Deep embeddings to comprehend and visualize microbiome protein space

Krzysztof Odrzywolek[1,2,*], Zuzanna Karwowska[3,*], Jan Majta[1,4], Aleksander Byrski[2], Kaja Milanowska-Zabel[1,$], Tomasz Kosciolek[3,^]

[1] - Ardigen, Podole 76, 30-394 Kraków, Poland

[2] - Institute of Computer Science, Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, Mickiewicza 30, 30-059 Kraków, Poland

[3] - Malopolska Centre of Biotechnology, Jagiellonian University, Gronostajowa 7A, 30-387 Krakow, Poland

[4] - Department of Computational Biophysics and Bioinformatics, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland

** co-first authors*
$ kaja.milanowska-zabel@ardigen.com
^ tomasz.kosciolek@uj.edu.pl

Abstract

Understanding the function of microbial proteins is essential to reveal the clinical potential of the microbiome. The application of high-throughput sequencing technologies allows for fast and increasingly cheaper acquisition of data from microbial communities. However, many of the inferred protein sequences are novel and not catalogued, hence the possibility of predicting their function through conventional homology-based approaches is limited. Here, we leverage a deep-learning-based representation of proteins to assess its utility in alignment-free analysis of microbial proteins. We trained a language model on the Unified Human Gastrointestinal Protein catalogue and validated the resulting protein representation on the bacterial part of the SwissProt database. Finally, we present a use case on proteins involved in SCFA metabolism. Results indicate that our model (ArdiMiPE) manages to accurately represent features related to protein structure and function, allowing for alignment-free protein analyses. Technologies such as ArdiMiPE that contextualize metagenomic data are a promising direction to deeply understand the microbiome.

Introduction

In just over a decade, a substantial body of evidence linked gut microbiome dysbiosis with diseases ranging from obesity[1], inflammatory bowel disease[2–4], diabetes[5,6], cancer[7,8], depression[9] and other psychiatric disorders[10,11]. It shows the profound impact of the microbiome on human health and is a testament to rapid technological progress in sequencing technologies. Since the mid-2000s, the bulk of our insight into the role of the microbiome came from high-throughput and cost-effective 16S rRNA marker gene sequencing experiments that allow for taxonomic discrimination between microorganisms. Though informative, microbiome analysis based solely on taxonomy is prone to bias, due to incomplete reference databases and does not provide detailed information about microbiome function[12]. One of the areas of high interest and relevance is our ability to deduce the gene function from sequence, as it provides more insight into the microbiome's role in human health. Functional analysis of microbiome data can be performed, based on high-throughput, large-scale shotgun metagenomics and other multi-omics experiments that are now becoming accessible for large-scale studies. Gene sequence fragments generated during a shotgun sequencing experiment can be functionally annotated, using homology-based tools such as BLAST or HMMER that search fragments of sequences against reference databases such as Pfam or Gene Ontology (GO)[13]. Similarly to 16S sequencing, functional assignment can be biased, due to incomplete reference databases; so far, only up to 50% of all microbial protein sequences may be annotated[14]. Despite remarkable progress in the last decades, developing precise methods for function prediction is still a major challenge in bioinformatics (see CAFA [15] initiative). The volume of metagenomic data is making the problem even more difficult to deal with. Thus, introducing an *in silico* method to help assign protein functions could prove highly beneficial for realizing the full potential behind metagenomics and multi-omics.

Deep learning is a proven technique for dealing with intricate problems and has been shown to work exceptionally well for tasks like speech recognition, natural language processing (NLP), or image classification[16]. Recently, it has been successfully employed for analyzing biological sequences, like genomes or proteomes[17]. Perhaps the best-known example of the use of deep learning in biology was the protein structure prediction problem. DeepMind's AlphaFold models[18,19] won the last two Critical Assessment of protein Structure Prediction (CASP) challenges - CASP13[20] and CASP14, bringing a seismic shift to this decades-old field. The main reason for the notable success of Deep Neural Networks in these areas of biology is their ability to process massive amounts of data, even unlabeled, and extract meaningful patterns from them. Deep learning can leverage the exponential growth of data available in biological databases, which may be limiting for traditional methods. In multi-omics, especially when considering protein information, the capability to learn from unlabeled data is particularly valuable. The gap between the number of unlabeled and labeled protein sequences is widening every year, also thanks to metagenomics, which enabled a more rapid acquisition of data and gave access to uncultured organisms (https://www.uniprot.org/statistics/TrEMBL).

So far, deep learning methods in protein bioinformatics were employed in two ways: to directly annotate the sequence (supervised learning) or to create a representation of a protein (for example, a sequence embedding using self-supervised learning). Annotation using deep learning is a natural extension of traditional methods, which aim to assign a label to a newly sequenced protein. The label is usually connected to an entry from a database of choice and may belong to curated ontologies (e.g., GO terms[21]) or classification schemes (e.g., EC numbers[22]). Accordingly, studies in the last decade show that deep learning can successfully predict EC numbers[23,24], GO terms[25–30], PFAM families[31,32], or multiple labels at once[33]. However, the labeled proteins are not only in shortage, limiting the potential of deep learning, but also skewed towards model organisms, which may result in biased models.

To overcome these obstacles, more recent approaches use massive unlabeled datasets (UniParc, BFD, PFAM) to train self-supervised models. These models analyze raw amino acid sequences in an alignment-free fashion to learn statistical representations of a protein. The representation can then be effectively used for downstream analyses and predictions of, e.g. secondary or tertiary structure, protein stability, contact map[34,35], protein function[36,37], localization[38,39], variant effect[40], protein engineering[40,41], remote homology detection[34] and more. Moreover, deep-learning-based methods can be used to analyze proteins that do not resemble any catalogued proteins, which is particularly useful in the case of the under-annotated microbiome protein space. Deep-learning-based representations are computationally efficient and accurate, hence they seem appropriate to leverage large amounts of data in high-volume metagenomic studies. Compared to standard bioinformatic tools used to functionally annotate sequences, such as BLAST and HMMER, deep-learning methods require a larger amount of resources, but only during the training phase.

Here, we describe the ArdiMiPE (Ardigen Microbial Protein Embeddings) model, based on BiLSTM (Bidirectional Long Short-Term Memory) architecture, which leverages deep sequence embeddings to understand their potential for solving metagenomic challenges. We trained the model on 20 million microbial proteins from the Unified Human Gastrointestinal Protein (UHGP) catalogue, and then demonstrated the utility of the proposed representations on the Bacterial SwissProt database.

In the first part of this paper, we evaluated the embedding space by recreating protein ontology labels from their nearest neighbors in the space. In the second part, we visualized the space using Uniform Manifold Approximation and Projection (UMAP)[42], which allowed for a better interpretation of the evaluation results. As an extension, we built an interactive visualization of the space, which is available at https://protein-explorer.ardigen.com. Reproducing this process on a large collection of proteins, such as metagenomic datasets, can facilitate their exploration. Finally, we showed how ArdiMiPE representation goes beyond sequence similarity on short-chain fatty acid kinases.

The use of deep protein representations can be beneficial in metagenomic studies, providing advantages over sequence homology-based approaches, both in terms of computation time and annotation coverage. A deep model can create a global protein space, strongly related to protein function, by making use of unannotated protein sequences in an unsupervised manner. Representing proteins in this space enables their rapid analysis, using a wide range of traditional methods operating on vector spaces and facilitates tasks, such as classification, clustering or semantic search. ArdiMiPE assigns functions based on learned abstract patterns that combine, but also go beyond protein sequence and domain architecture. The use of representation space enables ArdiMiPE to group even sequentially distant proteins into clusters of proteins sharing similar functions. The speed and accuracy of ArdiMIPE is promising in applied settings, such as predicting the mode of action of bacteria and their pathways, new therapeutic design, etc., where the time and ease of use of computational tools may provide accurate interpretations of generated data in a matter of minutes.

Results

Alignment-free deep protein embeddings represent structure- and function-related ontologies

Metagenomic data may generate an amount of information on the order of tens of millions of reads, which may be assembled into millions of protein sequences. For traditional sequence homology or profile-based approaches, this amount of data is manageable, but requires significant computing power. For deep learning, on the other hand, such a large amount of data provides an opportunity to be exploited for training and assures a robust representation of analysed sequences.

To build the deep representation, we trained the BiLSTM model on the Unified Human Gastrointestinal Protein catalog (UHGP), which contains 625 million microbial protein sequences clustered with MMseqs2 linclust into 20,239,340 representative sequences at 95% amino acid identity[14,43]. From the trained model, we take a hidden-state vector that acts as a protein representation (see Methods and Fig. 1A).

Although the representation was trained on metagenomic data, we need proteins with a specified function and origin to validate it. Therefore, for our analysis, we used bacterial proteins from the SwissProt database clustered into 201,622 representatives at 97% sequence identity. SwissProt is a reliable source, linking proteins to many ontologies that enable a multilevel description of sequences (e.g. Table 1). For simplicity, we call this collection of proteins Bacterial SwissProt (see Methods). We generated embeddings for all Bacterial SwissProt sequences using the ArdiMiPE model and then reduced the 2,048 dimensional vectors with PCA (Principal Component Analysis) to 50 dimensions. Such a representation is used in all our analyzes (Reduced Embeddings in Fig. 1B). Rationale for selected parameters can be found in the Methods section.

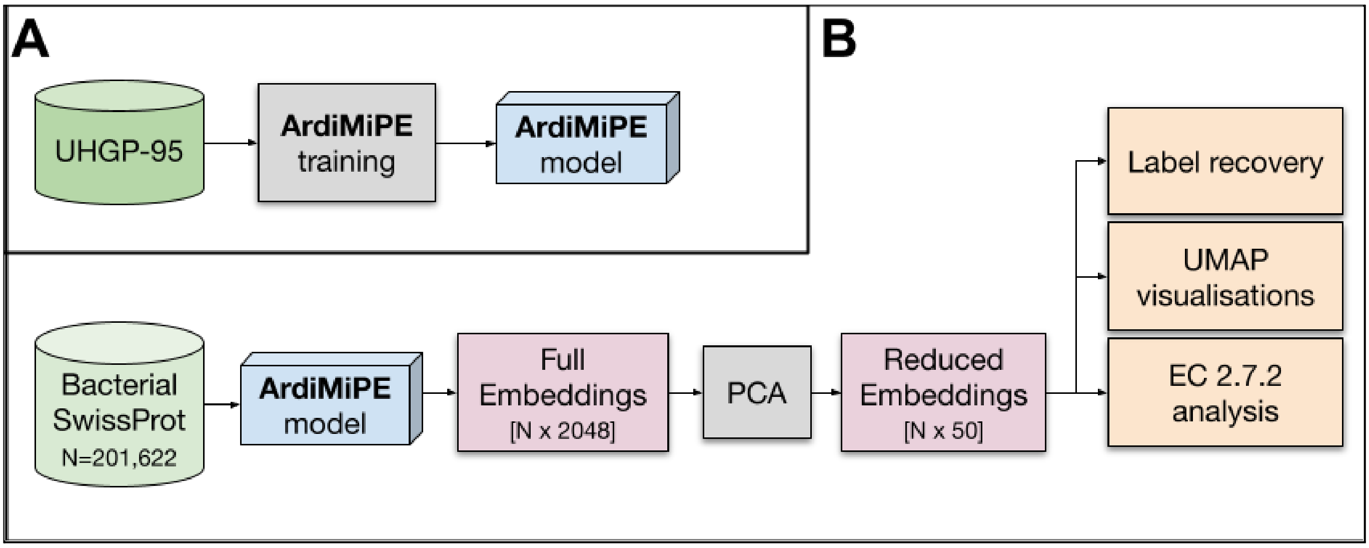


Figure 1. Workflow showing the training of the ArdiMiPE model and its subsequent use in analyzes.

To get a deeper understanding of the type of information encoded within deep representations, we created an evaluation task of recovering the label of a given protein from the labels of its nearest neighbours for a cross-section of various ontologies. If the label is correctly recovered, it indicates that the representation is consistent within this ontology (Fig. 2). This paper focuses on investigating the representation and its features, not aiming at creating a universal label predictor.

To evaluate the consistency of the representation, we selected a number of ontologies from Bacterial SwissProt, related to Function, Structure, or organism of Origin (Table 1). The ontologies significantly vary in the number of classes and Bacterial SwissProt coverage. Hence, the recovery task for each ontology may have a varying degree of difficulty. For this reason, we compared ArdiMiPE embeddings to k-mer (n-gram) embeddings (see Methods) - general sequence-based representation, which is often used as a baseline for sequence embedding methods[35,40].

In order to measure label recovery performance of our and baseline representations, we used a cross-validation-based approach. We removed labels of 20% randomly selected proteins in the dataset. Next, we trained a k-Nearest-Neighbor (kNN) classifier. Then, for every protein without the label, we predicted its label based on a majority vote from k nearest neighbors (k=51). We repeated this procedure 5 times.

Many proteins are annotated with more than one label within each ontology (for example, a protein may have multiple Pfam domains). To overcome this challenge, we used the Intersection over Union (IoU) metric. IoU is the ratio between the correctly predicted labels and the union of all predictions with all ground-truth labels for given protein (1). IoU ranges between 0 and 1, where 1 means perfect label recovery. For single-label tasks, IoU reduces to accuracy.

$$IoU = \frac{|\, predicted \cap ground{-}truth \,|}{|\, predicted \cup ground{-}truth \,|} \quad (1)$$

Table 1. Description of Bacterial SwissProt ontology databases. *For the label recovery task, we used a number of ontologies that can be assigned to a protein. These ontologies are based on 3D protein structure (SUPFAM, Gene 3D), domains (PFAM, InterPro), function (GO, KO, EC numbers) or provide information about organism of origin (taxonomy)*

Database	Category	Description	Bacterial SwissProt	
			#proteins	#classes
SUPFAM	Structure	SUPFAM associates sequence families from Pfam with SCOP structural families using profile matching to produce sequence superfamilies of known structure.	147,137	989
GENE 3D	Structure	GENE 3D contains protein domain assignments for sequences from all of the major sequence databases. Domains are predicted using a library of representative profile HMMs, derived from CATH superfamilies or directly mapped from structures in the CATH database.	116,919	1,173
InterPro	Sequence and domain	InterPro brings together 11 protein family databases (CATH-Gene3D, HAMAP, PANTHER, Pfam, PRINTS, ProDom, PROSITE Patterns, PROSITE Profiles, SMART, SUPERFAMILY, and TIGRFAMs). Each database provides a specific signature i.e. position-specific score matrices, hidden Markov models and profiles etc. to increase the sensitivity of protein classification.	198,677	12,244
KO (KEGG Orthology)	Function	KO is a database of molecular functions. Each molecular function is represented in terms of a manually defined functional ortholog that together create molecular networks (pathways). Each functional ortholog is defined from experimentally characterized genes and proteins in specific organisms, which are then used to assign orthologous genes in other organisms, based on sequence similarity.	177,018	6,614
GO (Gene Ontology)	Function	GO is a controlled terminology that can be used to consistently and structurally identify genes and gene products. The GO terms are organized within a directed acyclic graph (DAG), and each GO term has a described relationship to one or more other terms in the same domain (i.e. biological process, molecular function, or cellular location).	192,990	5,799
eggNOG	Function and taxonomy	eggNOG is a database of orthology relationships, gene evolutionary histories and functional annotations. It is built on the concept of OGs (orthologous groups) that are the result of a non-supervised analysis of thousands of genomes and relationships between all their genes.	162,261	15,932
EC number	Function	EC numbers are a manually assigned nomenclature that describes enzymes, based on the chemical reactions they catalyse.	193,198	3,005
Pfam	Sequence and domain	Pfam is a database of protein families and domains. Each Pfam family has a seed alignment that contains a representative set of sequences for the entry. This alignment is used to build a hidden Markov model profile and the profile is being searched in the sequence database called pfamseq using the HMMER software.	120,184	5,551

Taxonomy: Order	Taxonomy	Uniprot uses the NCBI taxonomic database to assign taxonomic identifiers to nucleotide sequences.	200,536	132
Taxonomy: Family			198,996	274
Taxonomy: Genus			200,615	660

Embedding performance on structure-, function- and taxonomy-related protein ontologies

Despite a varying number of classes in each task, the results from all ontologies unrelated to taxonomy were similar (Fig. 2). This suggests a comparable degree of difficulty between them, possibly due to the correlations amid labels (e.g. KOs are correlated with Pfam domains). ArdiMiPE and k-mers based representation performance drops for taxonomic labels i.e. genus, family, and order (Fig. 2).

Structure- and function-related ontologies

The representation generated by ArdiMiPE performs best at recovering labels from ontologies based on protein structures (Gene3D, SUPFAM), while function- or domain-related ontologies obtained a slightly lower metric. These results indicate that the representation space primarily encodes the structure of proteins and secondarily the function of the protein, which is a structure derivative[44]. ArdiMiPE's ability to approximate the protein structure based on its sequence may result from the nature of the model's training - during this process, the model has to predict which amino acid will be better fitted to the rest of the sequence. ArdiMiPE learns to predict the next amino acid in the sequence given all previous amino acids, but based on benchmarking results, we can conclude that ArdiMiPE uses secondary and tertiary structure of the protein that are contributing more to the function than sequence alone. It is understandable as domains and motifs provide more information about protein function than protein sequence alone.

Function annotation to proteins is a long-standing and open challenge[15]. The most straightforward methods are based solely on sequence alignment (with BLAST being the most prominent example). The main idea behind this approach is the hypothesis that proteins of similar sequences are usually homologous and thus, have a similar function. However, this hypothesis is frequently misleading, as closely related proteins do not always share the same function and it is difficult to pick a universal sequence similarity threshold that would delineate protein functions. In order to overcome those limitations, more sophisticated approaches and databases were developed (PROSITE, PFAM). These approaches focus on smaller subunits present in proteins, such as motifs or domains, as they are more robust determinants of protein function.

Taxonomy-related ontologies

In both k-mer and ArdiMiPE space, predicting taxonomic origin is more difficult than functional characteristics (Fig. 2). K-mers are based solely on protein sequence, so they can be more biased towards the organism of origin, rather than the functional aspects of proteins[45,46], https://www.eb.tuebingen.mpg.de/protein-evolution/protein-classification/). However, in the process of evolution, genes undergo displacement, duplication, and horizontal transfer, which causes an increase in the taxonomic distance between protein and its organism of origin. In the practice of taxonomy, assignment genes need to be universal, conserved and not undergo frequent horizontal gene transfer [47]. As such, a random protein is a poor indicator of the organism of origin. We also see this considering both models' performance on EggNOG ontology. It is combining information about function and taxonomy, achieving results between function- and taxonomy-related ontologies.

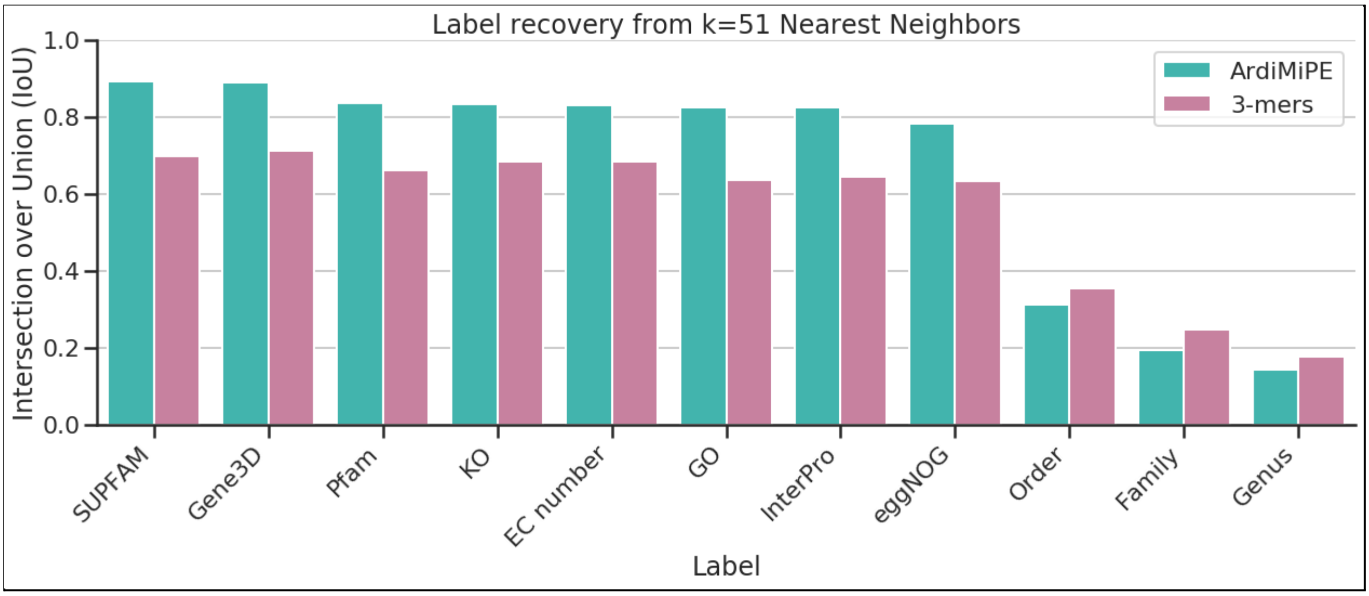


Figure 2. The degree of correctness in the recovery of labels using deep (ArdiMiPE) or k-mer-based representations. *The measure of the recovery is Intersection over Union (IoU) between original labels and a set of labels from 51 nearest neighbors.*

Low dimensional representation of protein sequence space goes beyond sequence similarity

Representations learned by deep models are information-rich, but more difficult to understand due to high dimensionality of the embedding (proteins are represented in a 2048 dimensional space). Further reduction in dimensionality with UMAP (down to two dimensions) allows us to plot and visually interpret the embedding space built by the ArdiMiPE model.

Deep embedding model creates a functionally structured representation space

To better understand which proteins were the easiest to recover based on the embedding, we defined Recovery Error Rate as *1 - average IoU* metric obtained on each protein across all ontologies. The use of this metric enabled us to localize regions with low & high Recovery Error Rates, which we visualized on the UMAP plot (see Fig. 3). In Fig. 3A, we show that proteins with low Recovery Error Rates are located in smaller clusters, while proteins with high error rates are concentrated in the center part of the UMAP visualization.

To investigate the functional structure of the representation space, we overlay it with labels defined by Kegg Orthology ID (KO) (see Fig. 4). The proteins that do not have a KO assigned are colored in grey - we see that they are placed in the central part of the plot. Most of the proteins are clearly clustered by their functional annotation. Furthermore, by focusing on specific space locations, we can see that close KO clusters share other functional features: domains (Fig 4A & 4B), EC number class (Fig 4A & 4D), or structural and molecular features (Fig 4E). It shows that ArdiMiPE representation does not focus only on one functional ontology, but rather on an abstract protein function defined on many levels. The visualisation explains the high label recovery results and expands analogous analysis conducted on a smaller scale with only 25 COGs[35]. Compared to the k-mer based representation, ArdiMiPE is significantly more structured (See Supplementary Fig. 1).

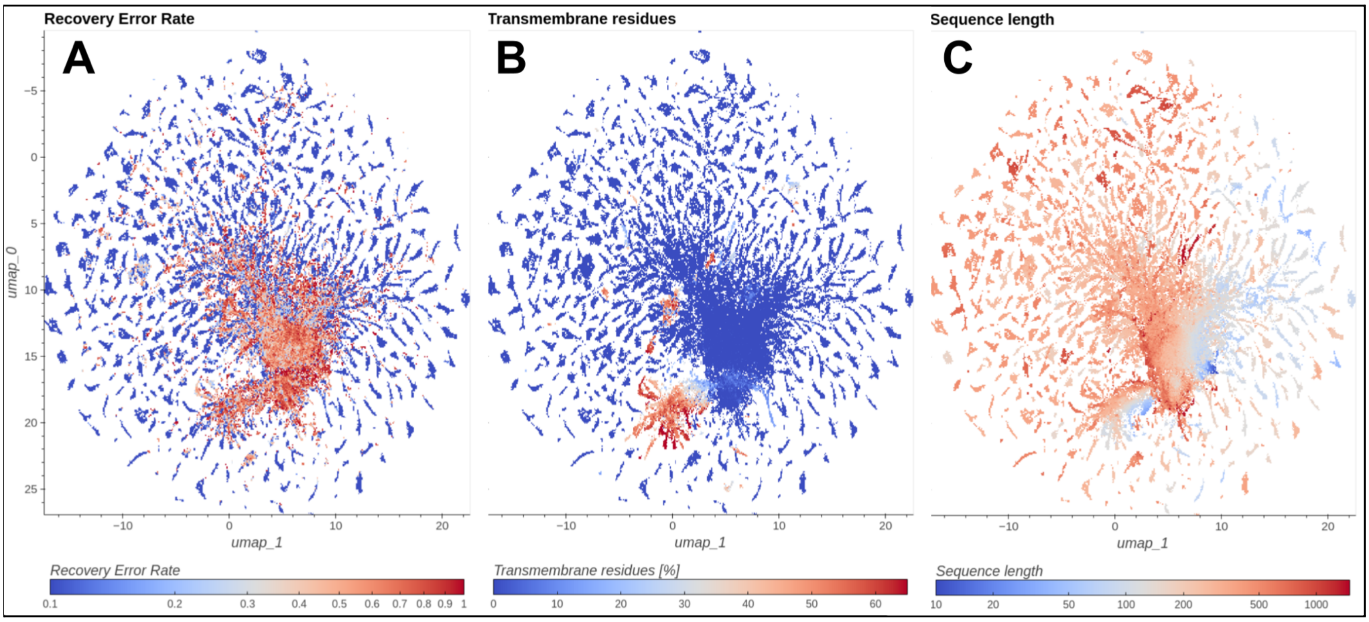


Figure 3. UMAP visualization of Bacterial SwissProt embeddings. *(**A**) Proteins colored by Recovery Error Rate. (**B**) Proteins colored by percentage of transmembrane residues. (**C**) Proteins colored by sequence length.*

We hypothesize that the regions of high Recovery Error Rate are occupied by rare proteins. Rare proteins form small functional classes in Bacterial SwissProt, and the smaller the functional class is, the more difficult it is to predict the label based on its neighbors. Additionally, their insufficient representation in the training set makes it difficult to model their sequences, as the embedding model can learn certain patterns only if they are shared by a sufficient number of proteins in the training dataset. Indeed, we observed that the Recovery Error Rate is negatively correlated (r=-0.715, N=200,115) with the log-average size of the functional class the protein belongs to (See Supplementary Fig. 2). Moreover, we noticed an increased frequency of the occurrence of words: *'Uncharacterized'*, '*Putative*' and '*Probable*' in SwissProt descriptions of error-causing proteins (25% for Error Score = 1 vs. 4% for Error Score = 0, See Supplementary Fig. 3), indicating less characterized proteins.

Short and transmembrane proteins

The embedding model is very sensitive to the length of the protein (Fig. 3C) and a significant number of short proteins is present in the central, lesser understood part of UMAP visualization. Short proteins (≤50 residues), underestimated for a long time, gained interest in recent years when it was discovered that they are involved in important biological processes such as cell signaling, metabolism, and growth[48]. The presence of a high Recovery Error Rate region might be a result of insufficient information on small proteins, which are still underrepresented in databases. Following Sberro et al., based on the NCBI GenPept database, over 90% of small protein families have no known domain and almost half are not present in reference genomes[49]. Additionally, due to their short length, they remain unnoticed by most bioinformatics tools that use a length threshold to minimize erroneous predictions.

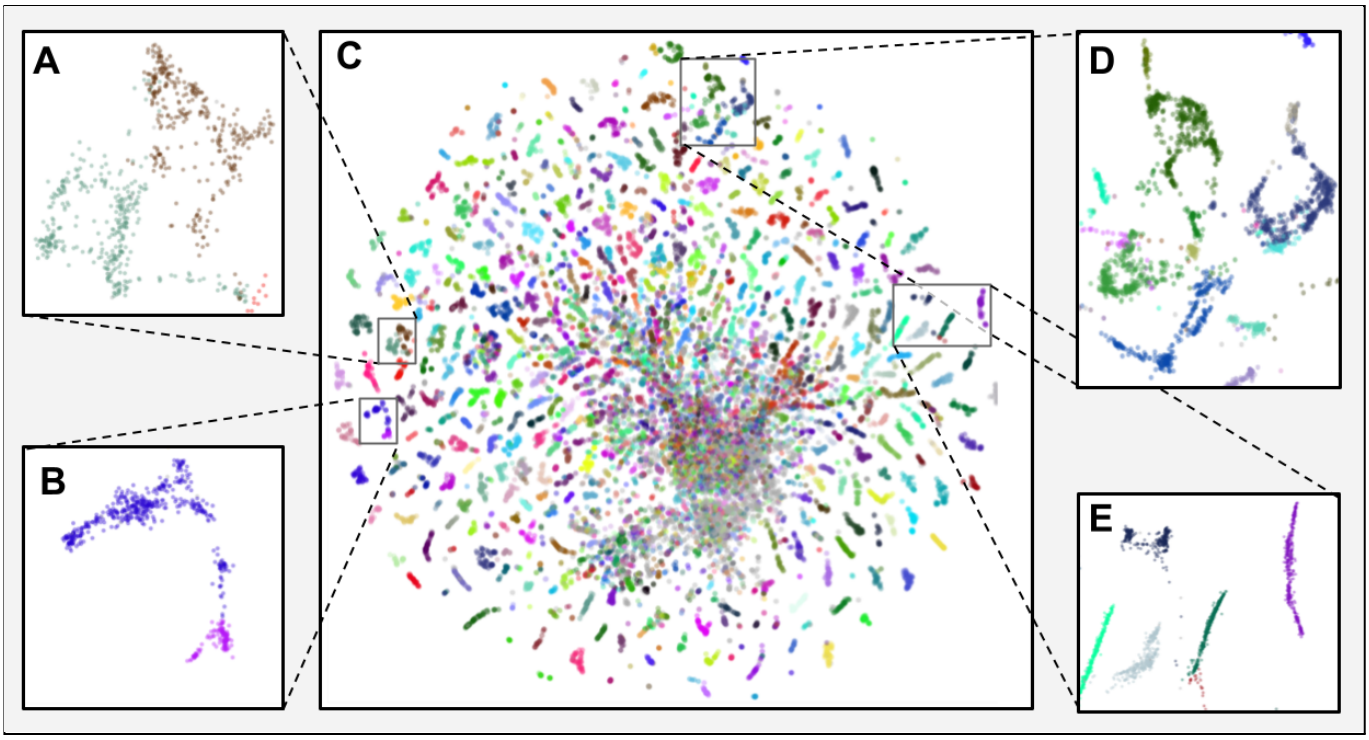


Figure 4. ArdiMiPE UMAP projection of Bacterial SwissProt colored by KO.

(A) *transferase proteins that share the same Pfam domain and belong to the EC 2.5.1 class - UDP-N-acetylglucosamine 1-carboxyvinyltransferase (K00790) in dark green, 3-phosphoshikimate 1-carboxyvinyltransferase (K00800) in brown*

(B) *GTP binding proteins sharing Pfam domains - Elongation Factor G (K02355) in purple, Peptide chain release factor (K02837) in pink.*

(C) *all Bacterial SwissProt proteins*

(D) *proteins that belong to the tRNA ligases class (EC 6.1.1) - Cysteine (K01883), Arginine (K01887), Glutamate (K01885), Glutamine (K01886), Glycine (K01880), Valine (K01873), and Isoleucine (K01870)*

(E) *ribosomal proteins - 30S ribosomal protein S1 (K02961) in light green, 50S ribosomal protein L14 (K02874) in light blue, 50S ribosomal protein L36 (K02919) in black, 50S ribosomal protein L35 (K02916) in dark green, and 50S ribosomal protein L15 (K02876) in purple.*

The whole space can be interactively explored in our application (https://protein-explorer.ardigen.com).

Transmembrane proteins constitute a very numerous group, containing approx. 30% of all known proteins. Unlike globular proteins, its members are on average larger and must exhibit a pattern of hydrophobic residues to fit into the cell membrane[50]. In order to define transmembrane proteins we used a transmembrane score (a percentage of transmembrane residues) adopted from Perdigão et al.[51]. In Figure 3B, we can see that the ArdiMiPE can easily separate transmembrane proteins, which is in line with previous research on deep protein representations[39,52]. However, part of transmembrane proteins lie within the high-recovery error region of the UMAP plot. Despite substantial pharmacological and biological relevance, they are less understood and underrepresented in databases, as structural experiments on them are difficult to conduct.

We believe that the deep embedding model trained on a more general catalog of metagenomic proteins (UHGP) is less biased towards well-known model organisms than SwissProt, hence, better suited for rare, short or transmembrane proteins.

The utility of a low-dimensional representation

We can see that ArdiMiPE space encompasses information about protein function based on secondary and tertiary structure, length and biochemical properties (e.g. transmembrane proteins). The fact that the ArdiMiPE allows us to understand more general features means that even when we do not know the features of a certain protein (e.g. due to the lack of experimental data) we can infer them by transferring annotations from the nearest known proteins in the space. Increasing effectiveness of predicting general features is one of the directions of further research. However, a low-dimensional representation analyzed here may not be able to capture all of the complex aspects of structure, function and relationships between proteins.

A sample use case - phosphotransferases (EC 2.7.2)

To demonstrate the use of embedding representation in a real-life scenario, we used a group of phosphotransferases. We have chosen them due to their importance in maintaining the human gut microbiome homeostasis. Acetate, butyrate, and propionate kinases are especially crucial in the process of forming short-chain fatty acids (SCFAs), which are secondary metabolites produced by gut microflora that play a vital role in maintaining intestinal homeostasis[53]. SCFAs are produced in the colon by bacteria during the fermentation of resistant starch and non-digestible fibers such as pectin or inulin. The most abundant SCFAs produced by intestinal bacteria are butyrate, propionate and acetate. Butyrate is produced by Firmicutes phylum, while Bacteroidetes phylum produces acetate and propionate[54–56]. SCFAs play an important role in maintaining gut homeostasis and their lowered level is often observed in patients suffering from irritable bowel diseases (IBD) such as Crohn's disease and ulcerative colitis. SCFAs serve as an important fuel for intestinal epithelial cells and participate in preserving gut barrier integrity. Moreover, recent findings indicate their role in energy metabolism (lipid metabolism), immunomodulation, regulation of intestinal epithelial cells, proliferation and cancer protection. Although promising, the research has been conducted mainly on murine or *in vitro* models, thus the results have to be interpreted with caution[54–56].

Proteins classified as phosphotransferases were chosen based on their EC number. We decided to use this annotation as EC numbers are a manually assigned nomenclature that describes enzymes based on the chemical reactions they catalyse. Their hierarchical structure allows for a fine-grained analysis. Proteins described by EC 2.7.2 class represent phosphotransferases with a carboxyl group as an acceptor. Even though there are 14 sub-subclasses in EC 2.7.2 subclass, we used only 8 of them, as our dataset focused on bacterial proteins and should not contain proteins described by other 6 sub-subclasses (Supplementary Table 1). Sub-subclasses used in this analysis are EC 2.7.2.1 (acetate kinase), EC 2.7.2.2 (carbamate kinase), EC 2.7.2.3 (phosphoglycerate kinase), EC 2.7.2.4 (aspartate kinase), EC 2.7.2.7 (butyrate kinase), EC 2.7.2.8 (acetylglutamate kinase), EC 2.7.2.11 (glutamate 5-kinase) and EC 2.7.2.15 (propionate kinase).

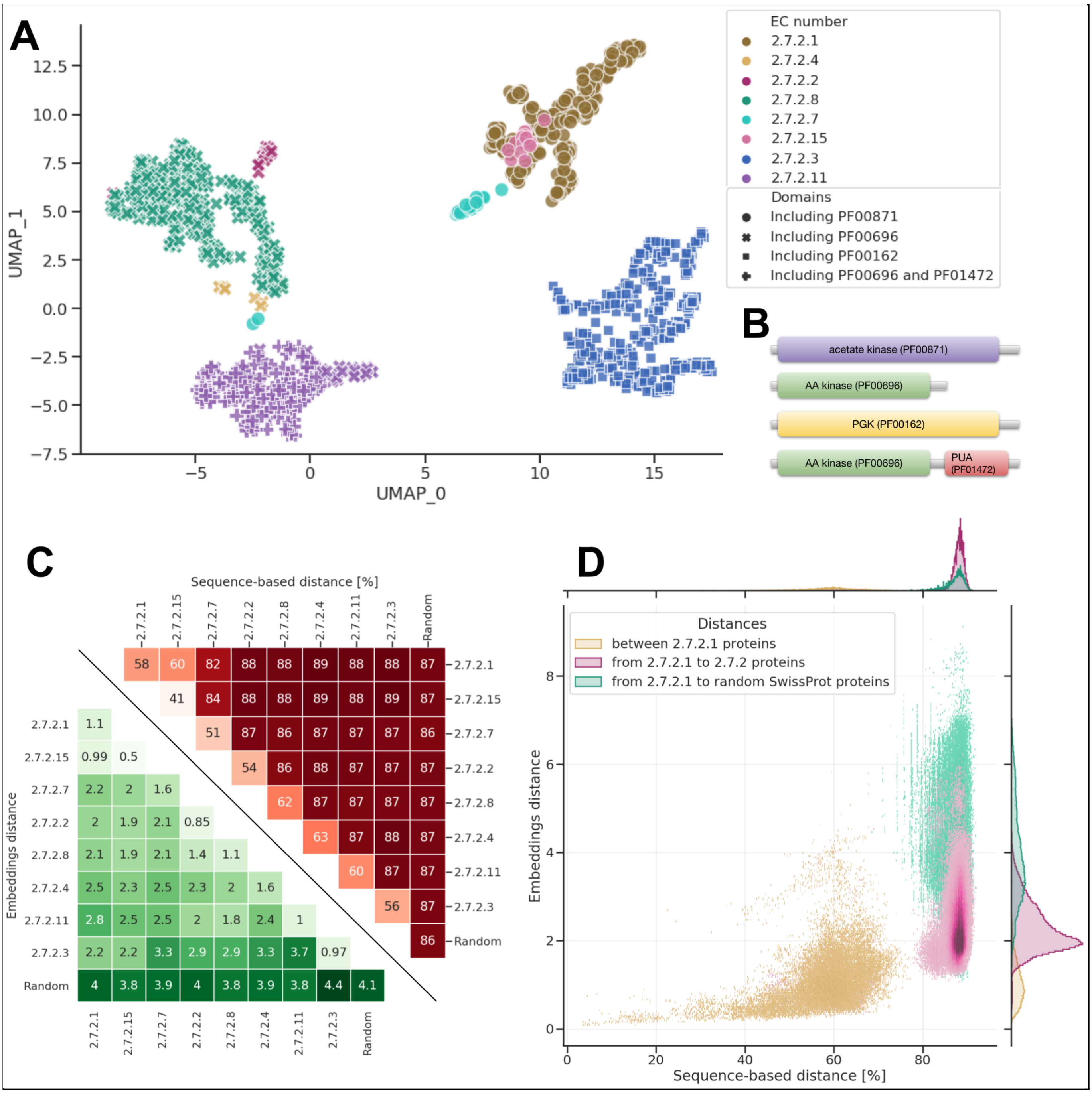
We examined the domain architecture of EC 2.7.2 proteins using the Pfam database of protein domains and families. The distribution of domains in a protein, called the domain architecture, is the main structure that defines a protein's function. The Pfam database is an extensive collection of protein families, represented by multiple sequence alignments and corresponding Hidden Markov Models (HMMs). We found that four domain architectures were dominant among analysed proteins. 31% of analyzed proteins contained one amino acid kinase domain (PF00696). Subsequently, 29% of proteins had one phosphoglycerate kinase domain (PF00162), 20% contained one acetate kinase domain (PF00871), and 18% of proteins had two coincident domains PF00696 & PF01472, i.e., amino acid kinase domain and PUA domain.

In total, we study 1,302 proteins exhibiting eight unique specific functions (ECs) and four distinct domain architectures. Different domain architectures suggest that these proteins have different amino acid sequences and would be difficult to identify as similar with baseline bioinformatic methods based on sequence similarity alone.

To investigate how accurately the embedding representation reflects the functional relationships between the proteins, we visualized them using UMAP (Fig. 5A). Almost all proteins were grouped according to their domain architecture, and proteins with similar domain architectures, such as proteins having only PF00696 domain and proteins having two domains PF00696 & PF01472, were also placed closer to each other. Despite clear domain-based grouping, proteins that share the same domain architecture, but catalyze different chemical reactions, are separated. The only exceptions are EC 2.7.2.1 and 2.7.2.15. One possible explanation for this exception is that these two enzymes can share substrates for their activity. Acetate kinases (EC 2.7.2.1) can accept propionate as an alternative substrate, and propionate kinases (EC 2.7.2.15) can accept acetate. Moreover, both EC 2.7.2.15 and EC 2.7.2.1 play essential roles in the production of propionate in bacteria[57]. The only inconsistency we can note are two 2.7.2.7 proteins that were placed far from their counterparts.

In conclusion, ArdiMiPE's representation reflects the functional similarities between proteins that are based on domain architecture (PFAM domains) or enzymatic activity (EC number). This emphasizes the significant advantage of deep embeddings, as they do not only focus on single, human-created ontology, such as e.g., EC numbers, but rather fuse all information to characterize proteins on multiple levels. It combines the strengths of approaches that focus on motifs, domains (PFAM), and 3D structure (GENE 3D) to understand protein function space comprehensively.

To better understand the differences between sequence-based distance and ArdiMiPE embeddings, we compared the Euclidean distance between EC 2.7.2 proteins and randomly chosen 500 proteins from the Bacterial SwissProt dataset. As a baseline, we selected sequence-based distance calculated with Clustal Omega[58]. The embedding-based distances within and between the EC 2.7.2 subclasses are smaller than to randomly selected proteins, which do not hold for the sequence-based distance (Fig. 5B & 5C). This proves that ArdiMiPE can go beyond sequence similarity and find relations between proteins with significantly different sequences and domain architectures. This property enables searching for proteins that are similar on a more abstract level and in the future may improve the annotation coverage of microbial proteins.



Sequence-based distance [%]

	2.7.2.1	2.7.2.15	2.7.2.7	2.7.2.2	2.7.2.8	2.7.2.4	2.7.2.11	2.7.2.3	Random
2.7.2.1	58	60	82	88	88	89	88	88	87
2.7.2.15		41	84	88	88	89	88	89	87
2.7.2.7			51	87	86	87	87	87	86
2.7.2.2				54	86	88	87	87	87
2.7.2.8					62	87	87	87	87
2.7.2.4						63	87	88	87
2.7.2.11							60	87	87
2.7.2.3								56	87
Random									86

Embeddings distance

	2.7.2.1	2.7.2.15	2.7.2.7	2.7.2.2	2.7.2.8	2.7.2.4	2.7.2.11	2.7.2.3	Random
2.7.2.1	1.1								
2.7.2.15	0.99	0.5							
2.7.2.7	2.2	2	1.6						
2.7.2.2	2	1.9	2.1	0.85					
2.7.2.8	2.1	1.9	2.1	1.4	1.1				
2.7.2.4	2.5	2.3	2.5	2.3	2	1.6			
2.7.2.11	2.8	2.5	2.5	2	1.8	2.4	1		
2.7.2.3	2.2	2.2	3.3	2.9	2.9	3.3	3.7	0.97	
Random	4	3.8	3.9	4	3.8	3.9	3.8	4.4	4.1

Figure 5. Visualization of functional clusters in the ArdiMiPE embedding space.

(***A***) *Deep embeddings of EC 2.7.2 proteins visualized with UMAP.*

(***B***) *Domain architecture of EC 2.7.2*

(***C***) *The mean distance between EC 2.7.2 proteins and 500 random proteins from the SwissProt space with distinction between embedding-based distance (green) and ClustalO distances (red). Values for both methods were calculated as averages of pairwise distances between all proteins within given clusters. The mean embedding-based distance between EC 2.7.2 proteins is significantly smaller compared to the distance between EC 2.7.2 proteins and 500 random proteins. Not only proteins from the same cluster group are closer to each other but also proteins from different EC 2.7.2 clusters are located significantly closer to proteins from other EC 2.7.2 clusters than to random proteins. Mean distance between proteins calculated using ClustalO does not reflect the clear separation between EC 2.7.2 and random proteins. Mean ClustalO distance between proteins from the same cluster is smaller than between EC 2.7.2 and random proteins, however ClustalO does not bring proteins closer from different EC 2.7.2 clusters.*

(***D***) *Comparison of embedding-based and sequence-based distance (ClustalO) to EC proteins 2.7.2.1. The distances were divided into those within the protein group EC 2.7.2.1, from EC 2.7.2.1 to other EC 2.7.2 proteins, and from EC 2.7.2.1 to randomly selected proteins. The embedding-based, as opposed to the sequence-based distance, differentiates the distances from EC 2.7.2.1 to other members of EC 2.7.2 and from EC 2.7.2.1 to random proteins.*

Discussion

The human microbiome plays a crucial role in human health, and changes in its composition can be related to various diseases, such as diabetes, cancer, or psychiatric disorders. To fully understand the complex relation between the microbiome and human health, it is necessary to look not just at the taxonomic level but also at a functional level. Despite various approaches to retrieve protein functions[59,60], a large portion of microbial proteins remain functionally uncharacterized. This paper presents a novel approach of using the Bidirectional LSTM model to visualize and contextualize the microbial protein space. We show that our model - ArdiMiPE - accurately represents protein features related to structure and function, overcoming some limitations of standard bioinformatics methods such as HMMER or BLAST.

ArdiMiPE creates an abstract, numerical representation of proteins in an embedding space. This embedding encodes information from various protein ontologies and combines knowledge on protein structure and function, overcoming the limitations of methods based on sequence similarity. At the same time, generating the embedding for a given protein is much more efficient computationally than using bioinformatic methods. On top of that, the embedding is also more suitable for a large range of further downstream algorithms, such as classification, clustering and visualization. Combining embeddings with a dimensionality reduction method, such as UMAP, may enable creating a reference protein map and facilitate protein research.

One of the significant challenges that any data-driven solution must face is data bias. We believe that using a catalog of metagenomic proteins (UHGP) for training made the model less biased towards well-known model organisms. Despite this, model validation required the use of experimentally verified data, which limited the scope of our validation to well-known proteins and prevented genuine validation on small or transmembrane proteins. We assume that with the growing interest in these proteins, their presence in the databases and number of their annotations will increase, which will allow for a more thorough validation.

We are witnessing rapid progress in both the deep learning field and in metagenomics, which generate massive amounts of data. We believe embedding models are an attractive alternative to database-bound, computationally intensive methods unsuitable for such influx of data. An appealing approach would be to join the strengths of computationally-cheap embedding models with other computational technologies that can accurately predict the features of individual genes (for example: protein 3D structure using AlphaFold[18]) and finally perform experimental validation on most promising targets. An approach such as ArdiMiPE enables such efficient contextualization of metagenomic data and may be used to better understand the microbiome for health.

Methods

ArdiMiPE Training

In the training, we took advantage of the Unified Human Gastrointestinal Protein catalog clustered at 95% sequence identity (UHGP-95) to limit the impact of the most common sequences. Moreover, it was proven that using unclustered sequences from the dataset does not increase the model quality[35]. UHGP-95 contains exactly 19,228,304 protein sequences, from which we randomly selected 5% to track training progress (validation set) and set aside another 5% for the final model evaluation (test set).

The rest of the data (18,266,888 sequences) was used to train the model. All proteins were clipped to 1,500 amino acids.

We used a 3-layered Bidirectional LSTMs (BiLSTM) model with 1024 hidden units in each layer. We have chosen the LSTM architecture as it gave the best results in Remote Homology detection in the TAPE benchmark[34] and achieved superior performance over Transformer-based architecture in the ProtTrans benchmark[39].

The model was trained by the AdamW optimizer for 225,331 weight updates with a mini-batch of size 1024, which corresponds to 12 epochs and approximately 48 hours on 4 Tesla V100 GPUs. The learning rate was set to 1e-3, except the first 8,000 steps that were used as a warmup. The process was implemented in the PyTorch library, based on the TAPE benchmark[34] repository (https://github.com/songlab-cal/tape).

Computing embeddings

To obtain a vector representation of a protein (embedding) from the BiLSTM model, we extracted vectors of hidden states for each amino acid and averaged them. This is in contrast to natural language processing practice, which uses the hidden state vector corresponding to the last word (here it would be the last amino acid) rather than the average representation of all words. However, there is evidence suggesting the superiority of averaged presentation in the field of protein processing[40]. This may be due to the fact that proteins are usually much longer than sentences, and LSTM-based models cannot fit the whole amino acid sequence in just one state.

Bacterial SwissProt

For evaluating the properties of the embedding space, we used the UniProtKB/Swiss-Prot 2019_02 database with 562,438 protein entries. For every entry, we parsed taxonomy lineage and functional labels (Table 1). Only proteins from the Bacteria domain were selected, leaving 331,523 proteins.

To remove near-identical protein sequences, we deduplicated the remaining set using *mmseq2 easyclust* with an identity threshold set to 97% and coverage set to 0.8. Removing duplicates ensured no cliques in the kNN graph, which we used in the kNN label recovery and UMAP visualizations. Cliques would lead to trivial solutions during kNN classification and “lonely islands'' in UMAP visualizations.

After the deduplication step, we obtained 201,622 proteins, and this set we named Bacterial SwissProt.

K-mer representations

For a general sequence-based baseline representation, we used the bag of k-mers method[61], which produces embedding for a protein by the following procedure: (a) generate all possible k-mers (subsequences of length k) from protein sequence, (b) count occurrences of each possible k-mers in the sequence, (c) sort counts alphabetically by k-mers sequence. Sorted counts form a vector representing the sequence.

Higher k leads to more specific representation but increases dimensionality, which is equal to the number of all possible k-mers ($N=20^k$). In our work, we choose k=3, which resulted in 8,000-dimensional vectors.

Label recovery

For the analysis, we used the Bacterial SwissProt described above. We generated deep and k-mer representations for each protein. Next, we reduced the dimensionality of both representations to 50 using the Principal Component Analysis (PCA) algorithm (Fig. 1B).

We narrowed down the set of analyzed proteins to only those with assigned labels in given ontology for each ontology analyzed. We divided these sets of proteins into five equal parts to estimate recovery efficiency through 5-fold cross-validation. For every fold, we constructed a kNN graph (https://github.com/lmcinnes/pynndescent) of the data from the four remaining folds. The graph was then used to predict classes for each protein in the fold, by querying the nearest proteins (N=51) and propagating their labels as a prediction. As the protein can be assigned to many classes (multi-label classification), we used the Intersection over Union (IoU) metric. A higher number of neighbors (N) taken into account causes the results to extend beyond the immediate neighborhood, which usually contains highly similar proteins. At the same time, the larger N is, the more challenging it is to predict a label for small classes, and it even becomes impossible for the classes smaller than N / 2 proteins. We choose 51 to balance these two properties.

UMAP visualizations

To visualise protein embedding space, we further reduced dimensionality of the PCA Reduced Embeddings with UMAP (Uniform Manifold Approximation and Projection; https://github.com/lmcinnes/umap), a nonlinear dimensionality reduction method.

UMAP was chosen over another common nonlinear dimensionality reduction method, t-SNE (t-distributed Stochastic Neighbor Embedding), as it preserves more of the global structure with superior run time performance[62].

2.7.2 cluster analysis

Selecting proteins. Proteins assigned to EC 2.7.2 subclass were chosen for analysis of ArdiMiPE performance. In the analysis, we used 8 available EC 2.7.2 sub-subclasses out of 14, as our bacterial dataset lacked proteins described by 6 other sub-subclasses. Sub-subclasses used in this analysis are EC 2.7.2.1 (acetate kinase), EC 2.7.2.2 (carbamate kinase), EC 2.7.2.3 (phosphoglycerate kinase), EC 2.7.2.4 (aspartate kinase), EC 2.7.2.7 (butyrate kinase), EC 2.7.2.8 (acetylglutamate kinase), EC 2.7.2.11 (glutamate 5-kinase) and EC 2.7.2.15 (propionate kinase). We assigned a Pfam ID to each protein using mapping available in SwissProt. 4 domain architectures were found dominant among 1,302 analysed proteins. 31% of analyzed proteins contained one amino acid kinase domain (PF00696), 29% had one phosphoglycerate kinase domain (PF00162), 20% one acetate kinase domain (PF00871) and 18% had two coincident domains (PF00696 and PF01472), i.e. amino acid kinase domain and PUA domain.

We visualized EC 2.7.2 proteins in the same manner as described above in *UMAP visualizations*.

Comparison to sequence (Clustal Omega for distance matrix). To infer about ArdiMiPEs ability to group more closely proteins sharing a function, we compared the distance between EC 2.7.2 proteins and 1,000 randomly chosen proteins from the Bacterial SwissProt database. We wanted to analyse if ArdiMiPEs distance between proteins is compatible with corresponding amino acid sequence distance. ArdiMiPEs distance was calculated as an Euclidean distance between 50 PCA components. Those 50 PCA components are the result of dimensionality reduction of 2,048 protein embeddings, generated by ArdiMiPE. Sequence distance was calculated using Clustal Omega [63], a bioinformatic tool for multiple sequence alignment. This tool takes a fasta file with unaligned protein amino acid sequences as input and calculates percent of sequence identity between those sequences giving a pairwise distance matrix. The distance measure used by Clustal Omega for pairwise distances of unaligned sequences is the k-tuple measure.

Acknowledgements

We would like to thank various people for their contribution to this project: Monika Majchrzak-Górecka and Paweł Biernat for reviewing the draft of the manuscript and for the helpful discussion; Karol Horosin and Mateusz Siedlarz for technical help with the interactive visualization.

KO, **JM**, **KMZ** are supported by NCRD Fast Track Grant POIR.01.01.01-00-0347/17 awarded to Ardigen S.A. **ZK** and **TK** are supported by the National Science Centre, Poland grant 2019/35/D/NZ2/04353. **AB** is supported by the funds of Polish Ministry of Science and Higher Education assigned to AGH University of Science and Technology.

Author contributions statement

KO - conceived the study, conducted analyses, analysed the data, wrote the paper
ZK - analysed the data, wrote the paper
JM - discussed the results
AB - helped supervise the project
KMZ - discussed the results, helped supervised the project
TK - conceived the study, wrote the paper, helped supervise the project
All authors discussed the results and contributed to the final manuscript.

Code availability

Code used in the analyses is available at https://github.com/ardigen/microbiome-protein-embeddings

Competing interests

The authors declare that they have no competing interests.

References

1. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).

2. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).

3. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

4. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol* **4**, 293–305 (2019).

5. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat Microbiol* **4**, 470–479 (2019).

6. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).

7. Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nat. Med.* **25**, 377–388 (2019).

8. Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Science* **371**, (2021).

9. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat Microbiol* **4**, 623–632 (2019).

10. Nguyen, T. T., Hathaway, H., Kosciolek, T., Knight, R. & Jeste, D. V. Gut microbiome in serious mental illnesses: A systematic review and critical evaluation. *Schizophr. Res.* (2019) doi:10.1016/j.schres.2019.08.026.

11. Cryan, J. F. & Dinan, T. G. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **13**, 701–712 (2012).

12. Jo, J.-H., Kennedy, E. A. & Kong, H. H. Research Techniques Made Simple: Bacterial 16S Ribosomal RNA Gene Sequencing in Cutaneous Research. *J. Invest. Dermatol.* **136**, e23–e27 (2016).

13. Prakash, T. & Taylor, T. D. Functional assignment of metagenomic data: challenges and applications. *Brief. Bioinform.* **13**, 711–727 (2012).

14. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0603-3.

15. Zhou, N. *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *bioRxiv* 653105 (2019) doi:10.1101/653105.

16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

17. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).

18. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning.

Nature **577**, 706–710 (2020).

19. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141–1148 (2019).

20. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins* **87**, 1011–1020 (2019).

21. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

22. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).

23. Li, Y. *et al.* DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* **34**, 760–769 (2018).

24. Zou, Z., Tian, S., Gao, X. & Li, Y. mlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning. *Front. Genet.* **9**, 714 (2018).

25. Kulmanov, M., Khan, M. A., Hoehndorf, R. & Wren, J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**, 660–668 (2018).

26. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).

27. Islam, S. M. S. & Hasan, M. M. DEEPGONET: Multi-label Prediction of GO Annotation for Protein from Sequence Using Cascaded Convolutional and Recurrent Network. *arXiv [cs.CV]* (2018).

28. Duong, D. *et al.* Annotating Gene Ontology terms for protein sequences with the Transformer model. *bioRxiv* 2020.01.31.929604 (2020) doi:10.1101/2020.01.31.929604.

29. Nauman, M., Ur Rehman, H., Politano, G. & Benso, A. Beyond Homology Transfer: Deep Learning for Automated Annotation of Proteins. *Int. J. Grid Util. Comput.* (2018) doi:10.1007/s10723-018-9450-6.

30. Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R. & Atalay, V. DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. *Sci. Rep.* **9**, 7344 (2019).

31. Seo, S., Oh, M., Park, Y. & Kim, S. DeepFam: deep learning based alignment-free method for

protein family modeling and prediction. *Bioinformatics* **34**, i254–i262 (2018).

32. Bileschi, M. L. *et al.* Using Deep Learning to Annotate the Protein Universe. *bioRxiv* 626507 (2019) doi:10.1101/626507.

33. Schwartz, A. S. *et al.* Deep Semantic Protein Representation for Annotation, Discovery, and Engineering. *bioRxiv* 365965 (2018) doi:10.1101/365965.

34. Rao, R. *et al.* Evaluating Protein Transfer Learning with TAPE. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 9689–9701 (Curran Associates, Inc., 2019).

35. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. 622803 (2020) doi:10.1101/622803.

36. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. 2020.09.04.282814 (2020) doi:10.1101/2020.09.04.282814.

37. Villegas-Morcillo, A. *et al.* Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *bioRxiv* 2020.04.07.028373 (2020) doi:10.1101/2020.04.07.028373.

38. Staerk, H., Dallago, C., Heinzinger, M. & Rost, B. Light attention predicts protein location from the language of life. *bioRxiv* (2021).

39. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. 2020.07.12.199554 (2020) doi:10.1101/2020.07.12.199554.

40. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).

41. Madani, A. *et al.* ProGen: Language Modeling for Protein Generation. *bioRxiv* 2020.03.07.982272 (2020) doi:10.1101/2020.03.07.982272.

42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

43. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**,

2542 (2018).

44. Gligorijevic, V. *et al.* Structure-Based Function Prediction using Graph Convolutional Networks. *bioRxiv* 786236 (2019) doi:10.1101/786236.

45. Facco, E., Pagnani, A., Russo, E. T. & Laio, A. The intrinsic dimension of protein sequence evolution. *PLoS Comput. Biol.* **15**, e1006767 (2019).

46. Riveros-Rosas, H., Julián-Sánchez, A., Villalobos-Molina, R., Pardo, J. P. & Piña, E. Diversity, taxonomy and evolution of medium-chain dehydrogenase/reductase superfamily. *Eur. J. Biochem.* **270**, 3309–3334 (2003).

47. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).

48. Miravet-Verde, S. *et al.* Unraveling the hidden universe of small proteins in bacterial genomes. *Mol. Syst. Biol.* **15**, e8290 (2019).

49. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* **178**, 1245–1259.e14 (2019).

50. Koehler Leman, J., Mueller, B. K. & Gray, J. J. Expanding the toolkit for membrane protein modeling in Rosetta. *Bioinformatics* **33**, 754–756 (2017).

51. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15898–15903 (2015).

52. Heinzinger, M. *et al.* Modeling the Language of Life – Deep Learning Protein Sequences. *bioRxiv* 614313 (2019) doi:10.1101/614313.

53. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **19**, 29–41 (2017).

54. Parada Venegas, D. *et al.* Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. *Front. Immunol.* **10**, 277 (2019).

55. Xiao, S., Jiang, S., Qian, D. & Duan, J. Modulation of microbially derived short-chain fatty acids on intestinal homeostasis, metabolism, and neuropsychiatric disorder. *Appl. Microbiol. Biotechnol.* **104**, 589–601 (2020).

56. Alexander, C., Swanson, K. S., Fahey, G. C. & Garleb, K. A. Perspective: Physiologic Importance of

Short-Chain Fatty Acids from Nondigestible Carbohydrate Fermentation. *Advances in Nutrition* vol. 10 576–589 (2019).

57. Palacios, S., Starai, V. J. & Escalante-Semerena, J. C. Propionyl coenzyme A is a common intermediate in the 1,2-propanediol and propionate catabolic pathways needed for expression of the prpBCDE operon during growth of Salmonella enterica on 1,2-propanediol. *J. Bacteriol.* **185**, 2802–2810 (2003).
58. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
59. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
60. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
61. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
62. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4314.
63. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
64. Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B. & Gough, J. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res.* **47**, D490–D494 (2019).
65. Lees, J. *et al.* Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* **40**, D465–71 (2012).
66. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
67. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* vol. 49 D545–D551 (2021).
68. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

69. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

70. Duvaud, S. *et al.* Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* **49**, W216–W227 (2021).

71. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

72. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, (2020).

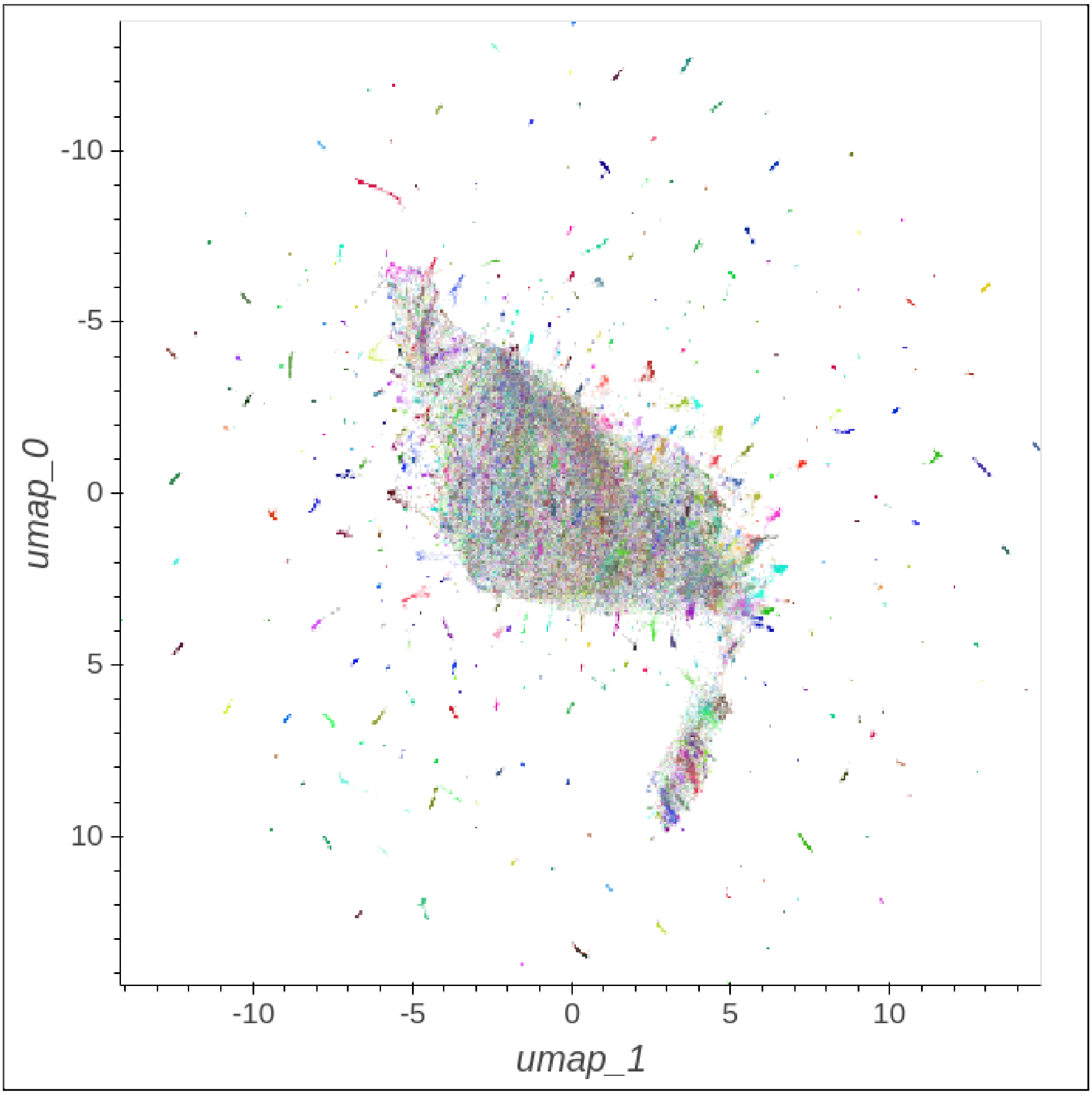
Supplementary information

Supplementary Table 1. *Description (i.e.: EC number, name, domain and number of proteins in Bacterial SwissProt) of the chosen EC 2.7.2 family that was used for a real life use case.*

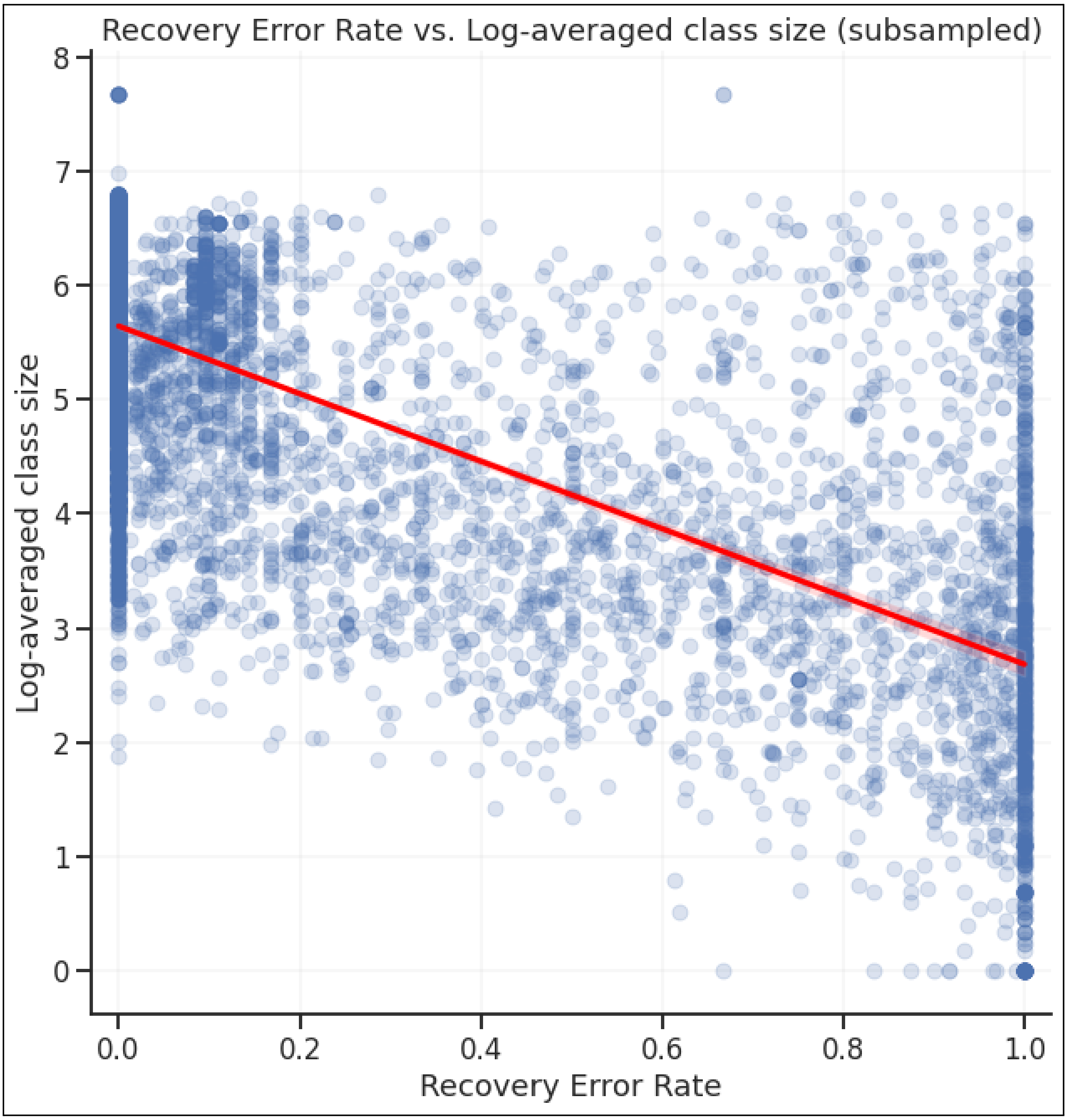
EC number	Name	Pfam domain architecture	Number of proteins
2.7.2.1	Acetate kinase	PF00871	226
2.7.2.2	Carbamate kinase	PF00871	18
2.7.2.3	Phosphoglycerate kinase	PF00162 PF00162||PF00121	377 1
2.7.2.4	Aspartate kinase	PF00696 PF00696||PF01842||PF13840 PF00696||PF13840 PF00696||PF01842 PF00696||PF00742||PF03447 PF00696||PF01842||PF13840||PF00742||PF03447 PF00696||PF13840||PF00742||PF03447	8 13 4 1 4 2 1
2.7.2.5	Transferred entry: 6.3.4.16	-	0
2.7.2.6	Formate kinase	-	0
2.7.2.7	Butyrate kinase	PF00871	27
2.7.2.8	Acetylglutamate kinase	PF00696 PF00696||PF04768	344 3
2.7.2.9	Transferred entry: 6.3.5.5	-	0
2.7.2.10	Phosphoglycerate kinase (GTP)	-	0
2.7.2.11	Glutamate 5-kinase	PF00696 PF00696||PF01472	34 230
2.7.2.12	Acetate kinase (diphosphate)	-	0
2.7.2.13	Deleted entry	-	0
2.7.2.14	Branched-chain-fatty-acid kinase	-	0
2.7.2.15	Propionate kinase	PF00871	14
2.7.2.16	2-phosphoglycerate kinase	-	0
2.7.2.17	[Amino-group carrier protein]-L-2-aminoadipate 6-kinase	-	0

Supplementary Table 2. *Annotation of proteins that are clustered in UMAP visualization to functional databases such as KEGG, GO, PFAM and annotation to EC number. We can see that proteins within a cluster share annotations.*

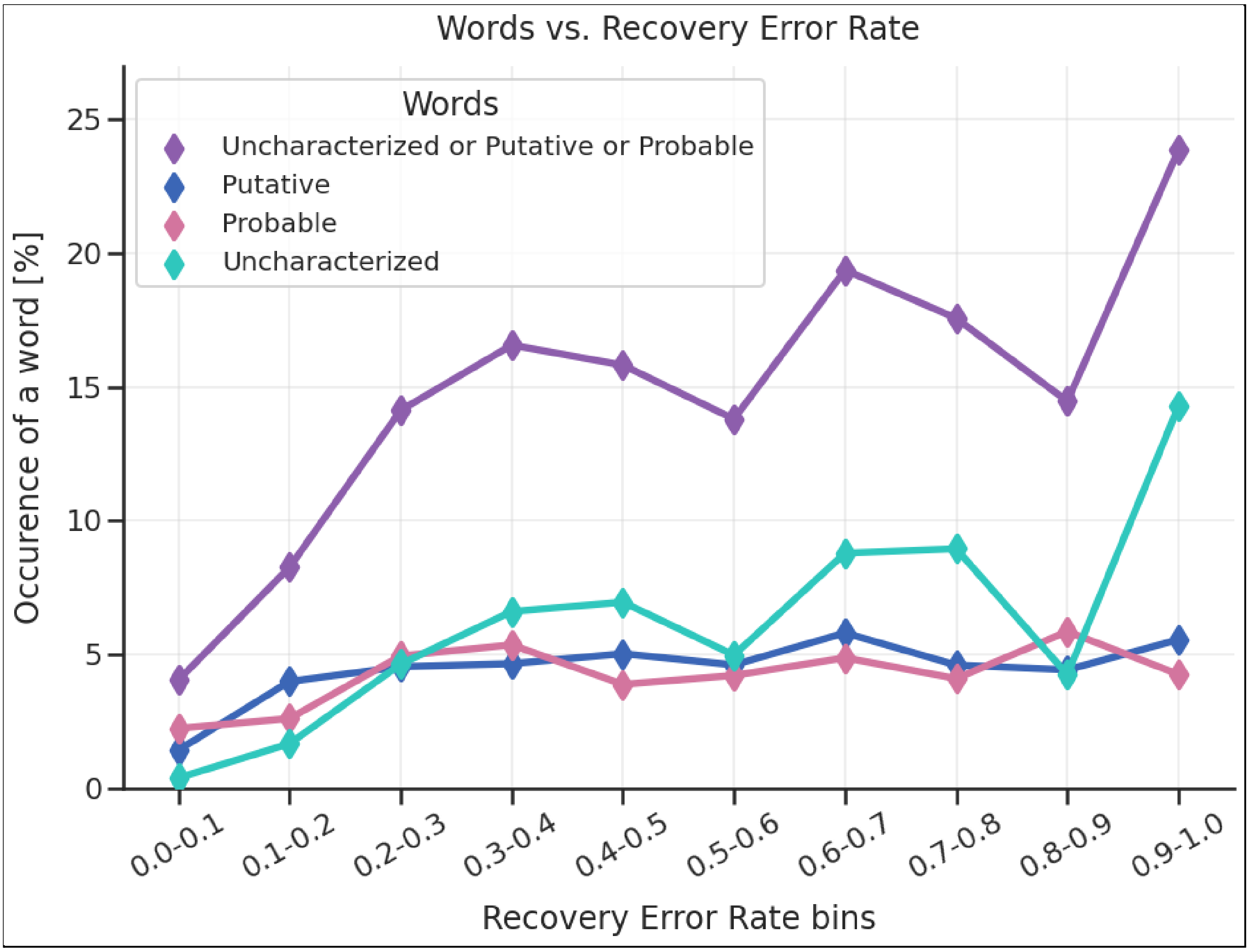
PROTEIN NAME	KO	EC number	PFAM
Fig 4A - proteins transferring alkyl or aryl groups, other than methyl groups			
UDP-N-acetylglucosamine 1-carboxyvinyltransferase	*K00790*	EC: 2.5.1.7	**PF00275**, EPSP_synthase
3-phosphoshikimate 1-carboxyvinyltransferase	*K00800*	EC: 2.5.1.19	
Fig 4B - GTP binding proteins			
Elongation factor G	*K02355*	-	**PF00679**, EFG_C **PF03764**, EFG_IV **PF00009**, GTP_EFTU **PF03144**, GTP_EFTU_D2
Peptide chain release factor 3	*K02837*	-	**PF00009,** GTP_EFTU (**PF03144**), GTP_EFTU_D2 **PF16658**, RF3_C
Fig 4E - ribosomal proteins			
30S ribosomal protein S1	*K02961*	- -	**PF00575**, S1
50S ribosomal protein L14	*K02874*		**PF00238**, Ribosomal_L14
50S ribosomal protein L36	*K02919*		**PF00444**, Ribosomal_L36
50S ribosomal protein L35	*K02916*		**PF01632**, Ribosomal_L35p
50S ribosomal protein L15	*K02876*		**PF00828**, Ribosomal_L27A
Fig 4D - tRNA ligases			
Cysteine--tRNA ligase	*K01883*	EC: 6.1.1.16	**PF09190**, DALR_2 **PF01406**, tRNA-synt_1e
Arginine--tRNA ligase	*K01887*	EC: 6.1.1.19	**PF03485**, Arg_tRNA_synt_N **PF05746**, DALR_1 **PF00750**, tRNA-synt_1d
Glutamate--tRNA ligase	*K01885*	EC: 6.1.1.17	**PF00749**, tRNA-synt_1c
Glutamine--tRNA ligase	*K01886*	EC: 6.1.1.18	**PF00749**, tRNA-synt_1c **PF03950**, tRNA-synt_1c_C
Glycine--tRNA ligase	*K01880*	EC: 6.1.1.14	**PF03129**, HGTP_anticodon **PF00587**, tRNA-synt_2b
Valine---tRNA ligase	*K01873*	EC: 6.1.1.9	**PF08264**, Anticodon_1 **PF00133**, tRNA-synt_1 **PF10458**, Val_tRNA-synt_C
isoleucyl-tRNA synthetase	*K01870*	*EC: 6.1.1.5*	**PF08264**, Anticodon_1 **PF00133**, tRNA-synt_1 **(PF06827),** zf-FPG_IleRS



Supplementary Figure 1. UMAP visualization of the k-mer protein representations space colored according to Kegg Orthology ID (KO). *We see separate groups of proteins, however, most of them are mixed up in the middle.*



Supplementary Figure 2. Relationship between Recovery Error Rate and the size of the class to which the protein belongs.



Supplementary Figure 3. The relationship between Recovery Error Rate and the occurrence of the words "Uncharacterized", "Putative", or "Probable"

Supplementary Table 3. *References to the databases used in our analyses.*

Database	Link to database
SUPFAM[64]	https://supfam.org
GENE 3D[65]	http://gene3d.biochem.ucl.ac.uk
InterPro[66]	https://www.ebi.ac.uk/interpro/
KO (KEGG Orthology)[67]	https://www.kegg.jp
GO (Gene Ontology)[68]	http://geneontology.org
eggNOG[69]	http://eggnog5.embl.de/#/app/home
EC number[70]	https://enzyme.expasy.org
Pfam[71]	http://pfam.xfam.org
Taxonomy[72]	https://www.ncbi.nlm.nih.gov/taxonomy

1. Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B. & Gough, J. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res.* **47**, D490–D494 (2019).
2. Lees, J. *et al.* Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* **40**, D465–71 (2012).
3. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
4. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research* vol. 49 D545–D551 (2021).
5. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
6. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
7. Duvaud, S. *et al.* Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* **49**, W216–W227 (2021).
8. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
9. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, (2020).