THE PROTEIN LANGUAGE VISUALIZER: SEQUENCE SIMILARITY NETWORKS FOR THE ERA OF LANGUAGE MODELS

Javier Espinoza-Herrera[1*], María F. Manríquez-García[2], Sofía Medina-Bermejo[3], Ailyn López-Jasso[4], Karry Shi[1], Dyllan Mead[1], Sarah M. Veskimägi[1], Maeve O'Connor[1], Adriana Siordia[1], Nathaniel Roethler[1], Adrian Jinich[1]

1. Department of Chemistry and Biochemistry, University of California San Diego, San Diego, United States of America
2. Unidad Profesional Interdisciplinaria de Ingeniería Campus Guanajuato IPN, Instituto Politécnico Nacional, Guanajuato, Mexico
3. Facultad de Medicina Mexicali, Universidad Autónoma de Baja California, Mexicali, Mexico
4. Escuela Superior de Medicina, Instituto Politécnico Nacional, Mexico City, Mexico

*jespinozaherrera@ucsd.edu

November 19, 2024

ABSTRACT

The advent of high-throughput sequencing technologies and the availability of biological "big data" has accelerated the discovery of new protein sequences, making it challenging to keep pace with their functional annotation. To address this annotation challenge, techniques such as Sequence Similarity Networks (SSNs) have been employed to visually group proteins for faster identification. In this paper, we present an alternative visual analysis tool that uses Protein Language Model (PLM) embeddings. Our PLVis pipeline employs dimensionality reduction algorithms to cluster similar sequences, enabling rapid assessment of proteins based on their neighbors. Through analysis using average Jaccard distance and cosine similarity metrics, we found that well-separated clusters (those with silhouette scores above 0.95) captured high-dimensional information better than other regions of the projection. While proteins in poorly defined "fuzzy" regions showed similar embeddings to those in neighboring clusters, we note that distances in these projections should not be directly interpreted. To make this pipeline accessible to a wider research community, we have created a Google Colab Notebook for the comparison of protein datasets.

Keywords Protein Language Models · Dimensionality Reduction · Protein Annotation

1 Introduction

High-throughput sequencing technologies have accelerated protein sequence discovery, vastly outpacing our ability to functionally annotate these proteins [1]. For instance, UniProtKB has over 248 million sequence entries, of which less than 1% belong to Swiss-Prot, the manually reviewed section of the database [2]. And despite UniProtKB's capability to automatically annotate its unreviewed entries, the function of over 30% of protein-encoding genes is still unknown, leaving their associated proteins as functionally unannotated [3, 4].

To help navigate these annotation challenges, visual and interactive analysis methods such as Sequence Similarity Networks (SSNs) have gained widespread adoption [5, 6, 7, 8, 9]. SSNs offer a conceptually simple but powerful approach to visualize protein relationships. In SSNs, nodes represent proteins connected by edges when their BLAST-based pairwise alignment exceeds a user-specified similarity threshold, enabling the identification of potential functional clusters [10]. Tools like the EFI-EST web server have made SSNs accessible and popular for studying protein sequence-function relationships [11]. While SSNs offer intuitive visual representations of protein similarities, they complement more advanced statistical methods in protein analysis. Profile Hidden Markov Models (HMMs), a technique originating from speech recognition and later adopted into Natural Language Processing (NLP) [12, 13], detect statistical patterns in protein sequences to help cluster them into families and identify sequence domains [14, 15]. Unlike SSNs, HMM-based visualizations for multiple proteins are often indirect and rarely used, through sequence logos, hierarchical clustering, or heat maps based on HMM scores and patterns [16, 17, 18].

Protein Language Models (PLMs), also adapted from NLP, are the conceptual successors of HMMs [19, 20]. They generally use a transformer neural network architecture [21, 22] and can generate high-dimensional vector representations (embeddings) of both individual amino acids (tokens) and full protein sequences. These PLM embeddings serve as powerful features for downstream prediction and classification tasks, including structure prediction, as well as protein generation and design [21, 23, 24, 25].

The rise of PLMs and their rich embeddings presents an exciting opportunity to design a new generation of interactive visualization tools for protein similarity. Here, we make accessible and systematic a simple yet powerful approach: the interactive exploration of two-dimensional projections of high-dimensional PLM embeddings. Despite the well-documented shortcomings and misuses of 2D reduction visualizations in biology—such as meaningless inter-cluster distances and misleading trajectory inferences [26, 27, 28, 29, 30]—we argue that this approach is valuable and underutilized for exploratory data analysis (EDA) of protein sequences in several contexts.

To validate our approach, we analyze clustering and proximity in 2D projections and compare them with family and domain annotations, as well as the similarity of the underlying high-dimensional embeddings. Well-separated, isolated clusters in the projections reliably capture similar sequences, protein families, and structures. For instance, we observed significantly better similarity scores for proteins within well-isolated clusters, indicating a higher similarity between sets of neighbors in the high-dimensional embeddings and low-dimensional representations. However, we caution against drawing inferences from large, poorly defined protein clusters in the 2D projections and emphasize that inter-cluster distances should not be interpreted as meaningful.

We propose that PLM projections are valuable for researching protein families, proteomes, and sequence-based comparisons across diverse organisms in the Tree of Life. We find these projections particularly useful for comparative analysis of full organism proteomes across different species. To demonstrate this, we present several case studies focusing on the EDA of protein families and proteome comparisons, including more in-depth analyses of Mycobacteria and Plasmodium species. To enhance the interactive power of these visualizations, our tool adds layers of functional annotation and interactivity, including protein structure visualization. We provide a public Google Colaboratory notebook containing our pipeline. This enables researchers to explore their protein datasets, facilitating broader adoption and application of our approach in diverse biological contexts.

2 Results

2.1 Exploring the protein sequence-function space with PLVis projections

We show the main framework for the PLVis pipeline in Figure 1. Each visualization is the result of obtaining the 1024-dimensional embeddings for each of the sequences in the protein set using a PLM (ESM, ProtTrans, etc.). Next, a dimensionality reduction algorithm decreases the number of dimensions for each embedding to 2. In this study, different reduction algorithms (UMAP, tSNE, TriMAP, and PaCMAP,) were tested to find a method that proficiently preserved information and presented it in a visually meaningful way. Finally, a clustering method (K-Means, DBSCAN, etc.) is used to identify groupings of proteins in the visualization, after which n-gram analysis is used to generate a name for each cluster. Our pipeline to generate PLVis projections is modular and lets users choose between different models for embedding generation, dimensionality reduction, and clustering [31, 32, 33, 34, 35].

We first make a head-to-head comparison of our pipeline to the standard BLAST-based approach, SSN. In the last decade, SSNs have been used to explore the sequence-function relationship between families and domains of proteins, this is done by finding large groups of connected nodes in the network (clusters of proteins) and classifying them by using their common traits [9, 36, 37, 38]. One key feature belonging to SSNs is their reliance on a user-set threshold to create the edges between the nodes. This threshold corresponds to the sequence alignment score and it can result in many single isolated proteins when set too high. A lower score can be used to decrease the number of isolated proteins, but you run the risk of overfitting when classifying the information. To analyze the effect of this binary threshold in comparison to a PLVis projection, we compared the visualizations using both methods for 10,000 randomly selected radical SAM (rSAM) enzymes (Fig. 2).

Figure 2A shows that the protein sequence embeddings of the 10,000 rSAM enzymes cluster into different groups in the two-dimensional space. We selected the 5 densest node clusters in the SSN and compared the placement of proteins belonging to those clusters in the PLVis projection, which was further analyzed using k-means clustering, resulting in 88 distinct protein clusters. For the most part, the SSN clusters appear conserved in the PLVis projection, a good example is SSN cluster 4, situated entirely in clusters 7 and 57 of the PLVis. A salient feature of the SSN is the 1,932 proteins (20% of the dataset) that appear as single disconnected nodes at the bottom. We mapped these isolated points on the PLVis, which are now present in 75 of the 88 k-means calculated clusters.

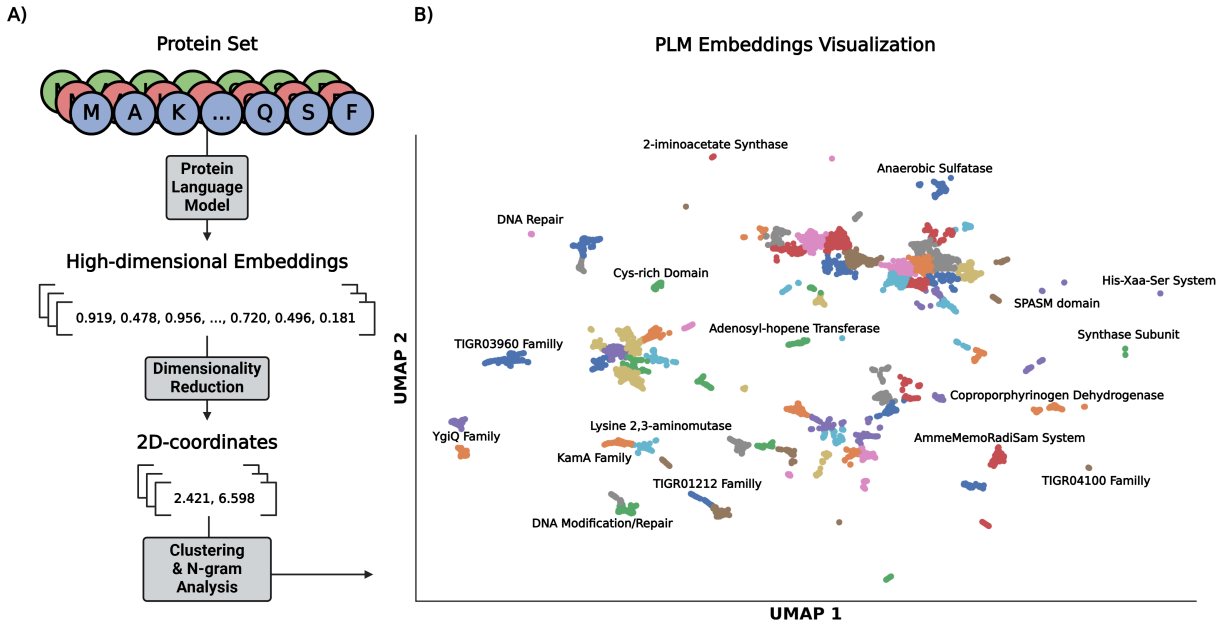

Figure 1: **Schematic overview of the PLVis pipeline.** A) A set of protein sequences is fed to a PLM to obtain embeddings. These embeddings are then reduced to 2 dimensions with a dimensionality reduction algorithm. Lastly, the data is clustered and n-gram analysis is performed to generate appropiate cluster titles to finalize the visualization. (B) Example visualization of the processed data. The arrows indicate the flow of information through the pipeline, allowing users to employ their preferred models for each step.

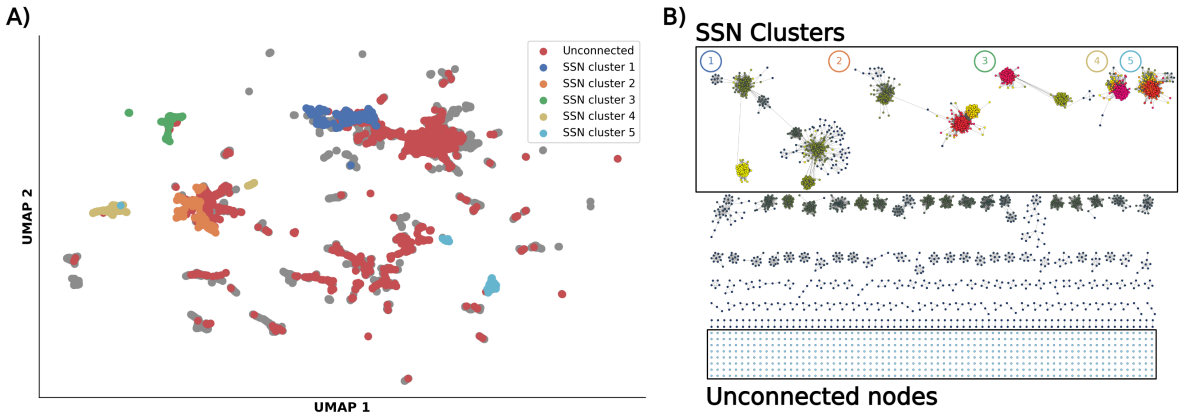

Figure 2: **Comparison of PLVis and SSN visualizations for a random set of 10,000 rSAM enzymes.** (A) Each color in the PLVis represents a cluster in the SSN (blue:1, orange:2, green:3, yellow:4, cyan:5). Proteins colored red in the PLVis represent proteins that are unconnected in the SSN. (B) The SSN was generated using a sequence alignment score of 70. The 5 densest clusters were selected and named accordingly. Nodes situated inside the unconnected region represent proteins that were cut from the threshold.

To assess in more detail how the PLVis projection groups protein information, we performed enrichment analysis on all the k-means calculated clusters using a hypergeometric test. In all cases, we compared the number of proteins with the most common InterPro entries (Domain, Family, and Other) within a cluster and compared them to the complete set of proteins in the projection. In the case of clusters with available InterPro information, 96% were enriched for a particular "Family" entry, 74% for "Domain" and 68% for "Other" entries. Out of the previously mentioned 1,932 previously unconnected proteins, 93% of them belong to clusters enriched for an InterPro "Family" entry. Such is the case of PLVis cluster 45, where 18 previously unconnected proteins are now associated with other proteins that share the InterPro "Family" entry IPR023821, which corresponds to TatD-associated rSAM enzymes. We believe that this analysis demonstrates the value of generating a PLVis plot to group proteins that may appear isolated in SSNs and that are typically discarded from analyses and hypothesis generation.

2.2 Well-separated clusters preserve high-dimensional embedding (ambient space) information in the 2D projection

Aware of the well-documented shortcomings of 2D projections in other biological contexts, most prominently in the field of single-cell genomics, we sought to analyze properties of PLM embedding projections. Previous studies in single-cell genomics have revealed limitations in how dimensionality reduction techniques like t-SNE and UMAP preserve information from high-dimensional data [26, 30]. These limitations can be understood in terms of local and global structure preservation. Local preservation refers to maintaining relationships between neighboring points within clusters, while global preservation concerns the meaningful arrangement of clusters relative to each other in the 2D space. One way to quantify local information loss is by measuring the Jaccard distance for each point in a dataset. This metric assesses the overlap between the N nearest Euclidean neighbors of a point in the original high-dimensional space (ambient space) and its low-dimensional projection (embedding space). In an analysis of 14 single-cell genomics datasets, t-SNE, and UMAP embeddings showed a large average Jaccard distance of 0.7, indicating a substantial loss of local neighborhood information in the dimensionality reduction process [26].

We sought to evaluate this same metric on the 2D projections of PLM embeddings to observe whether or not the projected clusters are maintaining high-dimensional information. By observing functional annotations across clusters through interactive exploration, we formulated the hypothesis that well-separated clusters contain closely related protein sequences, and are better representations of the ambient space than large, central clusters, which tend to have a poorer correlation between neighbors and similar protein families. Well-separated clusters can be intuitively detected as groups of proteins that are tightly packed together and distant from other clusters, and quantitatively measured using metrics such as silhouette scores. To test this, we used three different datasets: the dataset of 10,000 rSAM enzyme proteins discussed above; the *M. tuberculosis* full proteome; and a dataset overlaying the proteome of 8 species belonging to the *Mycobacterium* genus.

To measure the degree of cluster separation, we calculated the average silhouette score S of proteins within a cluster (at a fixed, optimized total cluster number), and defined as "well-separated" those above a threshold value of S>=0.95. These clusters are shown as blue data points in the UMAPs in Figure 3. We calculated the Jaccard distance for each protein in the data and compared the metric between well-separated clusters and the rest of the clusters. The Jaccard score compares the overlap between the N nearest Euclidean neighbors for each protein in ambient and low-dimensional embedding space, in our study N is the number of proteins inside a cluster (with a maximum of N=30). The violin plots in Fig. 3 show that for the three datasets, well-separated clusters have significantly lower average Jaccard distances than randomly selected ones (p-val<10-3, Mann-Whitney U). As a complementary measure of similarity, we compared the average cosine distance of within-cluster ambient PLM embeddings (Fig. 3). Cosine similarity distances are significantly lower (p-val<10-3, Mann-Whitney U) for well-separated clusters, reflecting their more similar PLM embeddings.

We note that the dataset corresponding to the 8 different Mycobacterium proteomes showed the highest number of well-separated clusters, with the lowest Jaccard distance and highest cosine similarity for proteins within these clusters. This reflects the fact that when comparing full organism proteomes, orthologous proteins—those that are highly similar across related species—tend to cluster together, distinct from the rest of the organism's proteome. In these multi-organism, full-proteome comparisons, it becomes easier to visually and quantitatively identify regions of closely related proteins. These findings, although expected and intuitive, highlight the value of using PL-viz to explore full proteome comparisons across organisms visually and interactively.

We then sought to validate the well-established fact that inter-cluster distances in non-linear projections are not particularly meaningful, by evaluating whether nearest neighboring clusters have more similar PLM embeddings than randomly selected clusters. Given that non-linear dimensionality reduction techniques like t-SNE and UMAP warp the shape of the data when projecting to lower dimensions, distances between data points should not be interpreted directly. Using the three previously mentioned datasets, we calculated the average cosine similarity for the embeddings

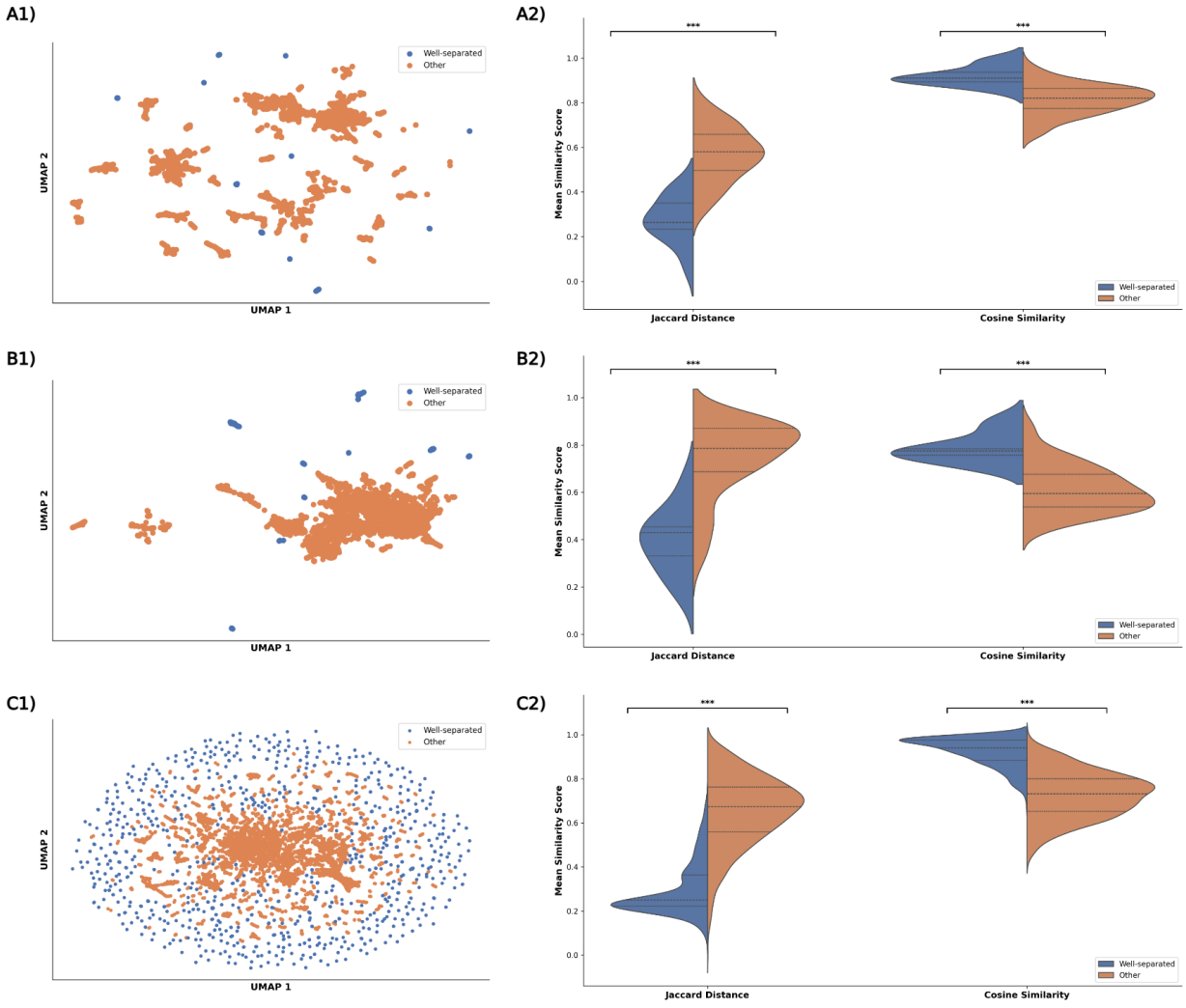

Figure 3: **Well-separated clusters of data are statistically better at conserving high-dimensional data.** (1) UMAP plots of PLM embeddings for (A) 10,000 radical SAM enzymes, (B) Mycobacterium tuberculosis proteome, (C) 8 Mycobacterium genus proteomes; blue - well-separated clusters, detected by silhouette score above threshold (S >= 0.95), orange - clusters with a silhouette score below the threshold. (2) Violin plots of the average Jaccard distance of proteins and cosine similarity of high-dimensional embeddings within well-separated clusters (blue) and the rest of the clusters (orange). Statistical comparison was performed using the Mann-Whitney U test (***p<0.001).

of proteins within each cluster and compared it to the inter-cluster cosine similarity with (1) the nearest neighboring cluster and (2) a randomly selected cluster (Figure 4).

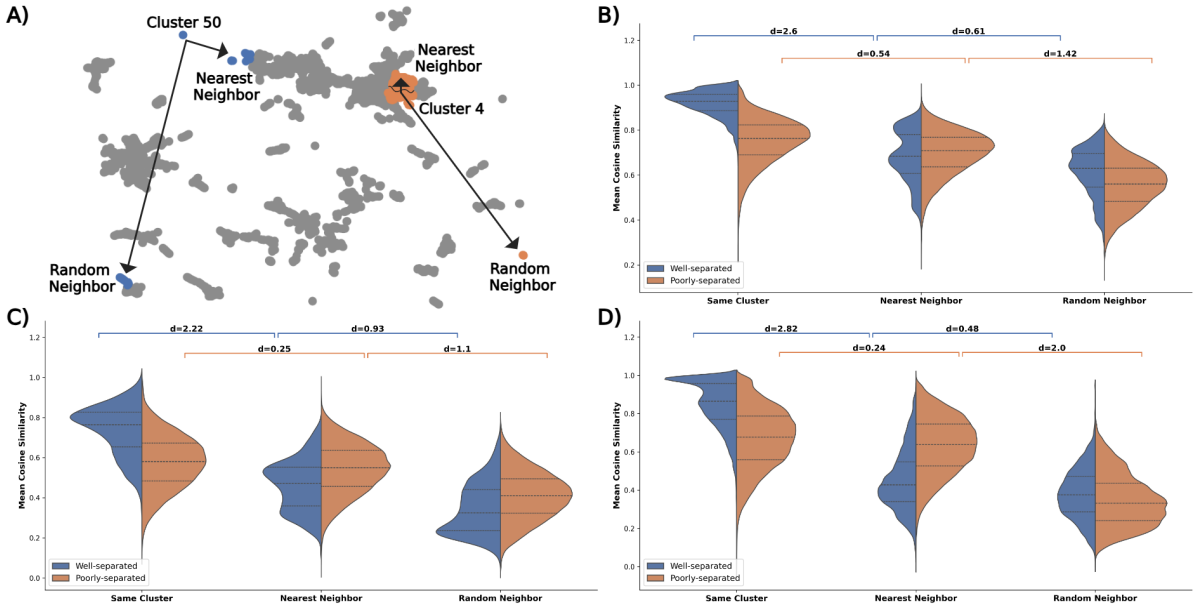

Figure 4: **Distance between clusters in the PLVis projection is not associated with sequence similarity.** (A) Example schematic of a well-separated cluster (blue) and a poorly-separated cluster (orange) and their relative positions to their corresponding closest neighboring cluster and a randomly selected cluster; well-separated clusters, detected by silhouette score above threshold (S >= 0.95), poorly-separated clusters with a silhouette below threshold (S < 0.5). (B, C, D) Violin plots of the mean sequence similarity score for each cluster when comparing its proteins with the nearest neighboring cluster and a randomly selected cluster for (B) 10,000 radical SAM enzymes, (C) Mycobacterium tuberculosis proteome, (D) 8 Mycobacterium genus proteomes. Significance bars represent the effect size between sets using Cohen's D.

In Fig. 4, the violin plots illustrate how cosine similarity varies as we move from proteins within the same cluster to those in the nearest neighboring clusters and finally to random clusters, highlighting trends for both well-separated and poorly-separated clusters. For all three datasets, cosine similarity is notably highest within well-separated clusters, aligning with previous observations on local similarity, while poorly-separated clusters show a more gradual decline. We used Cohen's D to measure the effect in two comparisons: (1) between intra-cluster similarity scores and neighboring-cluster similarity scores, and (2) between neighboring-cluster similarity scores and random-cluster similarity scores. These comparisons were performed separately for both well-separated and poorly-separated clusters. When measuring similarity with the neighboring cluster, proteins belonging to well-separated clusters show a significant drop in the mean, which is not as noticeable when observing the poorly-separated clusters. On the other hand, similar behavior can be observed as we move farther away from the cluster and measure the similarity of proteins with those in a random cluster, but this time, the proteins situated in a poorly-separated cluster show a more significant drop when compared to proteins in well-separated clusters. This implies that sequences in poorly-separated clusters, located in the "fuzzy", cloud-like aggregation of clusters, share a higher similarity with their surrounding proteins in the cloud-like formation. This pattern suggests that the spatial relationship in the final representation maintains some meaningful reflection of the underlying data structure, even though the absolute distances should not be interpreted directly. While the dimensional reduction serves primarily as a visualization tool, these patterns offer additional context for interpreting both local and global relationships between protein sequences in the visualizations.

2.3 PLM embeddings are a fast way to visualize sequence relationships in multi-organism proteome comparisons

Proteins in organisms evolved to fulfill a variety of biological functions. As we travel along a phylogenetic tree, the proteomic content of the species changes accordingly, making PLM embedding projections particularly valuable for

comparing protein families across different organisms in the Tree of Life. For this section, we compared the proteomes of species within genera of pathogenic importance, *Mycobacterium* and *Plasmodium*, responsible for tuberculosis and malaria in humans respectively.

We first generated a PLM embedding visualization for a subset of species from the genus *Mycobacterium*, a group of over 190 Gram-positive bacterial species belonging to the Actinobacteria phylum. These species range from relatively harmless organisms like *M. smegmatis* to dangerous human pathogens like *M. tuberculosis* and *M. leprae* [39, 40]. These bacteria were traditionally classified by their growth rate (slow or rapid), and recent taxonomic revisions have divided them into five distinct genera: *Mycolicibacterium*, *Mycolicibacter*, *Mycolicibacillus*, *Mycobacteroides*, and *Mycobacterium* [41]. To demonstrate the value that PLVis projections have in comparing proteomes across organisms, we analyzed and visualized the dataset containing the proteomes of eight *Mycobacterium* species: *M. smegmatis*, *M. fortuitum*, *M. kansasii*, *M. marinum*, *M. leprae*, *M. tuberculosis*, *M. bovis*, and *M. intracellulare* (shown in Figure 5).

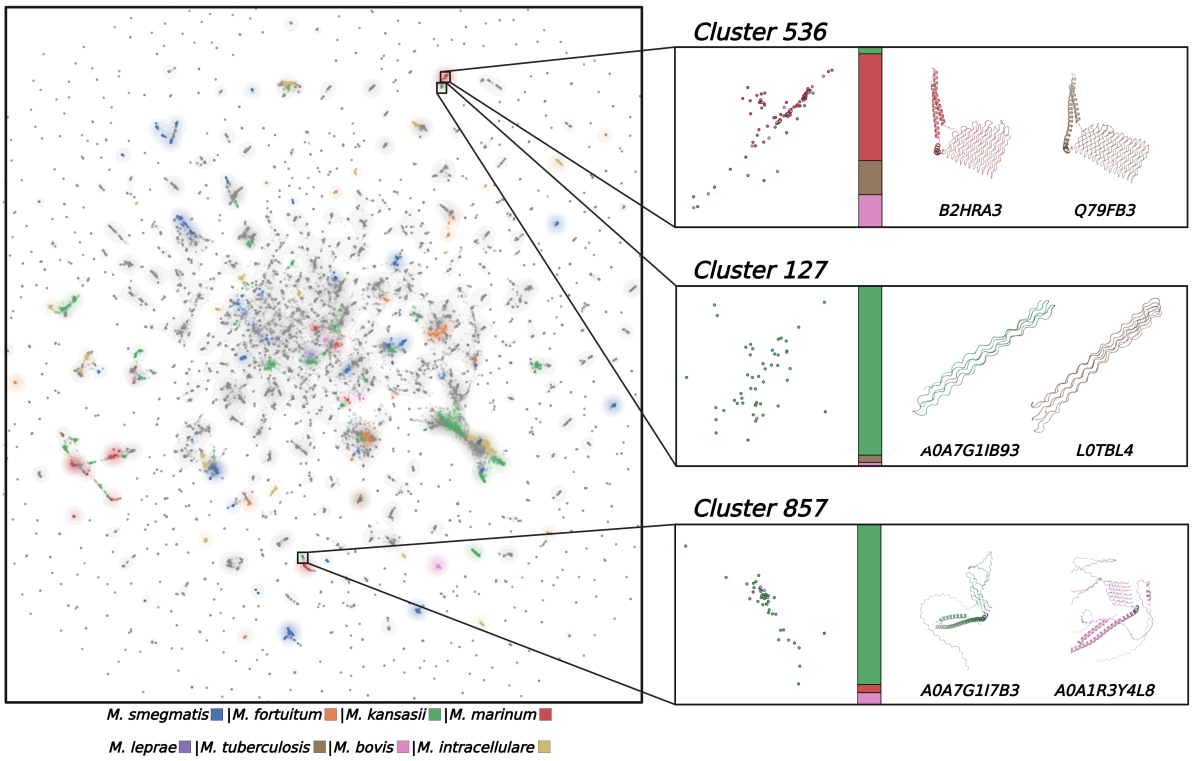

Figure 5: **PLVis for the proteomes of eight mycobacterium species and representative protein structures from three clusters.** Clusters in the visualization are colored by enrichment (blue: *M. smegmatis*, orange: *M. fortuitum*, green: *M. kansasii*, red: *M. marinum*, purple: *M. leprae*, brown: *M. tuberculosis*, pink: *M. bovis*, yellow: *M. intracellulare*), clusters in gray are not enriched for a Mycobacterium species. The three most enriched clusters in the projection (127, 536 & 857) are zoomed in and a color bar showing the fraction of organisms in the cluster is located on their right side. Protein structures generated with AlphaFold 3 are shown for the highlighted cluster and colored according to their source organism.

A key insight from visually comparing proteomes across related organisms is the ability to quickly identify which protein families are enriched or expanded in each organism. We thus performed a hypergeometric test with the Benjamini-Hochberg false discovery rate correction to identify the clusters enriched for a single organism. Out of the 1581 k-clusters, 184 (12%) are enriched and are colored according to their respective organisms in Fig. 5. We found that the three clusters with the lowest corrected p-value (clusters 127, 536 & 857) corresponded not only to closely related species (536: *M. marinum* & 127, 857: *M. kansasii*), but they all contained proteins belonging to the PE-Polymorphic GC-Rich (PE-PGRS) family. These proteins are glycine-rich with multiple GGA/GGN repeats and contain a PE domain near the N-terminus of the sequence as well as a high guanine and cytosine (GC) content of approximately 80% [42, 43, 44]. Cluster 857 contains five glycine-rich "uncharacterized proteins", one of which (A0A7G1IER6) fulfills all previously mentioned qualities (PE domain, GGX motif, and GC content) of a PE-PGRS

family protein. Furthermore, all three clusters were not categorized as well-separated, suggesting that they might be closely related to their neighboring clusters, which is further validated by their positions. Both clusters 127 and 536, are close together and linked with cluster 1389, another enriched cluster with PE-PGRS proteins. Cluster 857, although situated on the other side of the projection, is also surrounded by clusters enriched for PE-PGRS family proteins belonging to *M. marinum* (clusters 149 & 1511).

To understand why the previously mentioned clusters were positioned in separate parts of the visualization, we generated structures for randomly selected proteins using AlphaFold3, as shown in Fig. 5. A close look at the protein structures for each cluster reveals that the visualization separated the protein family according to similarities in their structure. Cluster 127 is characterized by smaller proteins in the group (less than 200 amino acids), with looping patterns that don't fit into the regular secondary structure classification. Clusters 536 and 857 on the other hand do contain a region with alpha-helix patterns near the N-terminal, with the main difference being that the proteins belonging to cluster 857 have a long disordered region near the C-terminus. We thus infer that the projections can separate proteins belonging to the same family according to their structure, which poses a significant advantage when looking for protein analogs to be used in experimental procedures. However, we reiterate that the distance between both groups of clusters is not a measure of their similarity.

Next, we analyzed the *Plasmodium* genus, consisting of protozoan parasites that require a vertebrate and an invertebrate host to complete its life cycle [45]. This genus is medically significant as it contains the parasitic species that cause malaria, a vector-borne infection. Five species within this genus are known to infect humans: *P. falciparum*, *P. malariae*, *P. ovale*, *P. vivax*, and *P. knowlesi* [46]. Similarly to the previous study, we visualized a dataset containing the proteomes of these five parasites, which is shown in Figure 6. Compared to the *Mycobacterium* visualization the *Plasmodium* PLViz has a larger and central poorly-separated/fuzzy region. (S<0.5). Of the 1,942 k-clusters, approximately 36% were poorly-separated, compared to 14% in the *Mycobacterium* projection.

For this dataset, we repeated the hypergeometric test with the Benjamini-Hochberg false discovery rate correction to identify clusters enriched for a single organism, which resulted in the identification of 375 (19%) enriched clusters. We identified 77 enriched clusters that contained proteins exclusively from a single species, a fact that further exemplifies the greater proteomic diversity of this dataset, due to the more complex organisms shown. Because of this greater diversity, one can quickly point out regions in the projection that highlight a specific family of proteins that belong exclusively to a single species in Figure 6. Such is the case for SICAvar proteins of *P. knowlesi*, Fam-L proteins of *P. malariae*, and RIFIN proteins belonging to *P. falciparum*.

It has been shown that RIFIN proteins are used by *P. falciparum* to evade the host immune system by binding to immune-inhibitory receptors [47]. Our analysis revealed that most RIFIN proteins were concentrated in three main clusters (38, 1448, and 1522), with only two RIFIN proteins found elsewhere in clusters 1582 and 1666. These outlier proteins are particularly interesting as they are surrounded by members of multiple protein families (RESA, tryptophan-rich antigen (TRAgs), and Maurer's clefts two transmembrane (PfMC-2TM) proteins) all of which, including RIFIN proteins, are associated with the infected erythrocyte's membrane [48, 49, 50, 51]. This clustering pattern suggests that the 12 "uncharacterized" proteins found in both outlier clusters might also function as erythrocyte surface antigens or membrane proteins. These observations and associated hypotheses showcase how our pipeline can help interactively navigate large-scale protein datasets to reveal biologically significant patterns, while simultaneously providing valuable insights into protein function prediction and pathogen biology.

2.4 Generating proteome comparisons with the PLVis Colab

To assist users in generating their comparisons using the pipeline for the studies above, we developed the PLVis Colab as a user-friendly tool that requires no programming knowledge. This interactive notebook enables users to create exploratory data analysis visualizations by simply uploading data frames and embedding files of the corresponding proteomes. While embeddings can be generated by any Protein Language Model (PLM), the file must be compressed into a GZ or H5 format, this is the default format for embeddings downloaded from the UniProt database.

The pipeline is designed to minimize user input, primarily requiring users to execute each notebook cell sequentially. After the dimensionality reduction of the embeddings, the program performs k-means clustering to determine the optimal number of clusters by analyzing average silhouette scores across different cluster numbers. These results are displayed in a line graph similar to the one shown in Figure 7. Users can either apply the automatically determined optimal number of clusters or specify their preferred number based on the graph. Once all the cells up to this point have been run, a CSV file is generated containing the coordinates for each of the proteins in the data frame.

Before visualizing the generated data frame, a dropdown menu containing the columns present in the data frame is shown to let the user select the coloring for the final visualization. This interactivity allows users to explore different aspects of the data by adjusting the visualization based on column-specific information. Furthermore, the users have the

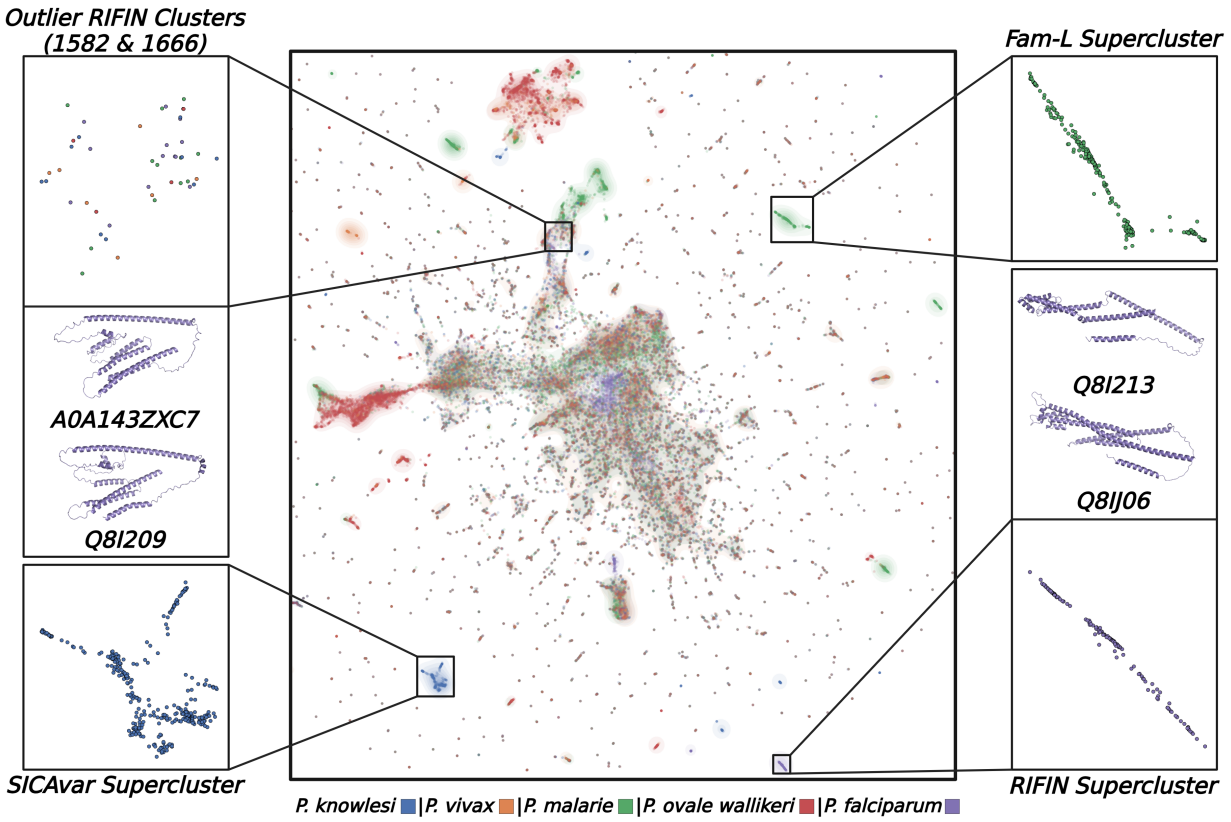

Figure 6: **PLVis projection of the PLM embeddings for the proteomes belonging to five plasmodium species.** Proteins in the visualization are colored according to their species (blue: *P. knowlesi*, orange: *P. vivax*, green: *P. malariae*, red: *P. ovale wallikeri*, purple: *P. falciparum*). The Fam-L, SICAvar and RIFIN superclusters are zoomed in. Protein structures generated with AlphaFold 3 are shown for the RIFIN supercluster and the two outliers in clusters 1582 and 1666.

option of selecting specific head-to-head comparisons from the selected column to filter the projection (e.g. species vs species, gene vs gene, pathway vs pathway) (Shown in Fig. 7). Lastly, after the main results, the Colab includes a quick structure comparison analysis, where the available AlphaFold-generated structures for a specified selection of proteins are retrieved and displayed (Fig. 7). We aim to make the PLVis Colab an invaluable resource for researchers seeking to understand complex protein relationships and patterns in their datasets.

3 Discussion

The PLVis pipeline presented here is an efficient and accessible alternative for the visual representation of protein data obtained from PLM embeddings. When used in conjunction with SSNs, these visualizations enhance protein functional annotation by effectively clustering proteins according to their family classifications. For instance, researchers investigating specific protein families and seeking to validate the function of poorly annotated proteins can utilize PLVis projections to rapidly categorize proteins into distinct subfamilies. This clustering facilitates the identification of promising candidates for experimental validation, particularly when minimally annotated proteins (confidence levels 1 or 2) are found in proximity to well-characterized proteins (confidence level 5).

While the primary strength of PLVis lies in its clustering capabilities, it's important to understand both its limitations and flexibility in practical applications. As stated before, due to the limitations of dimensionality reduction, distances in the visualizations aren't meaningful. However, this opens up opportunities for the users to have the liberty to modify cluster coordinates in their datasets, giving meaning to inter-cluster distance based on additional knowledge. For example, clusters can be spatially organized according to various biological parameters, such as gene expression patterns, protein essentiality profiles, or functional categories (e.g., positioning all redox enzymes in a specific region, or separating

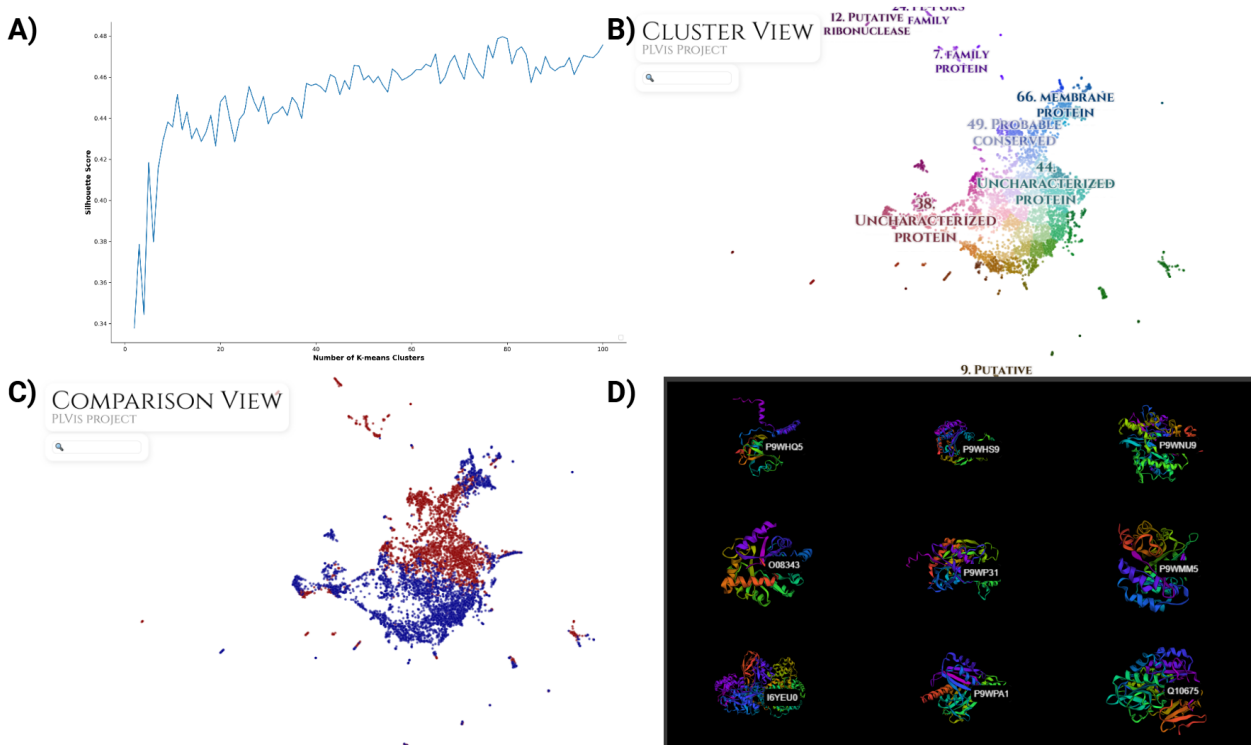

Figure 7: **PLVis Colab Notebook features.** (A) Silhouette score line graph shown in the PLVis Colab for the selection of K-means clusters in the final projection. (B) PLVis generated for the comparison between the proteomes of *M. tuberculosis* and *E. coli*. The color of each protein indicates its cluster in the 2D projection. (C) The PLVis Colab lets users assign a color to each protein based on a selected column in their data frame. In this example, the proteins are colored according to their source organism (red: *M. tuberculosis*, blue: *E. coli*). (D) AlphaFold generated structures for cluster 0 of the *M. tuberculosis* and *E. coli* PLVis comparison.

transcription factors, transporters, and enzymes). This flexibility in visualization emphasizes the importance of domain expertise and underscores the necessity for users to thoroughly understand both their biological data and the analytical tools at their disposal.

Beyond individual protein analysis and cluster organization, PLVis demonstrates remarkable utility in broader comparative studies. From a biological perspective, PLVis projections demonstrate optimal utility in comparative analyses of complete proteomes across different species. The resultant protein clustering patterns reveal significant biological insights, such as species-specific protein family absences or conserved patterns within taxonomic genera. This approach is particularly valuable for analyzing specific biological relationships, exemplified by host-pathogen interactions, where the visualization can identify clusters of proteins from both organisms that may be implicated in pathogenesis. Such protein clusters provide potential molecular signatures associated with disease mechanisms.

The PLVis Colab Notebook offers researchers a robust framework for the visual exploration and analysis of complex proteomic datasets, addressing a critical gap in the interpretation of high-throughput biological data. We encourage the broader scientific community to evaluate and implement PLVis Colab in their research workflows. Furthermore, we actively solicit community feedback and contributions to expand its analytical capabilities and applications. Through iterative development and collaborative refinement, PLVis Colab aims to become an integral component of the modern bioinformatics toolkit, facilitating deeper insights into proteomic data analysis across diverse biological investigations.

References

[1] Valérie de Crécy-lagard, Rocio Amorin de Hegedus, Cecilia Arighi, Jill Babor, Alex Bateman, Ian Blaby, Crysten Blaby-Haas, Alan J Bridge, Stephen K Burley, Stacey Cleveland, Lucy J Colwell, Ana Conesa, Christian Dallago, Antoine Danchin, Anita de Waard, Adam Deutschbauer, Raquel Dias, Yousong Ding, Gang Fang, Iddo Friedberg, John Gerlt, Joshua Goldford, Mark Gorelik, Benjamin M Gyori, Christopher Henry, Geoffrey Hutinet, Marshall

Jaroch, Peter D Karp, Liudmyla Kondratova, Zhiyong Lu, Aron Marchler-Bauer, Maria-Jesus Martin, Claire McWhite, Gaurav D Moghe, Paul Monaghan, Anne Morgat, Christopher J Mungall, Darren A Natale, William C Nelson, Seán O'Donoghue, Christine Orengo, Katherine H O'Toole, Predrag Radivojac, Colbie Reed, Richard J Roberts, Dmitri Rodionov, Irina A Rodionova, Jeffrey D Rudolf, Lana Saleh, Gloria Sheynkman, Francoise Thibaud-Nissen, Paul D Thomas, Peter Uetz, David Vallenet, Erica Watson Carter, Peter R Weigele, Valerie Wood, Elisha M Wood-Charlson, and Jin Xu. A roadmap for the functional annotation of protein families: a community perspective. *Database*, 2022:baac062, January 2022.

[2] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023.

[3] Constance J. Jeffery. Current successes and remaining challenges in protein function prediction. *Frontiers in Bioinformatics*, 3, July 2023. Publisher: Frontiers.

[4] Kimberly A Reynolds, Eduardo Rosa-Molinar, Robert E Ward, Hongbin Zhang, Breeanna R Urbanowicz, and A Mark Settles. Accelerating Biological Insight for Understudied Genes. *Integrative and Comparative Biology*, 61(6):2233–2243, December 2021.

[5] Janine N. Copp, Eyal Akiva, Patricia C. Babbitt, and Nobuhiko Tokuriki. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry*, 57(31):4651–4662, August 2018. Publisher: American Chemical Society.

[6] Nils Oberg, Timothy W. Precord, Douglas A. Mitchell, and John A. Gerlt. RadicalSAM.org: A Resource to Interpret Sequence-Function Space and Discover New Radical SAM Enzyme Chemistry. *ACS Bio & Med Chem Au*, 2(1):22–35, February 2022. Publisher: American Chemical Society.

[7] Remi Zallot, Nils Oberg, and John A. Gerlt. Discovery of New Enzymatic Functions and Metabolic Pathways Using Genomic Enzymology Web Tools. *Current opinion in biotechnology*, 69:77, January 2021.

[8] Amra Dhabalia Ashok, Jella N. Freitag, Iker Irisarri, Sophie de Vries, and Jan de Vries. Sequence similarity networks bear out hierarchical relationships of green cytochrome P450. *Physiologia plantarum*, 176(2):e14244, 2024.

[9] Audrey R. Long, Emma L. Mortara, Brisa N. Mendoza, Emma C. Fink, Francis X. Sacco, Matthew J. Ciesla, and Tyler M. M. Stack. Sequence similarity network analysis of drug- and dye-modifying azoreductase enzymes found in the human gut microbiome. *Archives of Biochemistry and Biophysics*, 757:110025, July 2024.

[10] Holly J. Atkinson, John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS ONE*, 4(2):e4345, February 2009.

[11] John A. Gerlt, Jason T. Bouvier, Daniel B. Davidson, Heidi J. Imker, Boris Sadkhin, David R. Slater, and Katie L. Whalen. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochimica et biophysica acta*, 1854(8):1019, April 2015.

[12] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. Conference Name: Proceedings of the IEEE.

[13] Bhavya Mor, Sunita Garhwal, and Ajay Kumar. A Systematic Review of Hidden Markov Models and Their Applications. *Archives of Computational Methods in Engineering*, 28(3):1429–1448, May 2021.

[14] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, January 1998.

[15] Simon C. Potter, Aurélien Luciani, Sean R. Eddy, Youngmi Park, Rodrigo Lopez, and Robert D. Finn. HMMER web server: 2018 update. *Nucleic Acids Research*, 46(Web Server issue):W200, June 2018.

[16] Benjamin Schuster-Böckler, Jörg Schultz, and Sven Rahmann. HMM Logos for visualization of protein families. *BMC Bioinformatics*, 5(1):7, January 2004.

[17] Travis J. Wheeler, Jody Clements, and Robert D. Finn. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, 15(1):7, January 2014.

[18] Adam Krejci, Ted R. Hupp, Matej Lexa, Borivoj Vojtesek, and Petr Muller. Hammock: a hidden Markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. *Bioinformatics*, 32(1):9–16, January 2016.

[19] Dan Ofer, Nadav Brandes, and Michal Linial. The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750, March 2021.

[20] Tristan Bepler and Bonnie Berger. Learning the Protein Language: Evolution, Structure and Function. *Cell systems*, 12(6):654, June 2021.

[21] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, August 2023. Publisher: Nature Publishing Group.

[22] Abel Chandra, Laura Tünnermann, Tommy Löfstedt, and Regina Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12:e82819, January 2023. Publisher: eLife Sciences Publications, Ltd.

[23] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, July 2022.

[24] Baohui Lin, Xiaoling Luo, Yumeng Liu, and Xiaopeng Jin. A comprehensive review and comparison of existing computational methods for protein function prediction. *Briefings in Bioinformatics*, 25(4):bbae289, July 2024.

[25] Robert Schmirler, Michael Heinzinger, and Burkhard Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, August 2024. Publisher: Nature Publishing Group.

[26] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, August 2023. Publisher: Public Library of Science.

[27] Dmitry Kobak and George C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, February 2021. Publisher: Nature Publishing Group.

[28] Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics*, 15(11):e1008432, November 2019. Publisher: Public Library of Science.

[29] Anant Dadu, Vipul K. Satone, Rachneet Kaur, Mathew J. Koretsky, Hirotaka Iwaki, Yue A. Qi, Daniel M. Ramos, Brian Avants, Jacob Hesterman, Roger Gunn, Mark R. Cookson, Michael E. Ward, Andrew B. Singleton, Roy H. Campbell, Mike A. Nalls, and Faraz Faghri. Application of Aligned-UMAP to longitudinal biomedical studies. *Patterns*, 4(6):100741, June 2023.

[30] Shu Wang, Eduardo D. Sontag, and Douglas A. Lauffenburger. What Cannot Be Seen Correctly in 2D Visualizations Of Single-Cell 'Omics Data? *Cell systems*, 14(9):723, September 2023.

[31] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[32] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, July 2024. Pages: 2024.07.01.600583 Section: New Results.

[33] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018.

[34] Ehsan Amid and Manfred K. Warmuth. TriMap: Large-scale Dimensionality Reduction Using Triplets, March 2022. arXiv:1910.00204.

[35] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *J. Mach. Learn. Res.*, 22(1):201:9129–201:9201, January 2021.

[36] Suwen Zhao, Ayano Sakai, Xinshuai Zhang, Matthew W Vetting, Ritesh Kumar, Brandan Hillerich, Brian San Francisco, Jose Solbiati, Adam Steves, Shoshana Brown, Eyal Akiva, Alan Barber, Ronald D Seidel, Patricia C Babbitt, Steven C Almo, John A Gerlt, and Matthew P Jacobson. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife*, 3:e03275, June 2014. Publisher: eLife Sciences Publications, Ltd.

[37] Katherine H. O'Toole, Barbara Imperiali, and Karen N. Allen. Glycoconjugate pathway connections revealed by sequence similarity network analysis of the monotopic phosphoglycosyl transferases. *Proceedings of the National Academy of Sciences of the United States of America*, 118(4):e2018289118, January 2021.

[38] Angela Giorgianni, Alice Zenone, Leander Sützl, Florian Csarman, and Roland Ludwig. Exploring class III cellobiose dehydrogenase: sequence analysis and optimized recombinant expression. *Microbial Cell Factories*, 23(1):146, May 2024.

[39] Enrico Tortoli, Tarcisio Fedrizzi, Conor J. Meehan, Alberto Trovato, Antonella Grottola, Elisabetta Giacobazzi, Giulia Fregni Serpini, Sara Tagliazucchi, Anna Fabio, Clotilde Bettua, Roberto Bertorelli, Francesca Frascaro, Veronica De Sanctis, Monica Pecorari, Olivier Jousson, Nicola Segata, and Daniela M. Cirillo. The new phylogeny of the genus Mycobacterium: The old and the news. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 56:19–25, December 2017.

[40] Nathan L. Bachmann, Rauf Salamzade, Abigail L. Manson, Richard Whittington, Vitali Sintchenko, Ashlee M. Earl, and Ben J. Marais. Key Transitions in the Evolution of Rapid and Slow Growing Mycobacteria Identified by Comparative Genomics. *Frontiers in Microbiology*, 10, January 2020. Publisher: Frontiers.

[41] Radhey S. Gupta, Brian Lo, and Jeen Son. Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus Mycobacterium into an Emended Genus Mycobacterium and Four Novel Genera. *Frontiers in Microbiology*, 9, February 2018. Publisher: Frontiers.

[42] Flavio De Maio, Rita Berisio, Riccardo Manganelli, and Giovanni Delogu. PE_pgrs proteins of Mycobacterium tuberculosis: A specialized molecular task force at the forefront of host-pathogen interaction. *Virulence*, 11(1):898–915, December 2020.

[43] Christopher D'Souza, Uday Kishore, and Anthony G. Tsolaki. The PE-PPE Family of Mycobacterium tuberculosis: Proteins in Disguise. *Immunobiology*, 228(2):152321, March 2023.

[44] Eliza Kramarska, Flavio De Maio, Giovanni Delogu, and Rita Berisio. Structural Basis of PE_pgrs Polymorphism, a Tool for Functional Modulation. *Biomolecules*, 13(5):812, May 2023.

[45] I W Sherman. Biochemistry of Plasmodium (malarial parasites). *Microbiological Reviews*, 43(4):453–495, December 1979. Publisher: American Society for Microbiology.

[46] Spinello Antinori, Laura Galimberti, Laura Milazzo, and Mario Corbellino. Biology of human malaria plasmodia including Plasmodium knowlesi. *Mediterranean Journal of Hematology and Infectious Diseases*, 4(1):e2012013, 2012.

[47] Fumiji Saito, Kouyuki Hirayasu, Takeshi Satoh, Christian W. Wang, John Lusingu, Takao Arimori, Kyoko Shida, Nirianne Marie Q. Palacpac, Sawako Itagaki, Shiroh Iwanaga, Eizo Takashima, Takafumi Tsuboi, Masako Kohyama, Tadahiro Suenaga, Marco Colonna, Junichi Takagi, Thomas Lavstsen, Toshihiro Horii, and Hisashi Arase. Immune evasion of Plasmodium falciparum by RIFIN via inhibitory receptors. *Nature*, 552(7683):101–105, December 2017. Publisher: Nature Publishing Group.

[48] A. F. Cowman, R. L. Coppel, R. B. Saint, J. Favaloro, P. E. Crewther, H. D. Stahl, A. E. Bianco, G. V. Brown, R. F. Anders, and D. J. Kemp. The ring-infected erythrocyte surface antigen (RESA) polypeptide of Plasmodium falciparum contains two separate blocks of tandem repeats encoding antigenic epitopes that are naturally immunogenic in man. *Molecular Biology & Medicine*, 2(3):207–221, June 1984.

[49] Bo Wang, Feng Lu, Yang Cheng, Jun-Hu Chen, Hye-Yoon Jeon, Kwon-Soo Ha, Jun Cao, Myat Htut Nyunt, Jin-Hee Han, Seong-Kyun Lee, Myat Phone Kyaw, Jetsumon Sattabongkot, Eizo Takashima, Takafumi Tsuboi, and Eun-Taek Han. Immunoprofiling of the Tryptophan-Rich Antigen Family in Plasmodium vivax. *Infection and Immunity*, 83(8):3083–3095, July 2015. Publisher: American Society for Microbiology.

[50] Iryna Tsarukyanova, Judy A. Drazba, Hisashi Fujioka, Satya P. Yadav, and Tobili Y. Sam-Yellowe. Proteins of the Plasmodium falciparum two transmembrane Maurer's cleft protein family, PfMC-2TM, and the 130 kDa Maurer's cleft protein define different domains of the infected erythrocyte intramembranous network. *Parasitology Research*, 104(4):875–891, March 2009.

[51] Mohamed S. Abdel-Latif, Klaus Dietz, Saadou Issifou, Peter G. Kremsner, and Mo-Quen Klinkert. Antibodies to Plasmodium falciparum rifin proteins are associated with rapid parasite clearance and asymptomatic infections. *Infection and Immunity*, 71(11):6229–6233, November 2003.