max} (%)	Kingdom	Peptide	ε_{max} (%)
Fungi	FPKCYW	34.58	Fungi	DANQDNY	92.50
Metazoa	KWMMYW	87.18	Metazoa	CKGFFKR	98.78
Viridiplantae	YPCFMW	82.88	Viridiplantae	KSCRLRW	98.05
Protista	WHDCHC WWEFYH	17.03	Protista	REENKWC	37.91

Supplementary Table 3: ε_M values for 6mer and 7mer taxonomic quasi-primes across Phyla in each Superkingdom are available at:
phyla_summary.csv

Supplementary Table 4: Quasi-prime 7mers identified in selected Chordata proteins

Protein	Peptide	Region in protein	ε-score (%)
Human Sodium channel 8A	ELPRWHM	913 - 919	93.86
	EWIETMW	936 - 942	98.9
	WIETMWD	937 - 943	98.74
	IETMWDC	938 - 944	98.74
	ETMWDCM	939 - 945	99.21
	TMWDCME	940 - 946	98.27
	MWDCMEV	941 - 947	98.27
	WDCMEVA	942 - 948	98.27
	IVEHNWF	1194 - 1200	91.02
	CLNMVTM	1533 - 1539	98.58
	NMVTMMV	1535 - 1541	98.43
	YFTIGWN	1580 - 1586	95.12
Human E3 ubiquitin-protein ligase ARIH2	CFKCRQM	257 - 263	91.50
	FKCRQMY	258 - 264	91.97
	KCRQMYH	259 - 265	91.97
Human bifunctional heparan sulfate N-deacetylase/N-sulfotransferase 1	KEFWWFP	382 - 388	90.08
	WFPHMWS	386 - 392	98.58
	FPHMWSH	387 - 393	98.58
	PHMWSHM	388 - 394	98.58

	HMWSHMQ	389 - 395	98.58
	MWSHMQP	390 - 396	98.27
	WQDPCED	582 - 588	90.71
	CEDKRHK	586 - 592	93.39
	YLFLGMH	621 - 627	91.50
	KGFWCQL	814 - 820	96.22
	FWCQLLE	816 - 822	90.87
Human Myosin 7	VKNWPWM	824 - 830	94.17
	KNWPWMK	825 - 831	93.86
	NWPWMKL	826 - 832	93.70
	WPWMKLY	827 - 833	90.71
	WMKLYFK	829 - 835	90.71

Supplementary Table 5: Complete protein domain names can be found at: domains_short_names.csv

Supplementary Table 6: Complete protein family names can be found at: families_short_names.csv

Bibliography

Agresti, A. 1999. “On Logit Confidence Intervals for the Odds Ratio with Small Samples.” *Biometrics* 55 (2): 597–602.

Al-Amrani, Safa, Zaaima Al-Jabri, Adhari Al-Zaabi, Jalila Alshekaili, and Murtadha Al-Khabori. 2021. “Proteomics: Concepts and Applications in Human Medicine.” *World Journal of Biological Chemistry* 12 (5): 57–69.

Alileche, Abdelkrim, Jayita Goswami, William Bourland, Michael Davis, and Greg Hampikian. 2012. “Nullomer Derived Anticancer Peptides (NulloPs): Differential Lethal Effects on Normal and Cancer Cells in Vitro.” *Peptides* 38 (2): 302–11.

Alileche, Abdelkrim, and Greg Hampikian. 2017. “The Effect of Nullomer-Derived Peptides 9R, 9S1R and 124R on the NCI-60 Panel and Normal Cell Lines.” *BMC Cancer* 17 (1): 533.

Ali, Nilufar, Cody Wolf, Swarna Kanchan, Shivakumar R. Veerabhadraiah, Laura Bond, Matthew W. Turner, Cheryl L. Jorcyk, and Greg Hampikian. 2024. “9S1R Nullomer Peptide Induces Mitochondrial Pathology, Metabolic Suppression, and Enhanced Immune Cell Infiltration, in Triple-Negative Breast Cancer Mouse Model.” *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 170 (January):115997.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology.” *Nature Genetics* 25 (1): 25–29.

Balbi, Teresa, Caterina Ciacci, and Laura Canesi. 2019. “Estrogenic Compounds as Exogenous Modulators of Physiological Functions in Molluscs: Signaling Pathways and Biological Responses.” *Comparative Biochemistry and Physiology. Toxicology & Pharmacology : CBP* 222 (August):135–44.

Barquera, Blanca. 2014. “The Sodium Pumping NADH:quinone Oxidoreductase (Na^+-NQR), a Unique Redox-Driven Ion Pump.” *Journal of Bioenergetics and Biomembranes* 46 (4): 289–98.

Beinert, H., R. H. Holm, and E. Münck. 1997. “Iron-Sulfur Clusters: Nature’s Modular, Multipurpose Structures.” *Science (New York, N.Y.)* 277 (5326): 653–59.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42.

Bettembourg, Charles, Christian Diot, and Olivier Dameron. 2015. “Optimal Threshold Determination for Interpreting Semantic Similarity and Particularity: Application to the Comparison of Gene Sets and Metabolic Pathways Using GO and ChEBI.” *PloS One* 10 (7): e0133579.

Birhanu, Alemayehu Godana. 2023. “Mass Spectrometry-Based Proteomics as an Emerging Tool in Clinical Laboratories.” *Clinical Proteomics* 20 (1): 1–20.

Breton, Christelle, Lenka Snajdrová, Charlotte Jeanneau, Jaroslav Koca, and Anne Imberty. 2006. “Structures and Mechanisms of Glycosyltransferases.” *Glycobiology* 16 (2): 29R – 37R.

Burki, Fabien, Miguel M. Sandin, and Mahwash Jamy. 2021. “Diversity and Ecology of Protists Revealed by Metabarcoding.” *Current Biology : CB* 31 (19): R1267–80.

Chantzi, Nikol, Manvita Mareboina, Maxwell A. Konnaris, Austin Montgomery, Michail Patsakis, Ioannis Mouratidis, and Ilias Georgakopoulos-Soares. 2024. “The Determinants of the Rarity of Nucleic and Peptide Short Sequences in Nature.” *NAR Genomics and Bioinformatics* 6 (2): lqae029.

Chapman, John R. 2013. *Protein and Peptide Analysis by Mass Spectrometry*. Humana Press.

Chen, Chuming, Zhiwen Li, Hongzhan Huang, Baris E. Suzek, Cathy H. Wu, and UniProt Consortium. 2013. “A Fast Peptide Match Service for UniProt Knowledgebase.” *Bioinformatics* 29 (21): 2808–9.

Cheng, Jun, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, et al. 2023. “Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense.” *Science*, September. https://doi.org/10.1126/science.adg7492.
Cheng, Nikki, Dana M. Brantley, and Jin Chen. 2002. “The Ephrins and Eph Receptors in Angiogenesis.” *Cytokine & Growth Factor Reviews* 13 (1): 75–85.
Cooper, John A., and David Sept. 2008. “New Insights into Mechanism and Regulation of Actin Capping Protein.” *International Review of Cell and Molecular Biology* 267:183–206.
Dey, Debayan, Dipanjana Dhar, Sucharita Das, Aditi Maulik, and Soumalee Basu. 2022. “Structural and Functional Implications of Leucine-Rich Repeats in Toll-like receptor1 Subfamily.” *Journal of Biosciences* 47. https://www.ncbi.nlm.nih.gov/pubmed/36222137.
Duda, David M., Jennifer L. Olszewski, Jonathan P. Schuermann, Igor Kurinov, Darcie J. Miller, Amanda Nourse, Arno F. Alpi, and Brenda A. Schulman. 2013. “Structure of HHARI, a RING-IBR-RING Ubiquitin Ligase: Autoinhibition of an Ariadne-Family E3 and Insights into Ligation Mechanism.” *Structure (London, England : 1993)* 21 (6): 1030–41.
Dupree, Emmalyn J., Madhuri Jayathirtha, Hannah Yorkey, Marius Mihasan, Brindusa Alina Petre, and Costel C. Darie. 2020. “A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field.” *Proteomes* 8 (3). https://doi.org/10.3390/proteomes8030014.
Feng, Xiaoyuan, Yinzhao Wang, Rahul Zubin, and Fengping Wang. 2019. “Core Metabolic Features and Hot Origin of Bathyarchaeota.” *Engineering (Beijing, China)* 5 (3): 498–504.
Frey, Perry A., Adrian D. Hegeman, and Frank J. Ruzicka. 2008. “The Radical SAM Superfamily.” *Critical Reviews in Biochemistry and Molecular Biology* 43 (1): 63–88.
Frishman, D., and P. Argos. 1995. “Knowledge-Based Protein Secondary Structure Assignment.” *Proteins* 23 (4): 566–79.
Gene Ontology Consortium, Suzi A. Aleksander, James Balhoff, Seth Carbon, J. Michael Cherry, Harold J. Drabkin, Dustin Ebert, et al. 2023. “The Gene Ontology Knowledgebase in 2023.” *Genetics* 224 (1). https://doi.org/10.1093/genetics/iyad031.
Georgakopoulos-Soares, Ilias, Ofer Yizhar-Barnea, Ioannis Mouratidis, Martin Hemberg, and Nadav Ahituv. 2021a. “Absent from DNA and Protein: Genomic Characterization of Nullomers and Nullpeptides across Functional Categories and Evolution.” *Genome Biology* 22 (1): 245.
———. 2021b. “Absent from DNA and Protein: Genomic Characterization of Nullomers and Nullpeptides across Functional Categories and Evolution.” *Genome Biology* 22 (1): 245.
Hajisharifi, Zohre, Moien Piryaiee, Majid Mohammad Beigi, Mandana Behbahani, and Hassan Mohabatkar. 2014. “Predicting Anticancer Peptides with Chou’s Pseudo Amino Acid Composition and Investigating Their Mutagenicity via Ames Test.” *Journal of Theoretical Biology* 341 (January):34–40.
Hampikian, Greg, and Tim Andersen. 2007. “Absent Sequences: Nullomers and Primes.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 355–66.
Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. “eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14.
Ibrahim, Iskander M., Huan Wu, Roman Ezhov, Gilbert E. Kayanja, Stanislav D. Zakharov, Yanyan Du, Weiguo Andy Tao, Yulia Pushkar, William A. Cramer, and Sujith Puthiyaveetil. 2020. “An Evolutionarily Conserved Iron-Sulfur Cluster Underlies Redox Sensory Function of the Chloroplast Sensor Kinase.” *Communications Biology* 3 (1): 13.
Jones, Meredith D. M., Irene Forn, Catarina Gadelha, Martin J. Egan, David Bass, Ramon Massana, and Thomas A. Richards. 2011. “Discovery of Novel Intermediate Forms Redefines the Fungal Tree of Life.” *Nature* 474 (7350): 200–203.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf

Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.
Kamaliyan, Zeeba, and Thomas L. Clarke. 2024. “Zinc Finger Proteins: Guardians of Genome Stability.” *Frontiers in Cell and Developmental Biology* 12 (July):1448789.
Kamke, Janine, Christian Rinke, Patrick Schwientek, Kostas Mavromatis, Natalia Ivanova, Alexander Sczyrba, Tanja Woyke, and Ute Hentschel. 2014. “The Candidate Phylum Poribacteria by Single-Cell Genomics: New Insights into Phylogeny, Cell-Compartmentation, Eukaryote-like Repeat Proteins, and Other Genomic Features.” *PloS One* 9 (1): e87353.
Kao, Tzu-Jen, and Artur Kania. 2011. “Ephrin-Mediated Cis-Attenuation of Eph Receptor Signaling Is Essential for Spinal Motor Axon Guidance.” *Neuron* 71 (1): 76–91.
Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.
Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, et al. 2018. “GOATOOLS: A Python Library for Gene Ontology Analyses.” *Scientific Reports* 8 (1): 10872.
Koulouras, Grigorios, and Martin C. Frith. 2021. “Significant Non-Existence of Sequences in Genomes and Proteomes.” *Nucleic Acids Research* 49 (6): 3139–55.
Krasowski, Matthew D., Kazuto Yasuda, Lee R. Hagey, and Erin G. Schuetz. 2005. “Evolutionary Selection across the Nuclear Hormone Receptor Superfamily with a Focus on the NR1I Subfamily (vitamin D, Pregnane X, and Constitutive Androstane Receptors).” *Nuclear Receptor* 3 (September):2.
Krieg, Michael, Alexander R. Dunn, and Miriam B. Goodman. 2014. “Mechanical Control of the Sense of Touch by β-Spectrin.” *Nature Cell Biology* 16 (3): 224–33.
Kullander, K., N. K. Mather, F. Diella, M. Dottori, A. W. Boyd, and R. Klein. 2001. “Kinase-Dependent and Kinase-Independent Functions of EphA4 Receptors in Major Axon Tract Formation in Vivo.” *Neuron* 29 (1): 73–84.
Lacerda, Carla M. R., and Kenneth F. Reardon. 2009. “Environmental Proteomics: Applications of Proteome Profiling in Environmental Microbiology and Biotechnology.” *Briefings in Functional Genomics & Proteomics* 8 (1): 75–87.
Liang, Lung-Yu, Onisha Patel, Peter W. Janes, James M. Murphy, and Isabelle S. Lucet. 2019. “Eph Receptor Signalling: From Catalytic to Non-Catalytic Functions.” *Oncogene* 38 (39): 6567–84.
Li, Leyuan, Tong Wang, Zhibin Ning, Xu Zhang, James Butcher, Joeselle M. Serrana, Caitlin M. A. Simopoulos, et al. 2023. “Revealing Proteome-Level Functional Redundancy in the Human Gut Microbiome Using Ultra-Deep Metaproteomics.” *Nature Communications* 14 (1): 3428.
Liu, Nannan, Yifan Wang, Ting Li, and Xuechun Feng. 2021. “G-Protein Coupled Receptors (GPCRs): Signaling Pathways, Characterization, and Functions in Insect Physiology and Toxicology.” *International Journal of Molecular Sciences* 22 (10). https://doi.org/10.3390/ijms22105260.
Magidovich, Hilla, and Jerry Eichler. 2009. “Glycosyltransferases and Oligosaccharyltransferases in Archaea: Putative Components of the N-Glycosylation Pathway in the Third Domain of Life.” *FEMS Microbiology Letters* 300 (1): 122–30.
Markson, Joseph S., and Erin K. O’Shea. 2009. “The Molecular Clockwork of a Protein-Based Circadian Oscillator.” *FEBS Letters* 583 (24): 3938–47.
Matsushima, Norio, Hiroki Miyashita, and Robert H. Kretsinger. 2021. “Sequence Features, Structure, Ligand Interaction, and Diseases in Small Leucine Rich Repeat Proteoglycans.” *Journal of Cell Communication and Signaling* 15 (4): 519–31.
McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold

Approximation and Projection for Dimension Reduction." http://arxiv.org/abs/1802.03426.
McMahon, Harvey T., and Emmanuel Boucrot. 2011. "Molecular Mechanism and Physiological Functions of Clathrin-Mediated Endocytosis." *Nature Reviews. Molecular Cell Biology* 12 (8): 517–33.
Meng, Elaine C., Thomas D. Goddard, Eric F. Pettersen, Greg S. Couch, Zach J. Pearson, John H. Morris, and Thomas E. Ferrin. 2023. "UCSF ChimeraX: Tools for Structure Building and Analysis." *Protein Science: A Publication of the Protein Society* 32 (11): e4792.
Moeckel, Camille, Manvita Mareboina, Maxwell A. Konnaris, Candace S. Y. Chan, Ioannis Mouratidis, Austin Montgomery, Nikol Chantzi, Georgios A. Pavlopoulos, and Ilias Georgakopoulos-Soares. 2024. "A Survey of K-Mer Methods and Applications in Bioinformatics." *Computational and Structural Biotechnology Journal* 23 (December):2289–2303.
Mouratidis, Ioannis, Fotis A. Baltoumas, Nikol Chantzi, Michail Patsakis, Candace S. Y. Chan, Austin Montgomery, Maxwell A. Konnaris, et al. 2024. "kmerDB: A Database Encompassing the Set of Genomic and Proteomic Sequence Information for Each Species." *Computational and Structural Biotechnology Journal* 23 (December):1919–28.
Mouratidis, Ioannis, Candace S. Y. Chan, Nikol Chantzi, Georgios Christos Tsiatsianis, Martin Hemberg, Nadav Ahituv, and Ilias Georgakopoulos-Soares. 2023. "Quasi-Prime Peptides: Identification of the Shortest Peptide Sequences Unique to a Species." *NAR Genomics and Bioinformatics* 5 (2): lqad039.
Navon, Sharon Penias, Guy Kornberg, Jin Chen, Tali Schwartzman, Albert Tsai, Elisabetta Viani Puglisi, Joseph D. Puglisi, and Noam Adir. 2016. "Amino Acid Sequence Repertoire of the Bacterial Proteome and the Occurrence of Untranslatable Sequences." *Proceedings of the National Academy of Sciences of the United States of America* 113 (26): 7166–70.
Ohama, Takashi. 2019. "The Multiple Functions of Protein Phosphatase 6." *Biochimica et Biophysica Acta, Molecular Cell Research* 1866 (1): 74–82.
Pao, S. S., I. T. Paulsen, and M. H. Saier Jr. 1998. "Major Facilitator Superfamily." *Microbiology and Molecular Biology Reviews : MMBR* 62 (1): 1–34.
Patel, Ami, Jessica C. Dong, Brett Trost, Jason S. Richardson, Sarah Tohme, Shawn Babiuk, Anthony Kusalik, Sam K. P. Kung, and Gary P. Kobinger. 2012. "Pentamers Not Found in the Universal Proteome Can Enhance Antigen Specific Immune Responses and Adjuvant Vaccines." *PloS One* 7 (8): e43802.
Paysan-Lafosse, T., M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, et al. 2023. "InterPro in 2022." *Nucleic Acids Research* 51 (D1). https://doi.org/10.1093/nar/gkac993.
Poznański, Jarosław, Jan Topiński, Anna Muszewska, Konrad J. Dębski, Marta Hoffman-Sommer, Krzysztof Pawłowski, and Marcin Grynberg. 2018. "Global Pentapeptide Statistics Are Far Away from Expected Distributions." *Scientific Reports* 8 (1): 1–18.
Rayment, I., W. R. Rypniewski, K. Schmidt-Bäse, R. Smith, D. R. Tomchick, M. M. Benning, D. A. Winkelmann, G. Wesenberg, and H. M. Holden. 1993. "Three-Dimensional Structure of Myosin Subfragment-1: A Molecular Motor." *Science (New York, N.Y.)* 261 (5117): 50–58.
Sarrazin, Stephane, William C. Lamanna, and Jeffrey D. Esko. 2011. "Heparan Sulfate Proteoglycans." *Cold Spring Harbor Perspectives in Biology* 3 (7). https://doi.org/10.1101/cshperspect.a004952.
Schmid, Jochen, Dominik Heider, Norma J. Wendel, Nadine Sperl, and Volker Sieber. 2016. "Bacterial Glycosyltransferases: Challenges and Opportunities of a Highly Diverse Enzyme Class Toward Tailoring Natural Products." *Frontiers in Microbiology* 7 (February):182.
Steenwyk, Jacob L., Thomas J. Buida 3rd, Yuanning Li, Xing-Xing Shen, and Antonis Rokas. 2020. "ClipKIT: A Multiple Sequence Alignment Trimming Software for Accurate Phylogenomic Inference." *PLoS Biology* 18 (12): e3001007.
Tsiatsianis, Georgios Christos, Candace S. Y. Chan, Ioannis Mouratidis, Nikol Chantzi, Anna

Maria Tsiatsiani, Nelson S. Yee, Apostolos Zaravinos, Verena Kantere, and Ilias Georgakopoulos-Soares. 2024. “Peptide Absent Sequences Emerging in Human Cancers.” *European Journal of Cancer* 196 (January):113421.

UniProt Consortium. 2023. “UniProt: The Universal Protein Knowledgebase in 2023.” *Nucleic Acids Research* 51 (D1): D523–31.

ValizadehAslani, Taha, Zhengqiao Zhao, Bahrad A. Sokhansanj, and Gail L. Rosen. 2020. “Amino Acid -Mer Feature Extraction for Quantitative Antimicrobial Resistance (AMR) Prediction by Machine Learning and Model Interpretation for Biological Insights.” *Biology* 9 (11). https://doi.org/10.3390/biology9110365.

Varadi, Mihaly, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, et al. 2024. “AlphaFold Protein Structure Database in 2024: Providing Structure Coverage for over 214 Million Protein Sequences.” *Nucleic Acids Research* 52 (D1): D368–75.

Wang, James Z., Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. 2007. “A New Method to Measure the Semantic Similarity of GO Terms.” *Bioinformatics (Oxford, England)* 23 (10): 1274–81.

Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. “Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench.” *Bioinformatics (Oxford, England)* 25 (9): 1189–91.

Wen, Jia, Raymond H. F. Chan, Shek-Chung Yau, Rong L. He, and Stephen S. T. Yau. 2014. “K-Mer Natural Vector and Its Application to the Phylogenetic Analysis of Genetic Sequences.” *Gene* 546 (1): 25–34.

Wild, David. 2013. *The Immunoassay Handbook: Theory and Applications of Ligand Binding, ELISA and Related Techniques*. Newnes.

Wu, Chuanyan, Rui Gao, Yusen Zhang, and Yang De Marinis. 2019. “PTPD: Predicting Therapeutic Peptides by Deep Learning and word2vec.” *BMC Bioinformatics* 20 (1): 456.

Yang, Quan, Jinyao Zhao, Dan Chen, and Yang Wang. 2021. “E3 Ubiquitin Ligases: Styles, Structures and Functions.” *Molecular Biomedicine* 2 (1): 23.