Genomic language model predicts protein co-regulation and function

Yunha Hwang[1*], Andre L. Cornman[2], Elizabeth H. Kellogg[3], Sergey Ovchinnikov[4*], Peter R. Girguis[1*]

[1] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
[2] Independent contributor
[3] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA
[4] John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA

* co-correspondence: yhwang@g.harvard.edu, so@fas.harvard.edu, pgirguis@oeb.harvard.edu

Abstract

Deciphering the relationship between a gene and its genomic context is fundamental to understanding and engineering biological systems. Machine learning has shown promise in learning latent relationships underlying the sequence-structure-function paradigm from massive protein sequence datasets. However, to date, limited attempts have been made in extending this continuum to include higher order genomic context information. Evolutionary processes dictate the specificity of genomic contexts in which a gene is found across phylogenetic distances, and these emergent genomic patterns can be leveraged to uncover functional relationships between gene products. Here, we trained a genomic language model (gLM) on millions of metagenomic scaffolds to learn the latent functional and regulatory relationships between genes. gLM learns contextualized protein embeddings that capture the genomic context as well as the protein sequence itself and encode biologically meaningful and functionally relevant information (e.g., enzymatic function, taxonomy). Our analysis of the attention patterns demonstrates that gLM is learning co-regulated functional modules (i.e., operons). Our findings illustrate that gLM's unsupervised deep learning of the metagenomic corpus is an effective and promising approach to encode functional semantics and regulatory syntax of genes in their genomic contexts and uncover complex relationships between genes in a genomic region.

Main

Introduction

Evolutionary processes result in the linkage between protein sequences, structure and function. The resulting sequence-structure-function paradigm[1] has long provided the basis for interpreting vast amounts of genomic data. Recent advances in neural network (NN)-based protein structure prediction methods[2,3], and more recently protein language models (pLMs)[4–6] suggest that data-centric approaches in unsupervised learning can represent these complex relationships shaped by evolution. To date, these models largely consider each protein as an independent and standalone entity. However, proteins are encoded in genomes alongside other proteins, and the specific genomic context that a protein occurs in is determined by evolutionary processes where each gene gain, loss, duplication and transposition event is subject to selection and drift[7–9]. These processes are particularly pronounced in bacterial and archaeal genomes where frequent horizontal gene transfers (HGT) shape genomic organization and diversity[10,11]. Thus, there exists an inherent evolutionary linkage between genes, their genomic context, and gene function[12], which can be explored by characterizing patterns that emerge from large metagenomic datasets.

Recently, two machine learning based approaches have shown predictive power of genomic context in gene function[13] and metabolic trait evolution[14] in bacterial and archaeal genomes. However, both these models represent genes as categorical entities, yet these genes exist in continuous space where multidimensional properties such as phylogeny, structure, and function are abstracted in their sequences. In order to close the gap between genomic-context and gene sequence-structure-function, we developed the first genomic language model (gLM) that, to our knowledge, represents proteins using pLM embeddings that encode relational properties[4] and structure information[15]. Our model is based on the transformer[16] architecture and is trained using millions of unlabelled metagenomic sequences via the masked language modeling[17] objective, with the hypothesis that its ability to attend to different parts of a multi-gene sequence will result in the learning of gene functional semantics and regulatory syntax (e.g. operons). Here, we report evidence of the learned contextualized protein embeddings and attention patterns capturing biologically relevant information. We demonstrate gLM's potential for predicting gene function and co-regulation, and propose future applications and research directions, including transfer learning capabilities of gLM.

Masked language modeling of genomic sequences

Language models, such as Bidirectional Encoder Representations from Transformers (BERT[17]), learn the semantics and syntax of natural languages using unsupervised training of a large corpus. In masked language modeling, the model is tasked with reconstructing corrupted input text[17], where some fractions of the words are masked. Significant advances in language modeling performance were achieved by adopting the transformer[16] neural network architecture, where each token (i.e. word) is able to attend to other tokens. This is in contrast to Long-Short-Term-Memory networks (LSTMs)[18] that sequentially processes tokens. To model genomic sequences, we trained a 19-layer transformer model (**Figure 1A;** for a detailed figure see **Extended Data 1**) on seven million metagenomic contig fragments consisting of 15 to 30 genes from the MGnify[19] database. Each gene in a genomic sequence is represented by a 1280 feature vector (context-free protein embeddings) generated by using ESM2 pLM[15], concatenated with an orientation feature (forward or backward). For each sequence, 15% of genes are randomly masked, and the model learns to predict the masked label using the genomic context. Based on the insight that more than one gene can legitimately be found in a particular genomic context, we allow the model to make four different predictions and predict their associated probabilities (**Extended Data 1**). Thus, instead of predicting their mean value, the model can approximate the underlying distribution of multiple genes that can occupy a genomic niche. We assess the model's performance using a pseudo-accuracy metric, where a prediction is considered correct if it is closest to the masked protein in Euclidean distance compared to the other proteins encoded in the sequence (see methods). We validate our model's performance on the *Escherichia coli* K-12 genome[20] by excluding from training 5.1% of MGnify subcontigs in which more than half of the proteins are similar (>70% sequence identity) to *E. coli* K-12 proteins. It is important to note that our goal was not to remove all *E. coli* K-12 homologs from the training, which would have removed a vast majority of

training data as many essential genes are shared across organisms. Instead, our goal was to remove as many *E. coli* K-12-like genomic contexts (subcontigs) from training, which is more appropriate for the training objective. gLM achieves 71.9% in validation pseudo-accuracy and 59.2% in validation absolute accuracy (**Extended Data 2)**. Notably, 53.0% of the predictions made during validation are with high confidence (with prediction likelihood > 0.75), and 75.8% of the high confidence predictions are correct, indicating gLM's ability to learn a confidence metric that corresponds to increased accuracy. We baseline our performance with a bidirectional LSTM model trained using the same language modeling task on the same training dataset, where validation performance plateaus at 28% pseudo-accuracy and 15% absolute accuracy (**Extended Data 2 and 3**).

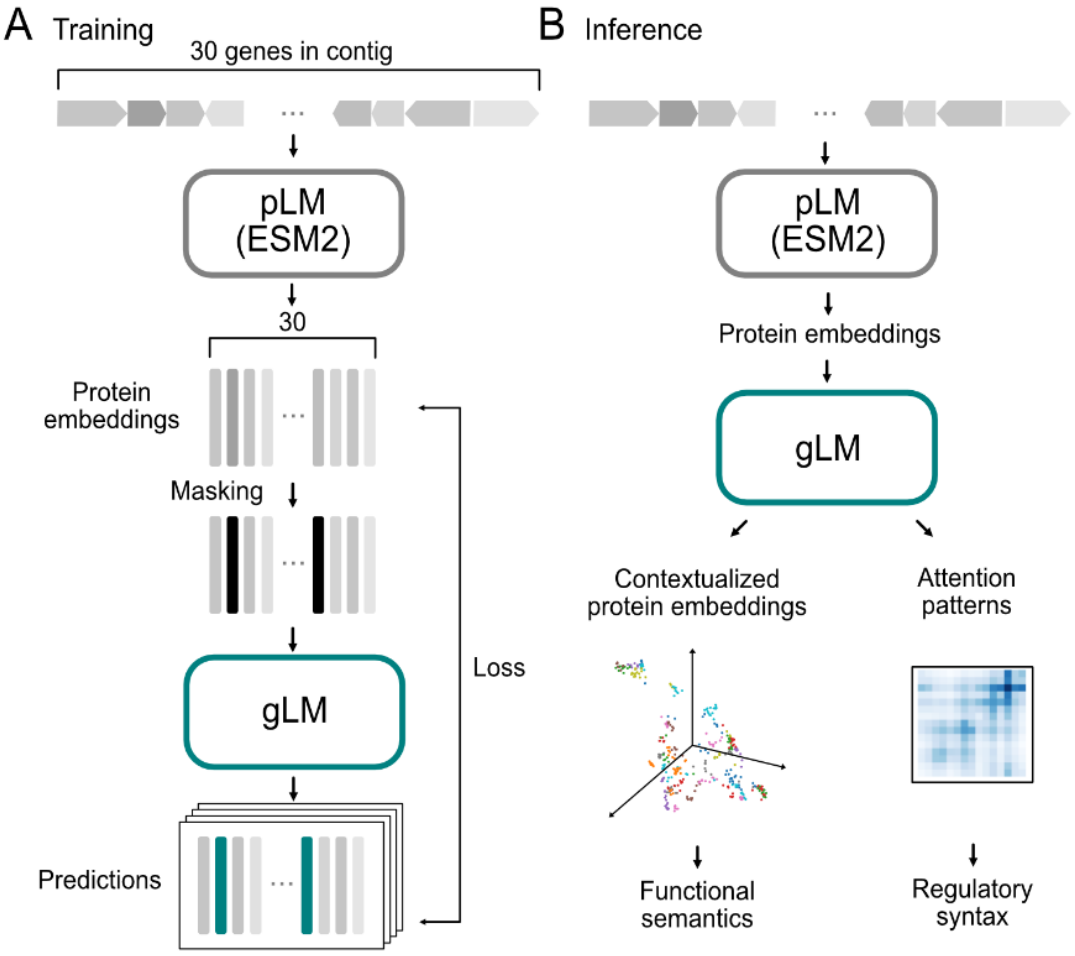


Figure 1. gLM training and inference schematics. A) For training, contigs (contiguous genomic sequences) containing up to 30 genes are first translated into proteins, which are subsequently embedded using a pLM (ESM2). Masked inputs are generated by random masking at 15% probability and gLM is trained to make four predictions for each masked protein, with associated likelihoods. Training loss is calculated on both the prediction and likelihoods. **B)** At inference time, inputs are generated from a contig using ESM2 output. Contextualized protein embeddings (last hidden layer of gLM) and attention patterns are used for various downstream tasks. See **Extended Data 1** for a detailed schematic.

Contextualized gene embeddings capture gene semantics

The mapping from gene to gene-function in organisms is not one-to-one. Similar to words in natural language, a gene can confer different functions[21] depending on its context[22], and many genes confer similar functions (i.e. convergent evolution[23], remote homology[24]). We used gLM to generate 1280-feature contextualized protein embeddings at inference time (**Figure 1B**), and we examined the "semantic" information captured in these embeddings. Analogous to how words are likely to have different meanings depending on the type of text in which they are found (**Figure 2A)**, we find that contextualized protein embeddings of genes that appear across multiple environments (biomes) tend to cluster based on biome types. We identified 31 proteins in our training database (MGYPs) that occurred more than 100 times and distributed with at least 20 occurrences in each "Host-associated", "Environmental", and "Engineered" biomes according to MGnify's designation. We find that gLM's contextualized protein embeddings capture biome information for the majority (n=21) of these multi-biome MGYPs. For instance, a gene encoding a protein annotated "translation initiation factor IF-1" occurs multiple times across biomes. While the input to gLM (context-free protein embedding; ESM2 output) is identical across all occurrences, gLM's output (contextualized protein embeddings) cluster with biome types (**Figure 2B**; see the 30 other multi-biome MGYP visualizations in **Extended Data 4**). This suggests that the diverse genomic contexts that a gene occupies are specific for different biomes, implying biome-specific gene semantics.

We explored an ecologically important example of genomic "polysemy" (multiple meanings conferred by the same word; **Figure 2C**) of methyl-coenzyme M reductase (MCR) complex. The MCR complex can carry out a reversible reaction (Reaction 1 in **Figure 2D**), whereby the forward reaction results in the production of methane (methanogenesis) while the reverse results in methane oxidation (methanotrophy). We first examine the McrA (methyl-coenzyme M reductase subunit alpha) protein in diverse lineages of ANME (ANaerobic MEthane oxidizing) and methanogenic archaeal genomes. These archaea are polyphyletic and occupy specific ecological niches. Notably, analogous to how a semantic meaning of a word exists on a spectrum and a word can have multiple semantically appropriate meanings in a context (**Figure 2C),** the MCR complex can confer different functions depending on the context. Previous reports demonstrate capacities of ANME (ANME-2 in particular) carrying out methanogenesis[25] and methanogens conducting methane oxidation in specific growth conditions[26]. The context-free ESM2 embedding of these proteins (**Figure 2E**) shows little organization, with little separation between ANME-1 and ANME-2 McrA proteins. However, contextualized gLM embeddings (**Figure 2F**) of the McrA proteins show distinct organization where ANME-1 McrA proteins form a tight cluster, while ANME-2 McrA proteins form a cluster closer to methanogens. This organization reflects the phylogenetic relationships between the organisms that McrAs are found in, as well as the distinct operonic and structural divergence of MCR complexes in ANME-1 compared to those found in ANME-2 and methanogens[27]. As proposed by Shao et al.[27], the preferred directionality in Reaction 1 (**Figure 2D**) in ANME-2 and some methanogens may be more dependent on thermodynamics.

We also demonstrate that contextualized gLM embeddings are more suitable for determining the functional relationship between gene classes. Analogous to how the words "dog" and "cat" are closer in meaning relative to "dog" and "train" (**Figure 2G)**, we see a pattern where Cas1 and Cas2 that appear diffuse in multiple subclusters in context-free protein embedding space (**Figure 2H)** cluster in contextualized embedding space (**Figure 2I**). This reflects their similarity in function (e.g., phage defense). This is also demonstrated in biosynthetic genes, lipopolysaccharide synthase (LPS) and polyketide synthase (PKS) genes clustering closer together in contextualized embedding space distinct from the Cas proteins (**Figure 2I)**. Contextualized protein embeddings are therefore able to capture relational properties semantic information[28], where proteins that are more similar in their function appear in more similar genomic contexts.

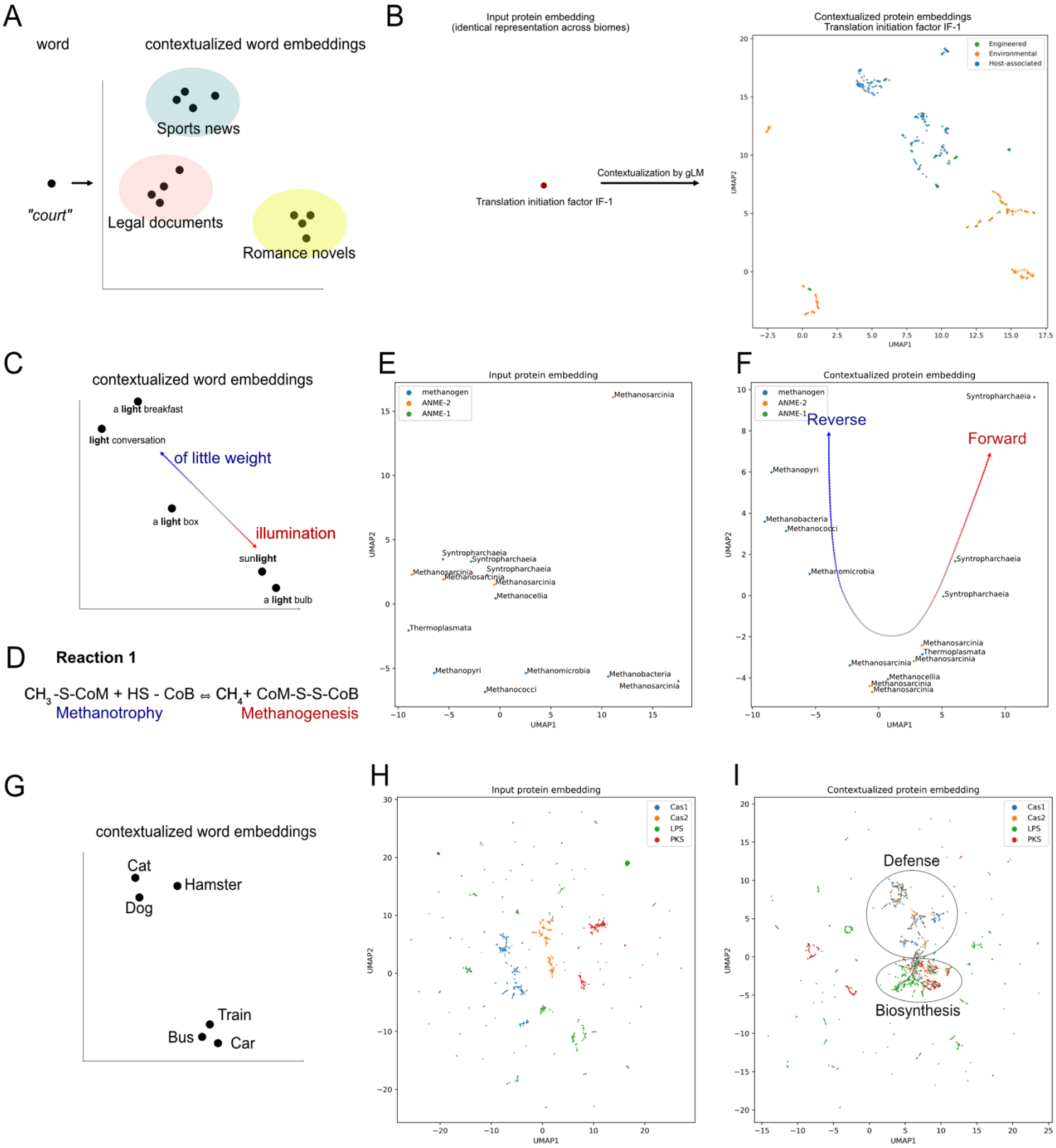


Figure 2. Contextualized protein embedding analysis and comparison with concepts in natural language modeling. A) A word upon contextualization can be mapped to embedding space. For many words, the semantic meaning varies in different types of literature, and therefore their contextualized

embeddings cluster with source text type. Figure was created for qualitative visualization. **B)** The input protein embedding (output of ESM2 and context-free protein embedding) is the same across all occurrences of the protein in the database. Upon contextualization with gLM, contextualized protein embeddings of the same protein (last hidden layer of gLM at inference time) cluster with biome type, analogous to the source text type in natural language. Contextualization of 30 other multi-biome MGYPs can be found in **Extended Data 4**. **C)** A word's meaning upon contextualization varies across a continuous spectrum and can be ambiguous even with contextualization (e.g., double entendre). **D)** Reaction 1, carried out by the MCR complex, either backward (Methanotrophy) or forward (Methanogenesis). **E)** Input protein embeddings of McrA sequences in genomes, colored by metabolic classification of the organism (ANME, methanogen) based on previous studies and labeled by class-level taxonomy. **F)** McrA protein embeddings, upon contextualization, cluster with the direction of Reaction 1 that the MCR complex is likely to carry out. **G)** Geometric relationship between contextualized protein embeddings based on the semantic closeness of words. **H)** Input (context-free) protein embeddings of Cas1, Cas2, lipopolysaccharide synthases (LPS) and polyketide synthases (PKS) showing clustering based on structural and sequence similarity. **I)** Clustering of contextualized protein embeddings where phage defense proteins cluster (Cas1 and Cas2) and biosynthetic gene products cluster (LPS and PKS).

Characterizing the unknown

Metagenomic sequences feature many genes with unknown or generic functions, and some are so divergent that they do not contain sufficient sequence similarity to the annotated fraction of the database[29]. In our dataset, of the 30.8M protein sequences, 19.8% could not be associated with any known annotation (see methods), and 27.5% could not be associated with any known Pfam domains using a recent deep learning approach (ProtENN[30]). Understanding the functional role of these proteins in their organismal and environmental contexts remains a major challenge because most of the organisms that house such proteins are difficult to culture and laboratory validation is often low throughput. We demonstrate the potential of using contextualized protein embeddings to assign putative functions to previously unannotated proteins. This approach is motivated by the observation that proteins that confer similar functions are found in similar genomic contexts due to selective pressures bestowed by functional relationships (e.g., protein-protein interactions, co-regulations) between genes. pLM-based context-free protein embeddings have previously been proposed for annotation transfer[31], particularly through remote homology detection[4]. However, we find that many of these unannotated proteins (orange) form large separate clusters away from the annotated fraction (blue) of the pLM embedding space (**Figure 3A**). This suggests that pLM-embedding based annotation transfer potential is limited for many unannotated proteins that are not close to any annotated proteins in representation space. In contrast, by using gLM-based contextualized protein embeddings, fewer unannotated proteins cluster with each other and can be associated with annotated proteins in representation space, resulting in higher annotation transfer potential (**Figure 3B**). This reflects the importance of context information for feasibly transferring validated knowledge and annotation in cultivable and well-studied strains (e.g., *E. coli*) to the vastly uncultivated metagenomic sequence space. Genomic context, along with molecular structure and phylogeny, appear to be important information to abstract in order to effectively represent sequences such that we can uncover hidden associations between the known and the unknown fractions of biology.

Contextualization improves enzyme function prediction

To test the hypothesis that the genomic context of the protein can be used to aid function prediction, we compared how much the addition of context information can improve the expressiveness of protein representations for enzyme function prediction. Specifically, we examined the capabilities of context-free (pLM) representation and context-added (pLM+gLM) representations in classifying enzyme commission (EC)[32] numbers of proteins. We trained a simple three-layer perceptron neural network (**Extended Data 5**) to predict all four hierarchies of EC number on a random 100,000 subcontigs in the training database, from which 115768 unique MGYPs could be assigned to 2637 different EC numbers (see methods). We find 2% and 3% increases in test prediction accuracy and precision respectively with the addition of contextualization (**Extended Data 6**). We found that the addition of context information results in largest performance (F1 score) gains (up to six-fold) in some EC classes that are known to feature structural diversity and phylogenetic heterogeneity (**Figure 3C**). For instance, EC 1.6.5.11 designates NADH:quinone oxidoreductases, which consists of structurally diverse enzymes with various substrate binding sites grouped together by the specific function of catalyzing electron transfer from NAD(P)H to quinones[33]. Similarly, ECs 2.7.7.65 and 2.7.11.1, denoting diguanylate cyclases and non-specific serine/threonine protein kinases respectively, feature diversity in substrates that they can bind[34,35].

In some EC classes, however, the additional context information seems to lower the prediction performance. For instance, ECs 4.1.1.81 and 5.4.99.60 designate pyridoxal-phosphate protein and cobalt-precorrin-8 methylmutase respectively, which are structurally and evolutionarily conserved[36]. Moreover, we also find that the additional context information leads to worse performance for pyruvate synthases (EC 1.2.7.1) suggesting the possibility that the diverse contexts that pyruvate synthases are found in may result in misclassification, or conversely, represent possible functional diversity[37]. Previous studies[37–39] have utilized deep learning to predict EC numbers directly from protein sequences and fine-tuning such approaches with contextualization could provide ways to infer function even for classes with fewer training examples or identify proteins with more than one function.

221
222 ***Horizontal transfer frequency corresponds to genomic context embedding variance***
223 A key process that shapes microbial genome organization and evolution is horizontal gene transfer (HGT).
224 The taxonomic range in which genes are distributed across the tree of life depends on their function and the
225 selective advantage they incur in different environments. Relatively little is known about the specificity in
226 the genomic region into which a gene gets transferred across phylogenetic distances. We examined the
227 variance of genomic context embeddings for proteins that occur at least one hundred times in the database.
228 Variance of genomic contexts are calculated by taking a random sample of 100 occurrences and then
229 calculating the mean pairwise distances between the hundred contextualized protein embeddings. We
230 conduct such independent random sampling and distance calculation ten times per protein and then
231 calculate the mean value and term this the genomic context variance. Our results show that the genomic
232 context variances are similarly distributed with a long right-hand tail for both annotated and unannotated
233 fractions of the genes **(Figure 3D)**. The most context-variant proteins in the right tail included phage
234 proteins and transposases, reflecting their ability to self-mobilize. Genomic context variances can be used
235 as a proxy for horizontal transfer frequencies and can be used to compare the fitness effects of the genomic
236 context on the evolutionary trajectory (e.g., gene flow) of genes and identify novel and functional
237 transposable elements.

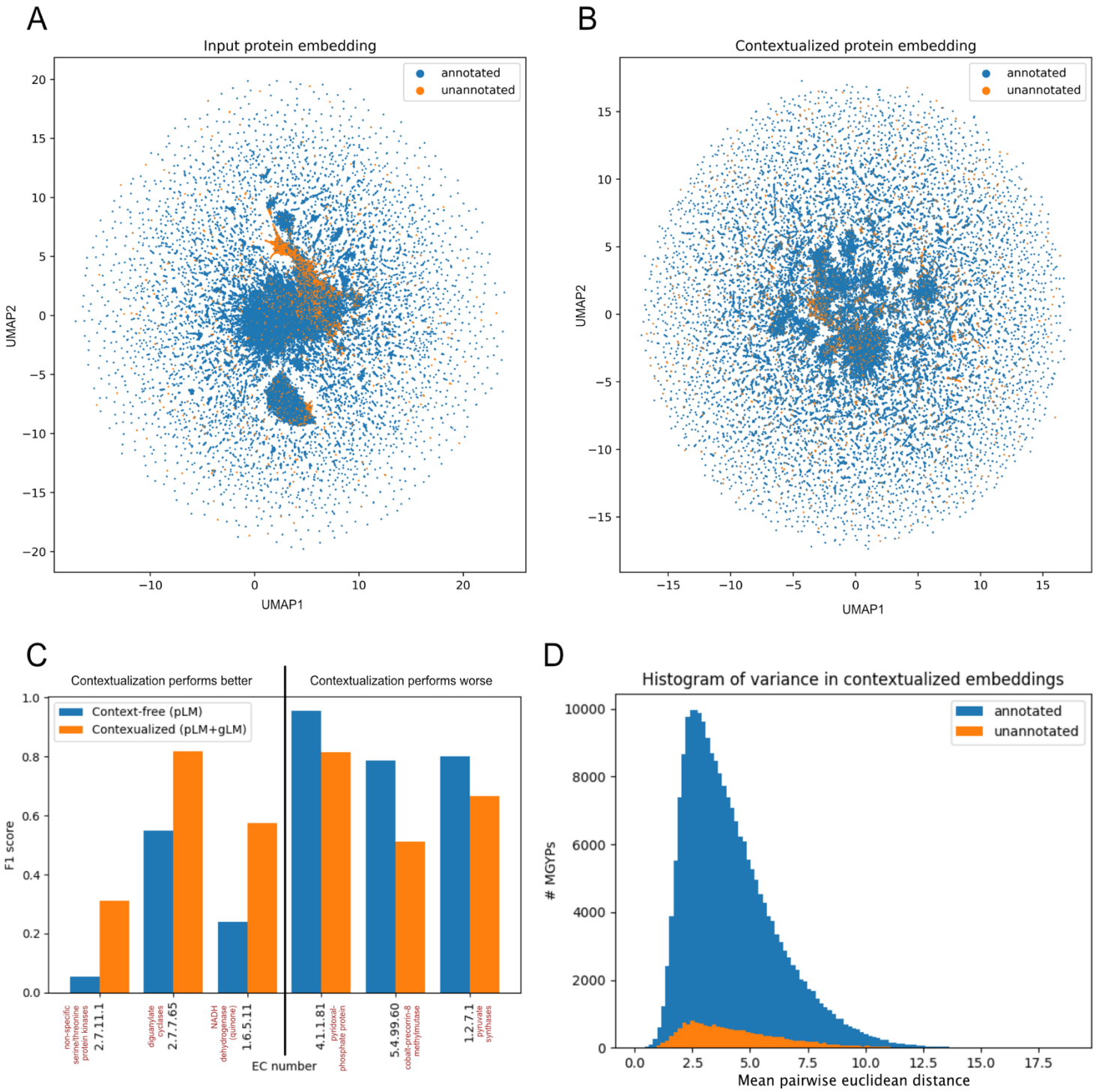


Figure 3. Contextualization of unannotated proteins. A) UMAP visualization of input protein embeddings (ESM2 output) of annotated proteins (blue) and proteins without annotation (orange) from a random 0.15% subset of the MGnify database. Many unannotated proteins cluster distinctly from the annotated proteins, indicating limited annotation transfer potential. **B)** UMAP visualization of contextualized embeddings of the same proteins. Unannotated proteins do not form distinct clusters from the annotated proteins, indicating higher annotation transfer potential. **C)** Three classes with highest

performance gain in EC prediction with contextualization (left) and three where contextualization results in diminished performance (right) **D)** Histogram of variance (# bins = 100) in contextualized embeddings of MGYPs that occur at least 100 times in the database. Histograms for unannotated and annotated fraction of the MGYPs are plotted separately and bars are not stacked. Annotated examples in the long right tail include phage proteins and transposases, reflecting their ability to self-mobilize.

Transformer's attention captures operons

The transformer attention mechanism[16] models pairwise interaction between different tokens in the input sequence. Previous examinations of the attention patterns of transformer models in natural language processing (NLP)[40] have suggested that different heads appear to specialize in syntactic functions. Subsequently, different attention heads in pLMs[41] have been shown to correlate to specific structural elements and functional sites in a protein. For our gLM, we hypothesized that specific attention heads focus on learning operons, a "syntactic" feature in genomes where multiple genes form regulatory modules. We used the *E. coli* K-12 operon database[42] consisting of 817 operons for validation. gLM contains 190 attention heads across 19 layers. We found that heads in shallower layers correlated more with operons (**Figure 4A, Extended Data 7)**, with raw attention scores in the 7th head of the 2nd layer [L2-H7] linearly correlating with operons with 0.44 correlation coefficient (Pearson's rho, Bonferroni adjusted p-value < 1E-5) (**Figure 4B**). We further trained a logistic regression classifier using all attention patterns across all heads. Our classifier predicts the presence of an operonic relationship between a pair of neighboring proteins in a sequence with high precision (mean average precision = 0.77, five-fold cross-validation) (**Figure 4C**).

Context dependency of AAA+ regulator functions in complex genetic systems

Understanding the functional role of a regulatory protein in an organism remains a challenging task because the same protein fold may carry out different functions depending on the context. For instance, AAA+ proteins (ATPases associated with diverse cellular activities) utilize the chemical energy from ATP hydrolysis to confer diverse mechanical cellular functions[43]. However, AAA+ regulators can also play very different, broad functional roles depending on their cellular interacting partners from protein degradation and DNA replication to DNA transposition. One particularly interesting example is the TnsC protein, which regulates DNA insertion activity[44] in Tn7-like transposon systems. Multiple bioinformatic efforts focused on discovery of novel transposons through metagenome search[45] and sequence searches of assembled genomes[46] aimed at identifying suitable homologs for genome-editing applications[47]. In order to test whether the methods developed here could identify novel Tn7-like transposition systems as well as distinguish these from other functional contexts, we explored the contextualized semantics of TnsC's structural homologs in the MGnify database. Without contextualization, there appears no clustering with associated transposase activity (**Figure 4E**). However, with added contextualization, previously identified TnsC (orange) and manually inspected TnsC-like structural homolog (red, labeled "TnsC-like") cluster together (**Figure 4F**). We further validate this visualization using embedding distance-based clustering (**Extended Data 8)**. Many structural homologs of TnsC were not involved in transposition and this is reflected in distinct clusters of gray data points away from known TnsC (oranges) and TnsC-like structural homologs (red) in **Figure 4F**. These clusters represent diverse and context-dependent AAA+ regulation activity that cannot be predicted from neither structure nor raw sequence alone. We predicted an operonic relationship between these AAA+ regulators and their neighboring genes and found many to be in operonic relationships with gene modules of diverse function including pilus assembly and viral host-nuclease inhibition (**Figure 4D, Extended Data 9**). In some cases, queried AAA+ proteins did not appear to be in an operonic association with the neighboring proteins, suggesting some AAA+ proteins are less likely to be functionally associated with their neighbors than others (**Extended Data 9, example 6)**. Using this example of AAA+ regulators, we illustrate that combining the contextualized protein embeddings and attention-based operon interaction may provide an important avenue for exploring and characterizing the functional diversity of regulatory proteins.

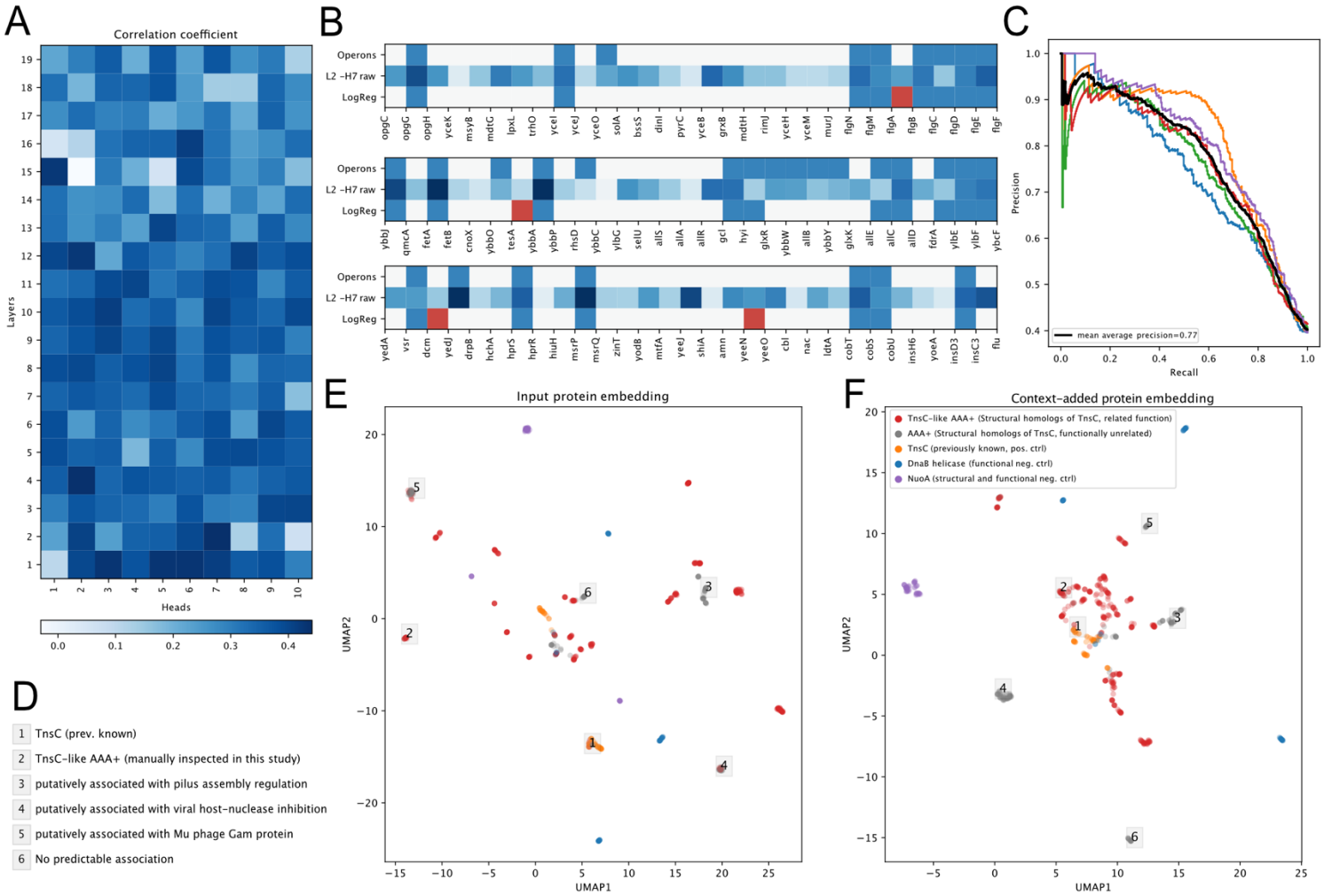


Figure 4. Attention analysis. A) Correlation coefficients (Pearson's rho) between attention heads across layers and operons. Darker color corresponds to stronger correlation with previously identified operons. Attention patterns of the second layer-seventh head [L2-H7] is most strongly correlated with the operons. **B)** Three random examples of contigs and predicted operonic relationship between neighboring genes. Proteins are listed in the order they are encoded in the contig. Top row (Operons): Ground truth E.coli K-12 operons based on RegulonDB; presence of operonic relationship in blue. Second row ([L2-H7]): raw attention scores in the attention head [L2-H7] most correlated with operons, with darker blue corresponding to higher attention scores. Third row (LogReg): logistic regression prediction using all attention heads, where the predicted operonic relationship between neighboring genes is marked in blue, while false positive predictions (or possibly misannotated ground truths in the case of flagellar proteins in the first example) are marked in red. **C)** Five-fold cross-validation precision-recall curves of logistic regression trained using all operons and attention heads. **D)** AAA+ regulator associations characterized using attention-based prediction of operons (**Extended Figure 7**) corresponding to labeled examples in panels E and F. **E)** ESM2 generated input protein embeddings of AAA+ regulator proteins that are structural homologs to TnsC (grey and red; using Foldseek[48]). Structural homologs of TnsC with confirmed involvement in Tn7-like transposons upon manual inspection were designated "TnsC-like AAA+ (manually inspected)" and are colored red. Other MGYP proteins annotated as "TnsC" against the UniRef90 database (orange) were added as positive controls for TnsC function. NuoA (NADH-quinone oxidoreductase subunit A; purple) were added as structural and functional negative controls. DnaB helicases (blues) were added as functional

negative controls, as these proteins have similar folds to TnsC but are not associated with transposition. **F)** Combined input protein and context embeddings of proteins in panel E. These embeddings are generated through concatenation of pLM (ESM2) embeddings and context (last layer of gLM) embeddings. Negative controls (NuoA and DnaB helicases) form distinct clusters in both E and F. Numbered labels in grey boxes indicate the AAA+ proteins with various functional association predictions listed in panel D and **Extended Data 9.** Raw distance-based cluster maps of the embeddings are shown in **Extended Data 8**.

gLM predicts paralogy in protein-protein interactions

Proteins in an organism are found in complexes and interact physically with each other. Recent advances in protein-protein interaction (PPI) prediction and structural complex research have largely been guided by identifying interologs (conserved PPI across organisms) and co-evolutionary signals between residues [49]. However, distinguishing paralogs from orthologs (otherwise known as the "Paralog matching" problem) in the expanding sequence dataset remains a computational challenge requiring queries across the entire database and/or phylogenetic profiling. In cases where multiple interacting pairs are found within an organism (e.g., histidine kinases (HK) and response regulators (RR)), prediction of interacting pairs is particularly difficult[50]. We reasoned that gLM, although not directly trained for this task, may have learned the relationships between paralogs versus orthologs. To test this capability, we used a well-studied example of interacting paralogs (ModC and ModA, **Figure 5A**) which form an ABC transporter complex. We queried gLM to predict the embedding of an interacting pair given no context except the protein sequence of either ModA or ModC. We find that without any fine-tuning gLM performs two orders of magnitude better than what is expected by random chance (see methods). Specifically, for 417 out of 2700 interacting pairs, gLM makes predictions that belong to the same cluster (50% sequence identity, n=2100 clusters) as the true label, and in 73 pairs, the gLM predicts a label that is closest to the exact interacting pair (**Figure 5B)**. Importantly, when considering only very high confidence predictions (prediction likelihood > 0.9, n=466), the gLM can match paralogs with an increased accuracy of 25.1%. When paralogs are correctly paired, gLM is more confident about the prediction (average confidence for correct prediction = 0.79, average confidence across all predictions = 0.53), while less certain predictions are either out of distribution, or closer to the mean of labels (**Figure 5C**). We attribute part of the inaccuracies in prediction to the fact that gLM was not trained on the task of predicting a masked gene given only a single gene as genomic context, though we expect the performance to improve with expanding the training sequence length range and fine-tuning the model specifically for the "paralog matching" problem.

Contextualized contig embeddings and potential for transfer learning

Contextualized protein embeddings encode the relationship between a specific protein and its genomic context, retaining the sequential information within a contig. We hypothesized that this contextualization adds biologically meaningful information that can be utilized for further characterization of the multi-gene genomic contigs. Here, we define a contextualized contig embedding as a mean-pooled hidden layer across all proteins in sequence, and a context-free contig embedding as mean-pooled ESM2 protein embeddings across the sequence (see methods). Both embeddings consist of 1280 features. We test our hypothesis by examining each of these embeddings' ability to linearly distinguish viral sequences from bacterial and archaeal sub-contigs. In metagenomic datasets, the taxonomic identity of assembled sequences must be inferred post-hoc, therefore the identification of viral sequences is conducted based on the presence of viral genes and viral genomic signatures[51]. However, such classification task remains a challenge particularly for smaller contig fragments and less characterized viral sequences. Here, we sampled random 30-protein sub-contigs from the representative bacterial and archaeal genome database and reference viral genomes in the NCBI and visualized their context-free contig embeddings (**Figure 5D)** and contextualized contig embeddings (**Figure 5E**). We observed more separation and taxonomic clusters at both domain- and class-levels (**Extended Data 10)**, suggesting that taxonomic signature is enhanced by encoding the latent relationships between proteins. This is further validated by training a logistic regression classifier on context-free and contextualized contig embeddings for class-level taxonomy (**Extended Data 10AB**), where we see a statistically significant improvement in average precision (**Figures 5F,** see **Extended Data 10CD** for confusion matrices). This emphasizes the biological importance of a protein's relative position in the genome and its relationship with the genomic environment, and further indicates that this information can be effectively encoded using gLM. Contextualized contig embeddings present opportunities for transfer learning beyond viral sequence prediction, such as improved metagenomically assembled genome (MAG) binning and assembly correction.

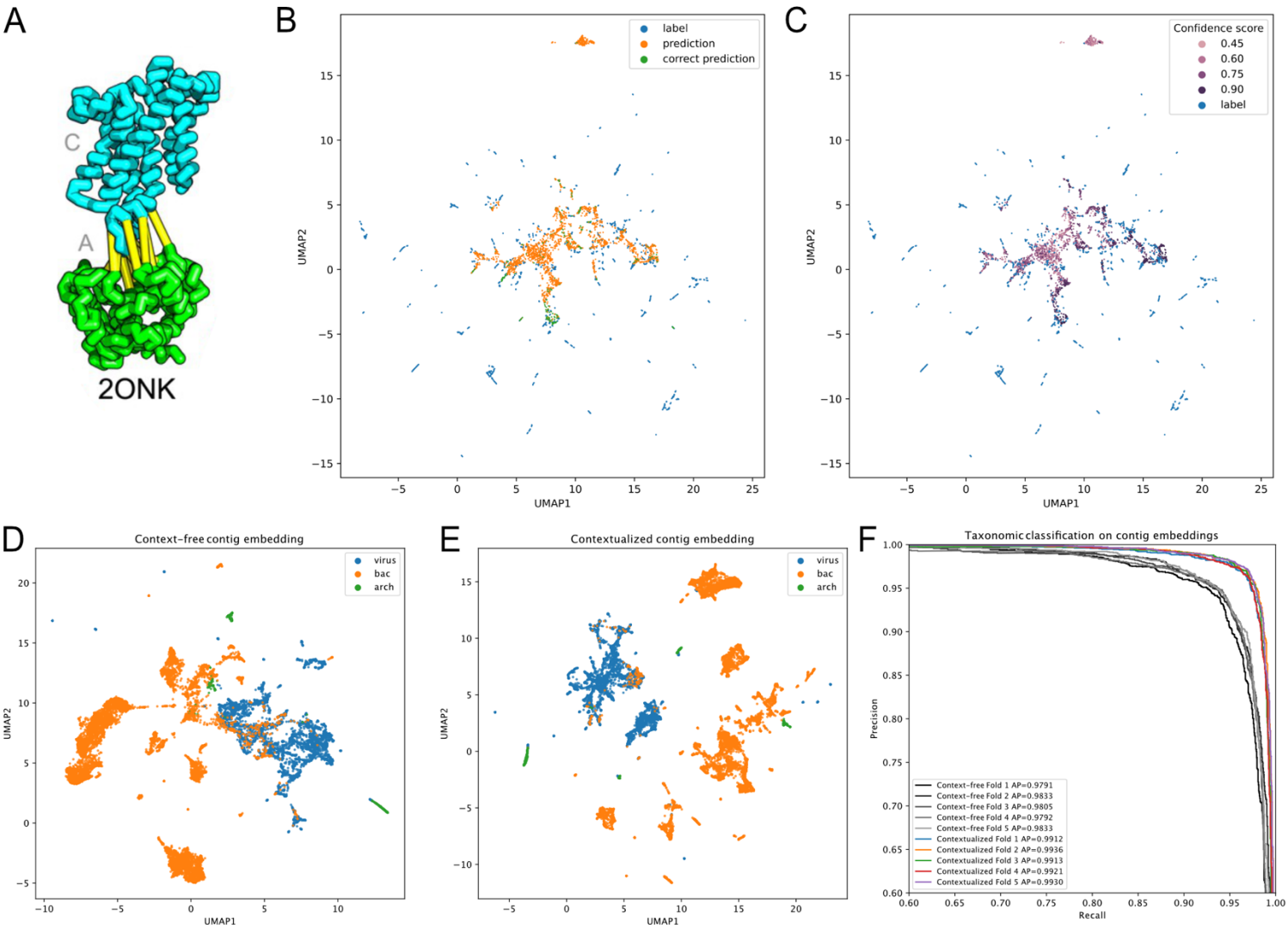


Figure 5. Potential for transfer learning. A) ModA and ModC interaction (adapted with permission from Ovchinnikov et al.[49]) **B)** UMAP projection of predictions (orange) and labels (blues) of paralogs (ModAC shown in B), where correct predictions are colored in green. **C)** Predicted embeddings are colored based on the predicted confidence. Out of distribution predictions and predictions closer to the mean are generally of lower confidence, while correct predictions are of higher confidence. **D&E)** Random 30-gene contigs from representative bacterial ("bac") and archaeal ("arch") genomes and reference viral ("vir") genomes were embedded by mean-pooling ESM2 protein embeddings (context-free contig embeddings, D) and by mean-pooling the last hidden layer of gLM (contextualized contig embeddings, E). **F)** Micro-averaged precision-recall curves and average precisions for logistic regression classifiers trained using context-free contig embeddings (grey lines) and contextualized contig embeddings (color lines) for class-level taxonomy classification task. Each line represents a fold in stratified k-fold cross-validation (k=5). Class-level taxonomy for each contig is shown in **Extended Data 10AB** and the confusion matrices for logistic regression classifiers are shown in **Extended Data 10CD.**

387 ***Discussion***
388 The unprecedented amount and diversity of metagenomic data, coupled with advances in deep learning
389 presents exciting opportunities for building models that can learn hidden patterns and structures of
390 biological systems. Such models build upon the conceptual and statistical frameworks that evolutionary
391 biologists have developed for the past century. With capabilities of abstracting much larger amounts of
392 data, these models can disentangle the extraordinary complexity of organismal genomes and their encoded
393 functions; this is a key step in furthering our understanding of biological processes. The work presented
394 here demonstrates and validates the concept of genomic language modeling. Our implementation of the
395 masked genomic language modeling illustrates the feasibility of training such a model and provides
396 evidence that biological information is being captured in learned contextualized embeddings and yielding
397 meaningful interpretations of the attention patterns. We show that gLM can be used for diverse downstream
398 tasks including enzyme function prediction, operon prediction, paralog matching and contig taxonomy
399 prediction. Furthermore, we demonstrate gLM's ability to illuminate context dependency in functions
400 across structural and sequence homology through the example of AAA+ regulators. Taken together, gLM
401 presents a highly promising direction for interpreting biology and we propose key areas for further
402 development: First, the transformer architecture has shown to be successful in efficient scaling; in both
403 natural language[52] and protein language processing[15], increasing the number of parameters in the model
404 along with the training dataset size have been shown to lead to vastly improved performance and
405 generalizability. Our model consists of ~1B parameters which is at least a magnitude smaller compared to
406 state-of-the-art pLMs. With further hyperparameter tuning and scaling, we expect better performance of
407 the model. Second, our model currently uses protein-level pLM embeddings to represent proteins in the
408 input. These embeddings are generated by mean-pooling the residue-level hidden states across the protein
409 sequence, and therefore the residue specific information is likely obscured. Future approaches for scaling
410 should consider using the residue-level embeddings, which contain richer information that can be used to
411 link residue-to-residue co-evolutionary interactions between proteins. Third, the task of reconstructing
412 masked protein embeddings requires modeling a distribution over possible embeddings; our method
413 approximates this distribution using a fixed number of predictions. Future work could improve upon this
414 by using a generative approach, such as a diffusion or GAN model. This may allow for better prediction
415 accuracy and greater generalizability for unseen datasets. Fourth, adding non-protein modalities (e.g. non-
416 coding regulatory elements[13]) as input to gLM may also greatly improve gLM's representation of biological
417 sequence data, and can learn protein function and regulation conditioned upon other modalities[53].
418
419 One of the most powerful aspects of the transformer-based language models is their potential for transfer
420 learning and fine-tuning. We tested some of the capabilities of gLM and successfully showed that higher
421 order biological information including gene function and regulation can be learned using genomic
422 sequences. Our results highlight the importance of contextualization of biological data, particularly as we
423 scale our modeling efforts from biomolecules to whole organisms. We propose the following promising
424 future directions for applying gLM for advancing biological research. 1) Feature-based transfer learning for
425 predicting protein function (e.g., Gene Ontology [GO] term, EC number), particularly those with limited
426 sequence and structural homology. 2) Fine-tuning gLM for the protein-protein-interactome prediction task.
427 3) Using gLM features to encode genomic contexts as additional input for improved and contextualized
428 protein structure predictions. In conclusion, genomic language modeling is a powerful tool to unbiasedly
429 condense important biological information from full metagenomic sequences. Coupled with the advances
430 in long-read sequencing, we expect a drastic increase in the input data quality, quantity, and diversity.
431 Genomic language modeling presents an avenue to bridge the gap between atomic structure and organismal
432 function, and thereby brings us closer to modeling biological systems, discovering novel biology, and
433 ultimately, precise bioengineering (e.g., genome editing, synthetic biology).
434
435

Methods

Sequence database

The genomic corpus was generated using the MGnify[19] dataset (released 2022-05-06 and downloaded 2022-06-07). First, genomic contigs with greater than 30 genes were divided into 30 gene non-overlapping subcontigs resulting in a total of 7,324,684 subcontigs with lengths between 15 and 30 genes (subcontigs < 15 genes in length were removed from the dataset). Each gene in the subcontig was mapped to a representative protein sequence (representative MGYP) using mmseqs/linclust[54], with coverage and sequence identity thresholds set at 90% (pre-computed in the MGnify database), resulting in a total of 30,800,563 representative MGYPs. Each representative MGYP was represented by a 1280-feature protein embedding, generated by mean-pooling the last hidden layer of the ESM2[55] "esm2_t33_650M_UR50D" model. Due to the memory limitation in computing embeddings for very long sequences, 116 of the MGYP sequences longer than 12290 amino acids were truncated to 12290 amino acids. ESM2 embeddings were subsequently standardized and clipped at (-10,10). A small fraction (0.4%) of the genes could not be mapped to a representative MGYP and therefore the corresponding sequence information could not be retrieved from the MGnify server; these sequences were assigned a 1280 feature vector of ones. For each gene in the sub-sequence, we added a gene orientation feature to the standardized MGYP protein embedding, where 0.5 denotes "forward" orientation relative to the direction of sequencing, and -0.5 denotes "reverse" orientation. Thus, each gene was represented by a 1281 feature vector in our corpus.

gLM architecture and training

gLM was built on the huggingface implementation of the RoBERTa[56] transformer architecture. gLM consisted of 19 layers with hidden size 1280 and ten attention heads per layer, with relative position embedding ("relative_key_query")[57]. For training, 15% of the tokens (genes) in the sequence (subcontig) were randomly masked to a value of -1. We then tasked the model with the objective of predicting the label of the masked token, where the label consists of a 100-feature vector that consists of the PCA whitened 99 principal components of the corresponding ESM2 protein embedding concatenated with its orientation feature. Specifically, gLM projects the last hidden state of the model into four 100-feature vectors and four corresponding likelihood values using a linear layer. Total loss is calculated using the following equation:

$$MSE(closest\ prediction, label) + \alpha * CrossEntropyLoss(likelihoods, closest\ prediction\ index)$$

The closest prediction is defined as the prediction that is closest to the label, computed by L2 distance. We set α = 1e-4. gLM was trained in half precision with batch size 3000 with distributed data parallelization on four NVIDIA A100 GPUs over 1,296,960 steps (560 epochs) including 5000 warm-up steps to reach a learning rate of 1e-4 with AdamW[58] optimizer.

Performance metric and validation

To evaluate the model quality and its generalizability beyond the training dataset, we use a pseudo-accuracy metric, where we deem a prediction to be "correct" if it is closest in Euclidean distance to the label of the masked gene relative to the other genes in the subcontig. Thus,

$$pseudo\ accuracy = \frac{\#count(argmin(dist(prediction, labels\ in\ subcontig)) == index(masked\ gene))}{\#masked\ genes}.$$

We chose to validate our metric and subsequent analyses on the best annotated genome to date: *E. coli* K12[59]. In order to remove as many *E. coli* K12 like subcontigs from the training dataset, we removed 5.2% of the subcontigs in which more than half of the genes were > 70% similar (calculated using mmseqs2 search[54]) in amino acid sequence to *E. coli* K12 genes. We validate our pseudo accuracy metric by calculating the absolute accuracy on the *E. coli* K12 genome for which each gene was masked sequentially:

$$absolute\ accuracy = \frac{\#count(argmin(dist(prediction, all\ genes\ in\ E.coli\ K12)) == index(masked\ gene))}{\#genes\ in\ E.coli\ K12}$$

485
486 ***Contextualized embedding calculation and visualization***
487 Contextualized protein embedding of a gene is calculated by first inputting a 15-30 gene subcontig
488 containing the gene of interest, and then running inference on the subcontig using the trained gLM without
489 masking. We then use the last hidden layer of the model corresponding to the gene as the embedding
490 consisting of 1280 features.
491
492 ***Gene annotation***
493 Genes were annotated using Diamond v2.0.7.145[60] against the UniRef90 database[61] with an e-value cut-off
494 1E-5. Genes were labeled as "unannotated" if either 1) no match was found in the UniRef90 database, or
495 2) the match was annotated with following keywords: "unannotated", "uncharacterized", "hypothetical",
496 "DUF" (domain of unknown function).
497
498 ***McrA protein analysis***
499 McrA protein encoding Methanogens and ANME genomes were selected from the accession ID list found
500 in the supplement of Shao et al[27]. Sub-contigs containing *mcrA* were extracted with at most 15 genes before
501 and after *mcrA.* The context-free and contextualized embeddings of McrA were calculated using the ESM2
502 and gLM respectively.
503
504 ***Enzyme Commission number prediction***
505 Enzyme Commission (EC) numbers were assigned to MGYPs by searching against the "KNN dataset"
506 DeepRe using mmseqs2[54]. Approximately 18% of the MGYPs could be assigned an EC number. We then
507 selected a random set of 100,000 subcontigs from the MGnify training data. First occurrences of MGYPs
508 with EC number were encoded into context-free (pLM) embeddings, using ESM2 mean-pooled last hidden
509 layer. Context-added (pLM+gLM) embeddings are generated by running gLM inference on the subcontig,
510 and then concatenating the pLM embedding with the last hidden layer of gLM corresponding to that MGYP.
511 We remove redundancies by making sure each data point consists of a unique MGYP. EC classes where
512 not all four numbers are assigned (e.g., 1.2.3._.) are removed from the dataset. All EC classes with less than
513 3 data points are removed, and subsequently split into train, validation, and test sets with 0.72/0.08/0.2 split,
514 with stratification on labels. A three-layer perceptron (**Extended Data 3**) is trained with ReLu activation,
515 batch normalization, and 0.2 dropout; EC number prediction is made by applying SoftMax on the final
516 layer.
517
518 ***Variance of contextualized protein embedding analysis***
519 Contextualized protein embeddings are generated at inference time. Variances of contextualized protein
520 embeddings were calculated for MGYPs that occur at least 100 times in the dataset, excluding the
521 occurrences at the edges of the subcontig (first or last token). For each such MGYP, we take 10 random
522 independent samples consisting of 100 occurrences and calculate the mean pairwise Euclidean distances
523 between the contextualized embeddings.
524
525 ***Attention analysis***
526 Attention heads (n = 190) were extracted by running inference on unmasked subcontigs, and the raw
527 attention weights were subsequently symmetrized. *E. coli* K12 RegulonDB[59] was used to probe heads with
528 attention patterns that correspond the most with operons. Pearson's correlation between symmetrized raw
529 attentions and operons were calculated for each head. We trained a logistic regression classifier that predicts
530 whether two neighboring genes belong to the same operon based on the attention weights across all attention
531 heads corresponding to the gene pair.
532
533 ***TnsC structural homolog analysis***
534 TnsC structural homologs were identified by searching ShCAST TnsC (PDB 7M99 chain H) against the
535 MGYP database using Foldseek on ESM Atlas (https://esmatlas.com/). The contigs containing these

536 homologs in the MGnify database were used to calculate the contextualized protein embeddings of the
537 identified structural homologs. Contigs with less than 15 genes were excluded from the analysis. Contigs
538 encoding proteins that were previously identified as "TnsC" using the UniRef90 database (see **Gene**
539 **annotation** methods section above) were included in the database. "TnsC-like" contigs were manually
540 annotated based on the presence of transposase genes (TnsB) and TniQ. Fifty random examples of MGnify
541 contigs containing MGYPs annotated as NuoA and DnaB were added as negative controls for the UMAP
542 visualization.
543
544 ***Paralogy and orthology analysis***
545 UniProt IDs from ABC transporter ModA and ModC protein interacting paralog pairs (n = 4823) were
546 previously identified by Ovchinnikov et al[49] and were downloaded from
547 https://gremlin.bakerlab.org/cplx.php?uni_a=2ONK_A&uni_b=2ONK_C and subsequently used to
548 download raw protein sequences from the UniProt server. Only pairs (n = 2700) where both raw sequences
549 were available for download, and where the UniProt ID differed by one (indicating adjacent positioning in
550 the reference genome) were selected for subsequent analyses. We constructed test contigs consisting of
551 three genes, where first and third genes are masked, and the second gene encodes one of the pair in forward
552 direction. We then queried gLM to predict the two neighboring masked genes, and considered the prediction
553 to be correct if either of the proteins closest to masked genes' highest confidence prediction in embedding
554 space belongs to the same sequence cluster as the interacting protein (50% amino acid sequence identity,
555 calculated using CD-HIT v4.6[62]). Random chance correct prediction rate was simulated using 1000
556 iterations of random predictions generated within the standard normal distribution and performing the same
557 operation as above to compute the rate of correct predictions.
558
559 ***Taxonomic analysis and visualization***
560 4551 bacterial and archaeal representative genomes and 11660 reference viral genomes were downloaded
561 from the RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq) on 12 Feb 2023. A random 30-gene
562 subcontig is chosen and encoded using ESM2, which then were subsequently concatenated with an
563 orientation vector and then used as input for the trained gLM. The last hidden layer was mean-pooled across
564 the sequence to retrieve 1280-feature contextualized contig embeddings. The ESM2 output protein
565 embeddings were also mean-pooled across the sequence to retrieve 1280-feature context-free contig
566 embeddings. We trained a logistic regression classifier to predict the class-level taxonomy of subcontigs
567 and evaluated the performance using stratified k-fold cross-validation (k=5).

Data and code availability

Data availability

Dataset used for training is available for download from the MGnify server (http://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/2022_05/). The model is available at zenodo under accession number 10.5281/zenodo.7855545.

Code availability

Training and inference code and analysis scripts are available at https://github.com/y-hwang/gLM.

Declarations

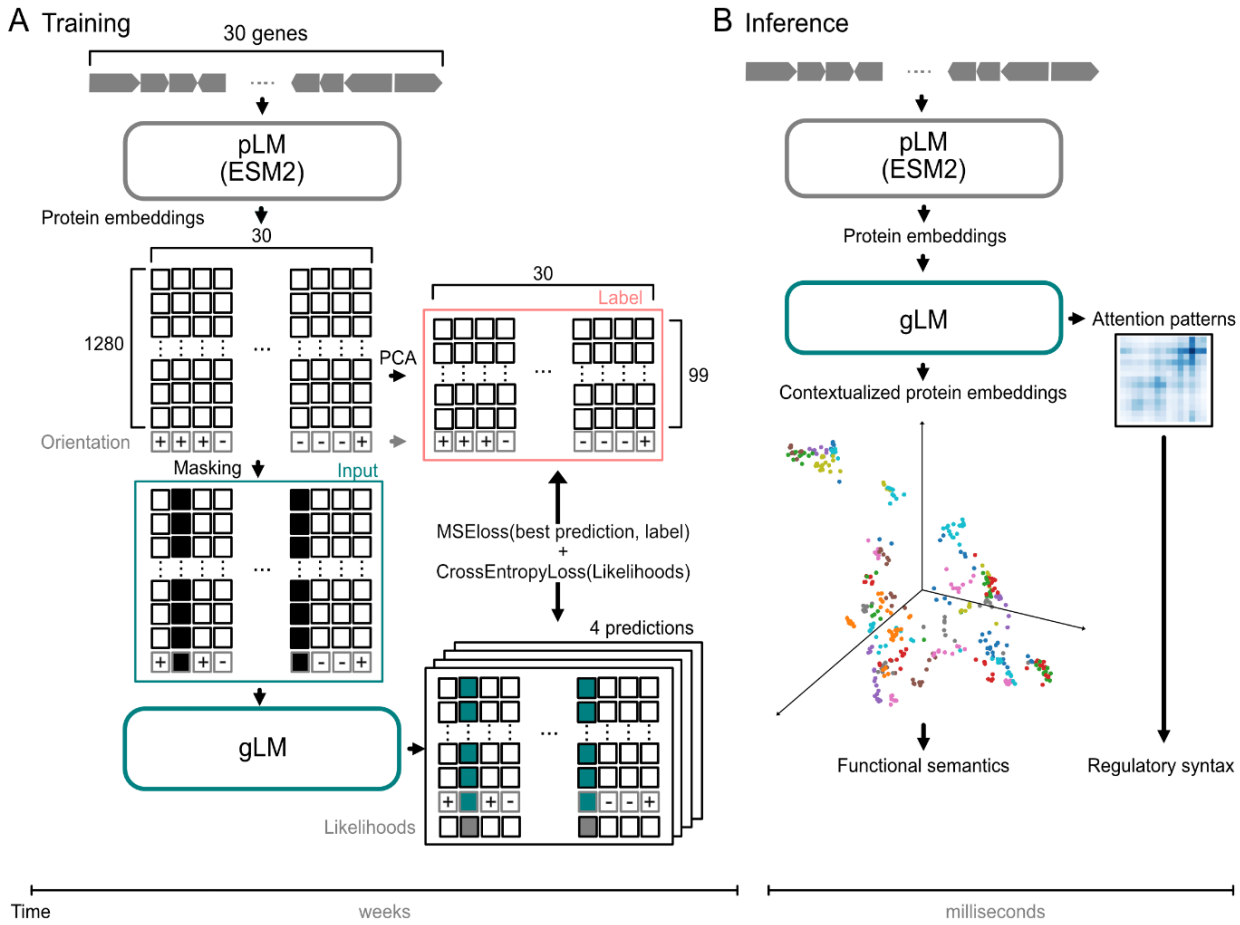
Acknowledgements. We would like to thank the EBI MGnify team for generating and maintaining the metagenome database. We would also like to thank Meta AI's ESM developers who made both the folded MGnify proteins structures and source-code openly available. We also thank Simon Roux and Landen Goszashti for insightful discussions.

Author contributions. YH prepared the datasets and trained the model with support from ALC and SO; ALC, SO and PRG provided input in analysis and data interpretation; EK and YH conducted AAA+ regulator analysis; YH wrote the manuscript with input from all authors; All authors read and approved the final manuscript.

Funding information. This work was supported by the Gordon and Betty Moore Foundation grant #9208 to P.R.G., NSF OCE-1635365 to P.R.G, and by the National Aeronautics and Space Administration under grant no. 80NSSC18K1140 and 80NSSC19K1427 issued through the NASA Network for Life Detection program to P.R.G. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

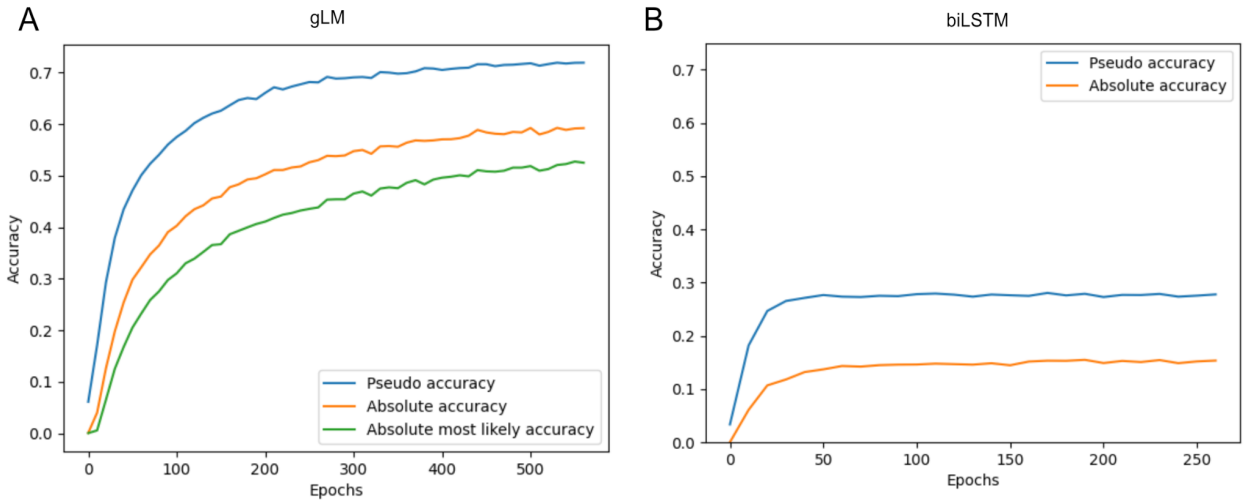
Conflicting interests. The authors declare that they have no conflict of interest.

Extended data



Extended Data 1. Detailed gLM training and inference schematics. A) Training begins with converting 15-30 gene metagenomic subcontigs to protein embeddings using ESM2. Orientation feature is concatenated for each protein and 15% of the proteins are masked randomly to generate training inputs. Labels are generated by applying PCA dimensionality reduction on the ESM2 protein embeddings, and concatenating the orientation feature. gLM is trained to make four possible predictions for the masked tokens, and their associated likelihoods. Training loss is calculated on both the prediction and likelihoods. The training stage takes several weeks on four NVIDIA A100 GPUs. **B)** At inference time, inputs are generated from a metagenomic subcontig using ESM2 output concatenated with an orientation feature. Hidden states and attention patterns of the trained gLM can be used for various downstream tasks.

607



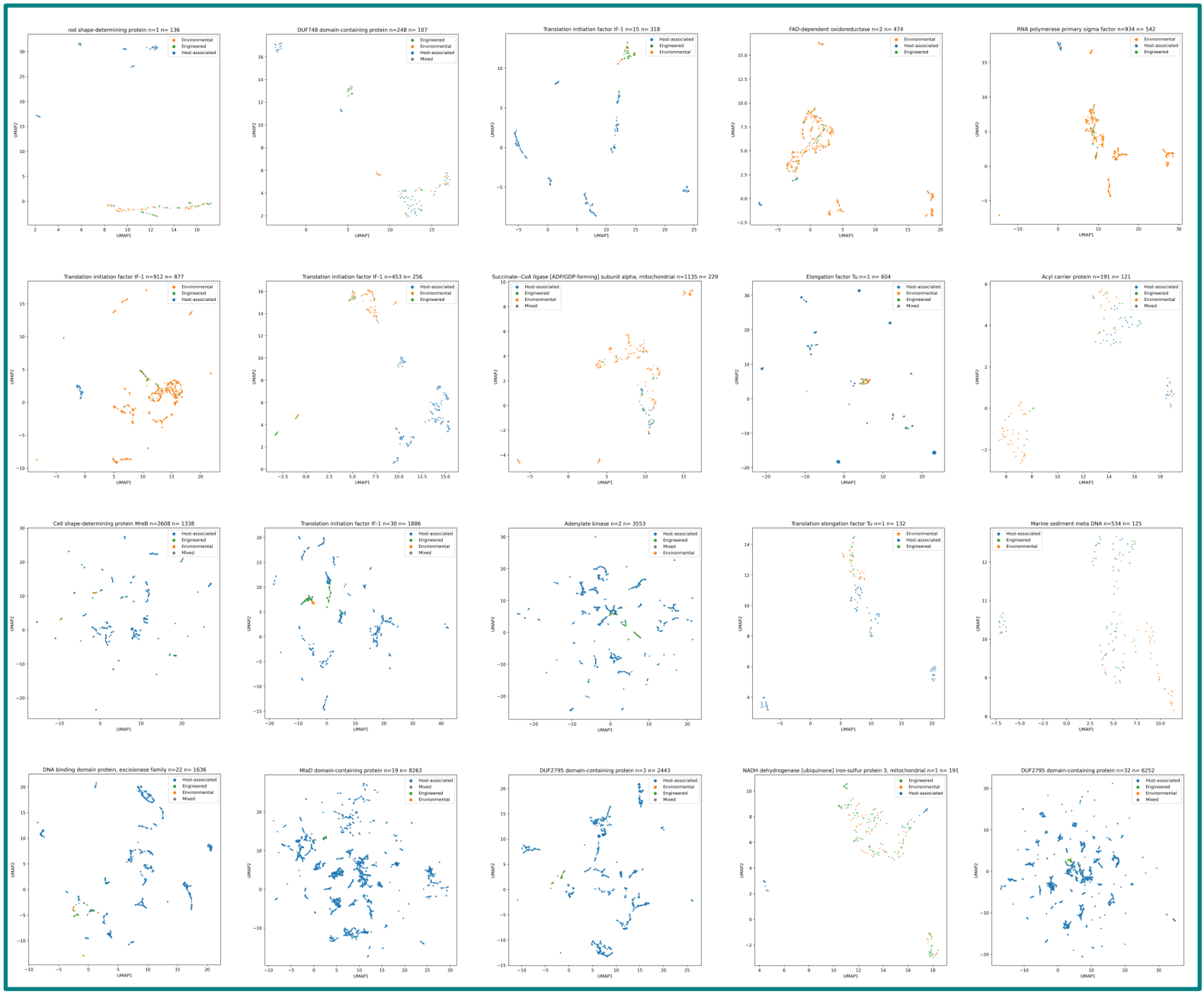
608

609 **Extended data 2. Validation accuracy curves for gLM (A) and biLSTM baseline (B).**

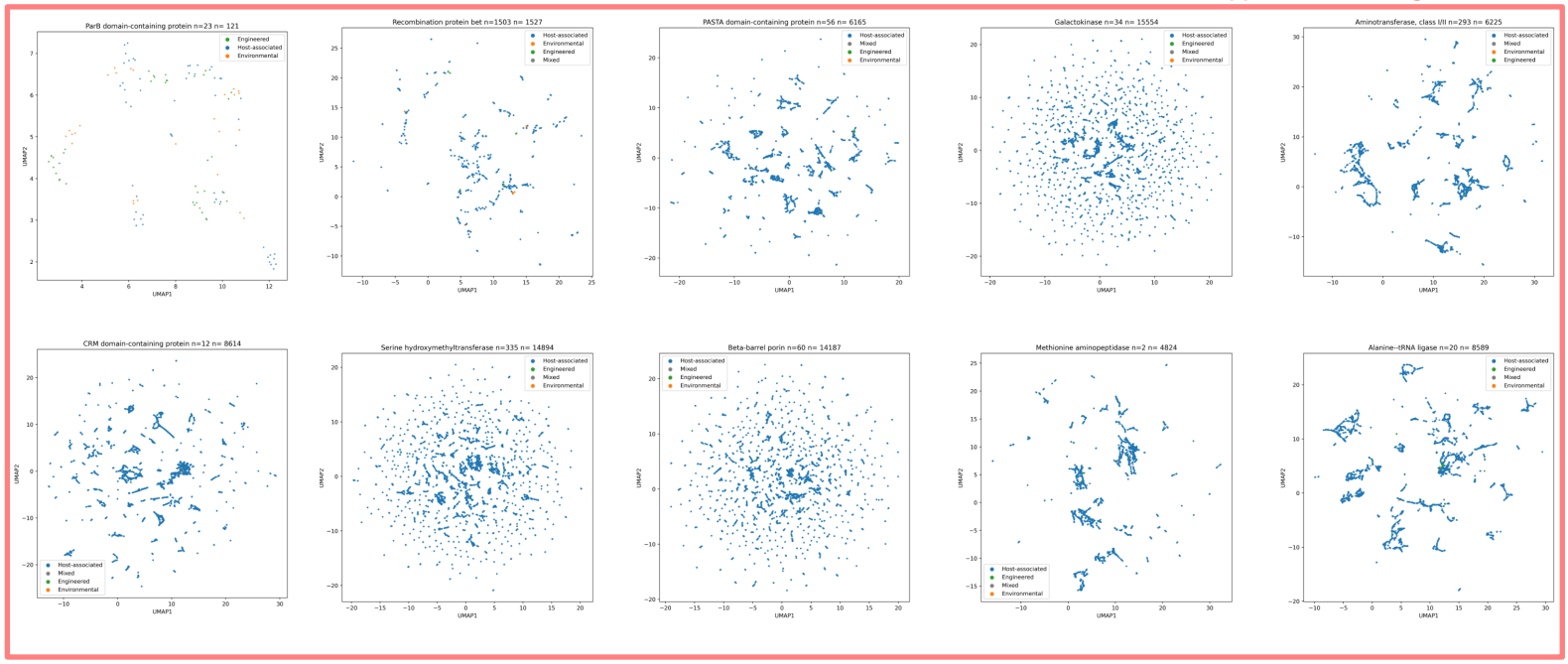
Extended data 3. Comparison of biLSTM baseline model with transformer-based gLM architecture and validation performances.

	biLSTM	gLM
Number of layers	5	19
Attention heads	N/A	20
Input embedding dimension	1281	1281
Hidden size	1280	1280
Batch size	4000	3000
Learning rate	1e-4	1e-4
Warm up steps	5000	5000
Training steps	1	1296960
Number of predictions	1	4
Number of parameters	27811840	954736916
Pseudo-accuracy (validation)	27.9	71.9
Absolute accuracy (validation)	14.78	59.2

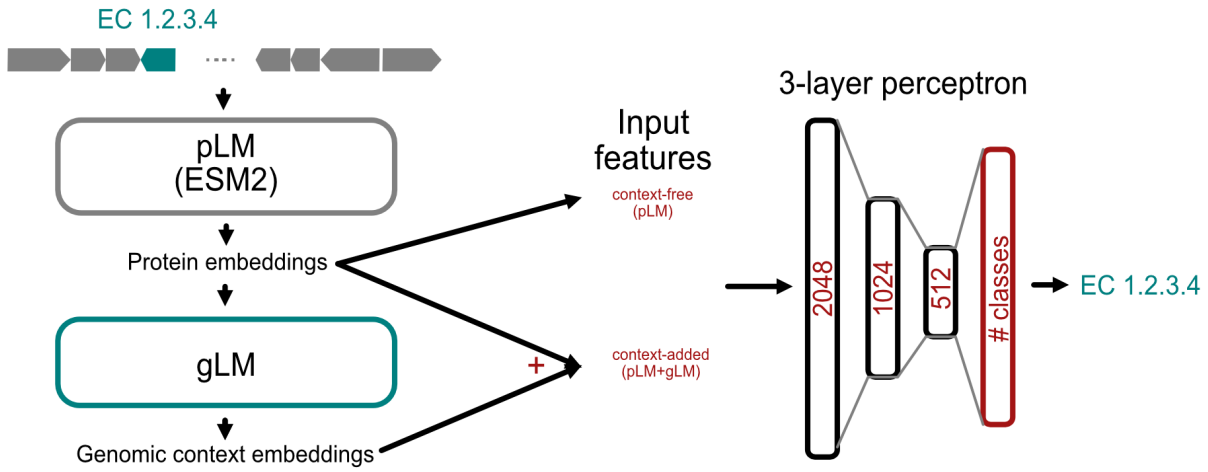
Clustering with biomes



No apparent clustering with biomes



615 **Extended Data 4. Contextualized protein embeddings for all 30 other multi-biome MGYPs. Biome**
616 color legend is identical to **Figure 2B**. Blue: Host-associated, Green: Engineered, Orange: Environmental,
617 Grey: Mixed.

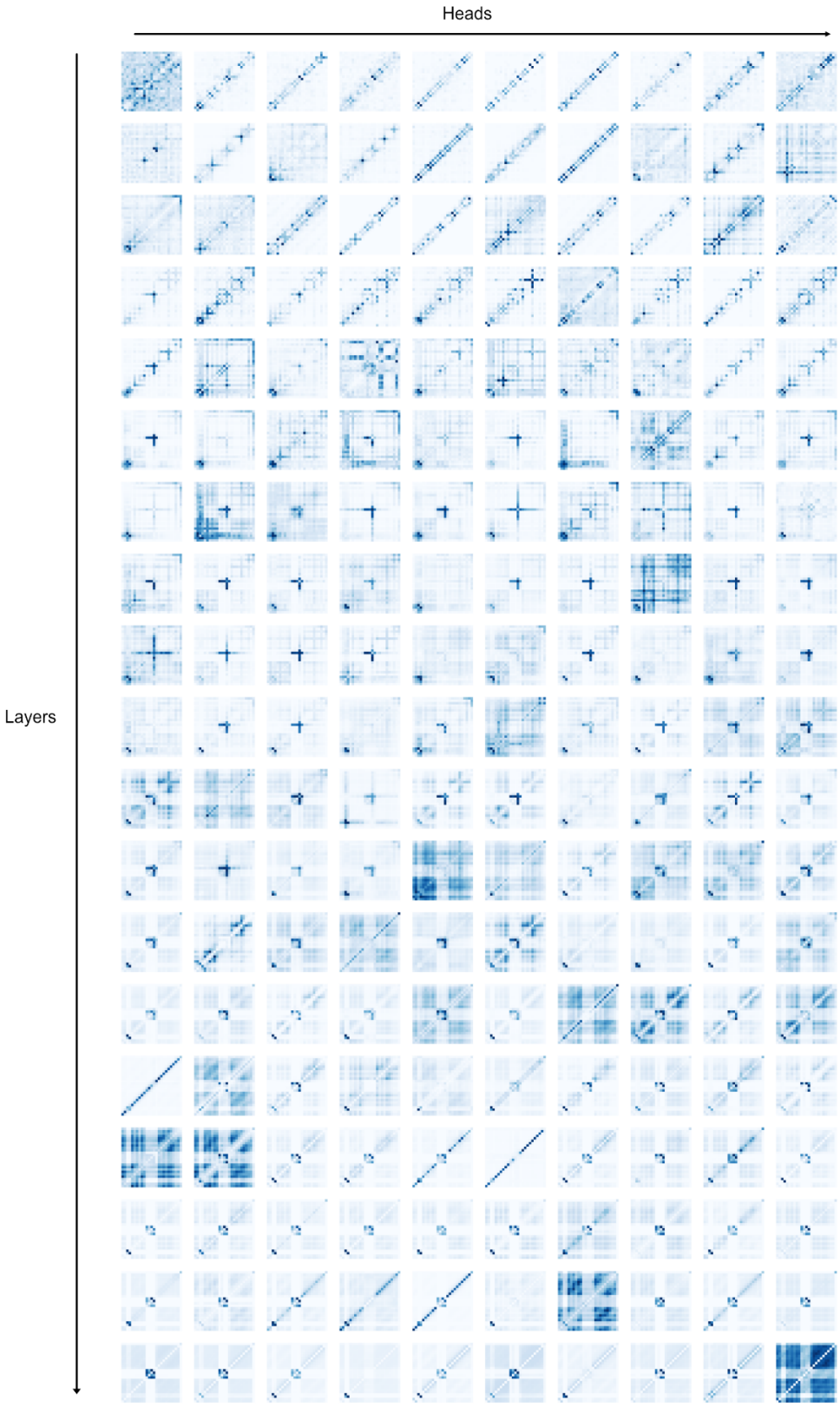


Extended Data 5. EC number prediction using a three-layer perceptron.

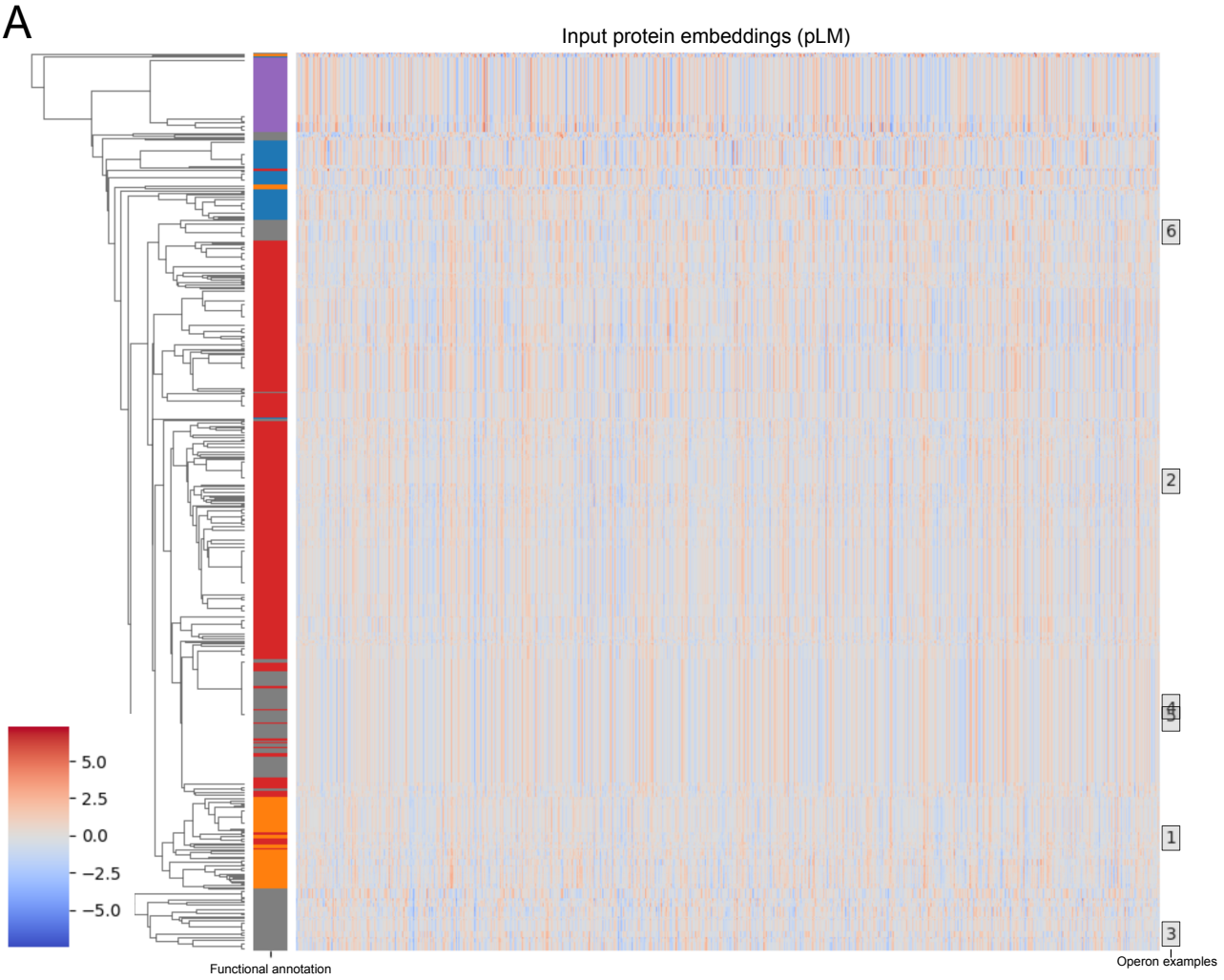
621 **Extended data 6. EC number prediction perceptron results.**
622
623

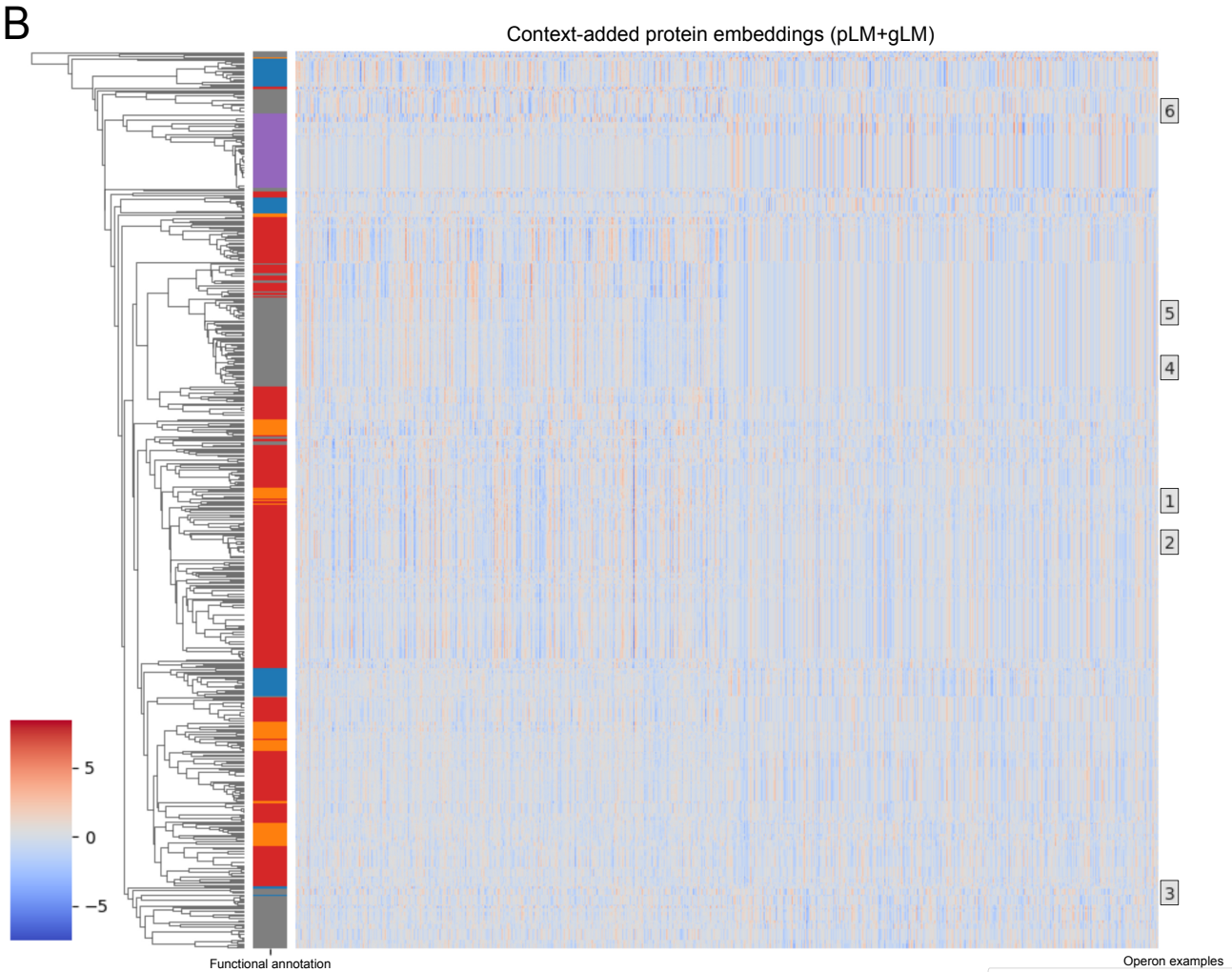
	Context-free representation (pLM)	Contextualized representation (gLM+pLM)
Validation accuracy	0.85	**0.875**
Test accuracy	0.850	**0.868**
Test mean average precision (“micro”-average)	0.725	**0.757**
Number of training epochs until peak validation accuracy (point of early stopping)	65	80

624



Extended data 7. Visualization of attention heads for a randomly chosen sequence. Increasing layer depth down the figure. Darker color represents larger attention scores.

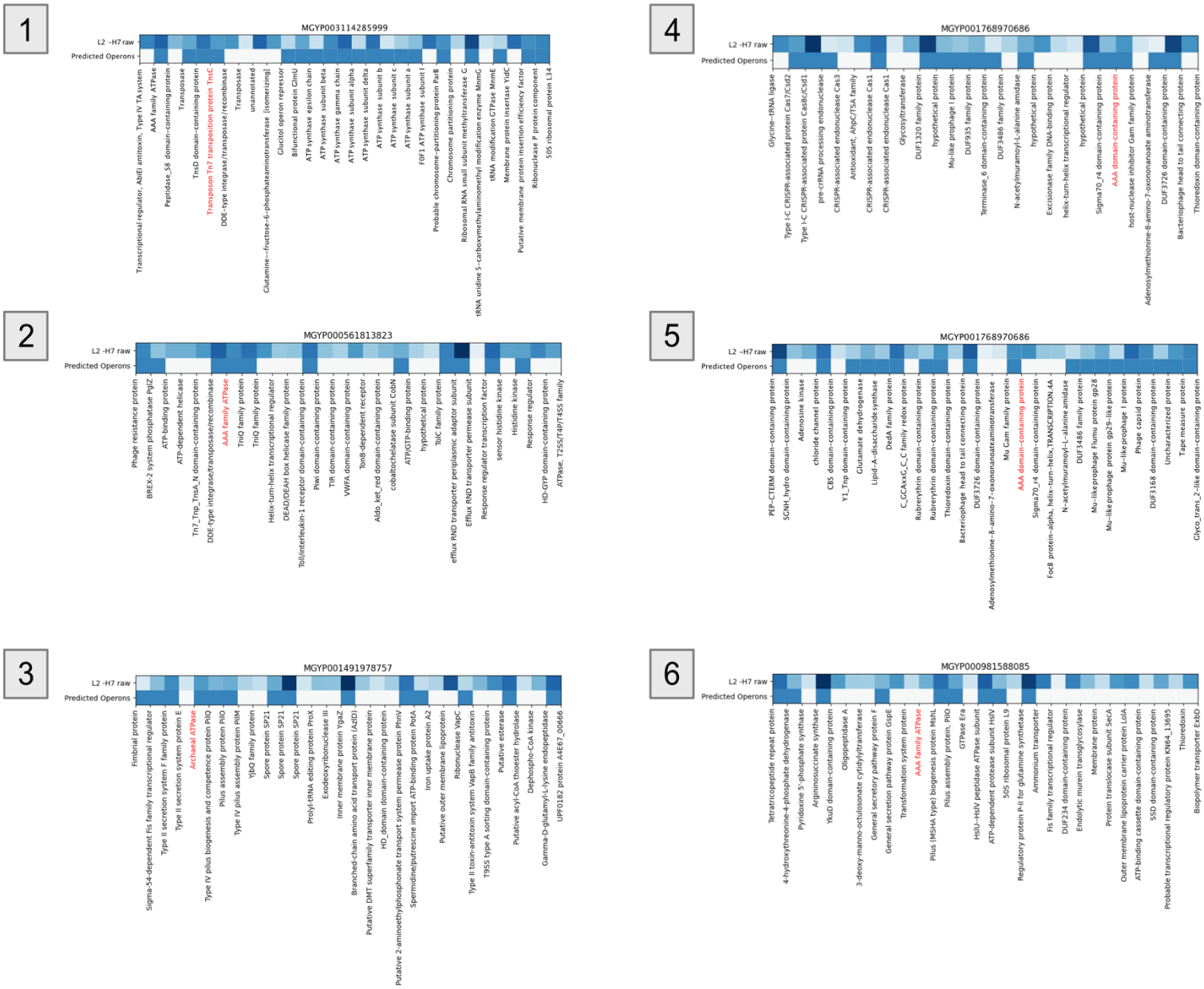




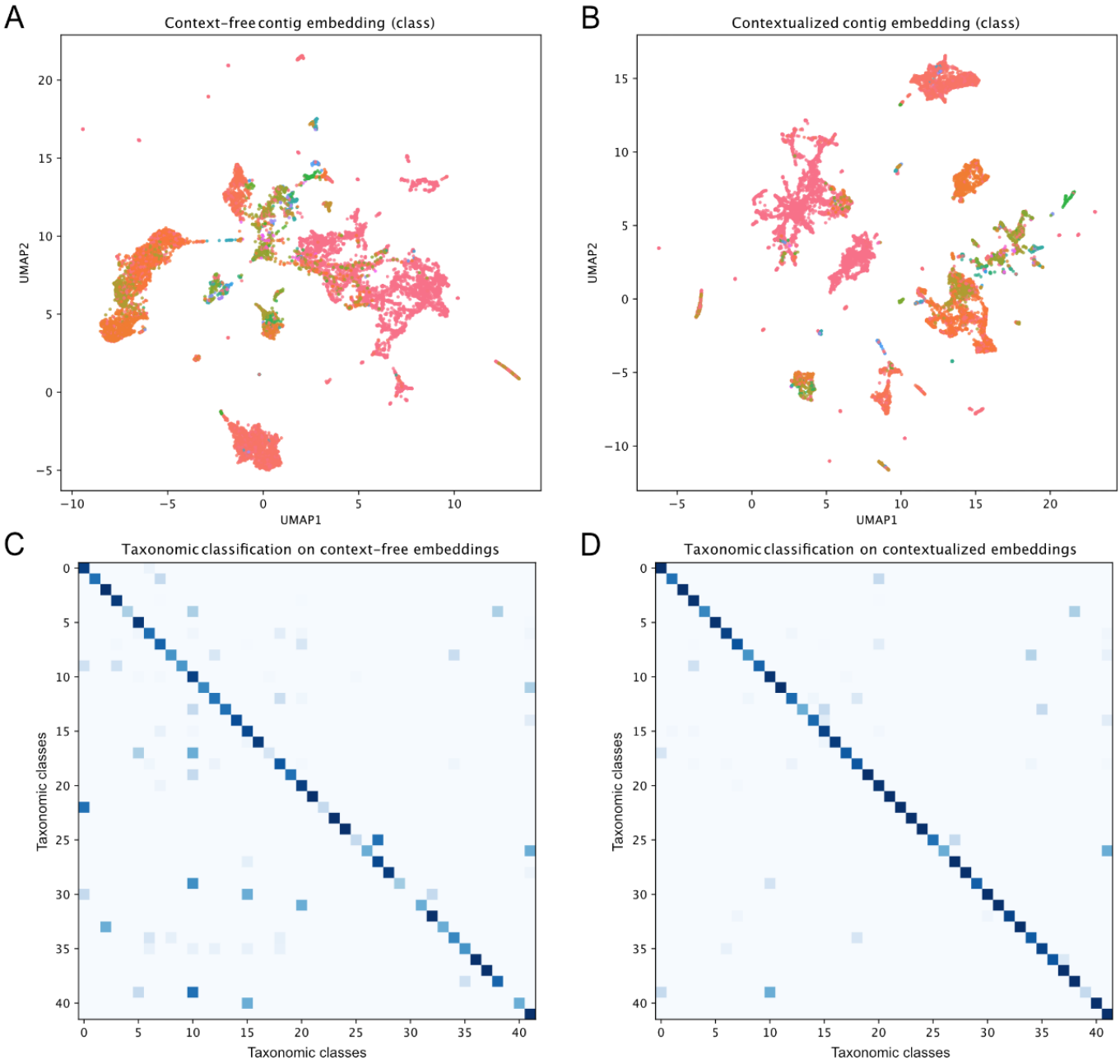
- TnsC-like AAA+ (manually inspected)
- AAA+ (Structural homologs of TnsC, functionally unrelated)
- TnsC (previously identified, pos. ctrl)
- DnaB helicase (functional neg. ctrl)
- NuoA (structural and functional neg. ctrl)

1. TnsC (prev. annotated)
2. TnsC-like (manually inspected in this study)
3. putatively associated with pilus assembly regulation
4. putatively associated with viral host-nuclease inhibition
5. putatively associated with Mu phage Gam protein
6. No predictable association

Extended Data 8. Distance based clustering of raw embeddings of AAA+ regulators and controls shown in Figure 4EF. Row colors and labels correspond to the point colors and designation used in **Figure 4EF** and **Extended Data 9**.



Extended Data 9. Attention-based operon prediction of AAA+ regulators. Annotations were made by searching against the UniRef90 database. Protein annotation in Red indicates the location of the AAA+ regulator identified based on a structural search of TsnC. Raw attention in the most correlated head (2nd layer, 7nd head, "L2-H7") is shown in the first row of the heatmap, and the second row is the predicted operon using the logistic regression predictor trained on *E.coli* K12 operon dataset.



Extended Data 10. Contextualization of contig embeddings and taxonomic classification. Comparison of context-free (A) and contextualized (B) class-level taxonomy of 30-gene contigs as shown in **Figure 5DE,** different colors represent different taxonomic classes. Confusion matrices for context-free (C) and contextualized (D) logistic regression classifier as shown in **Figure 5F**.

References

1. Redfern, O. C., Dessailly, B. & Orengo, C. A. Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* **18**, 394–402 (2008).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
4. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
5. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *arXiv [cs.LG]* (2020).
6. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* 1–8 (2023).
7. Wright, S. On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* **2**, 279–294 (1948).
8. Lynch, M. & Conery, J. S. The Origins of Genome Complexity. *Science* **302**, 1401–1404 (2003).
9. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).
10. Treangen, T. J. & Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
11. Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48–51 (2012).
12. Kountz, D. J. & Balskus, E. P. Leveraging Microbial Genomes and Genomic Context for Chemical Discovery. *Acc. Chem. Res.* **54**, 2788–2797 (2021).
13. Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language

processing. *Nat. Commun.* **13**, 5731 (2022).

14. Konno, N. & Iwasaki, W. Machine learning enables prediction of metabolic system evolution in bacteria. *Sci Adv* **9**, eadc9130 (2023).
15. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
16. Vaswani, Shazeer & Parmar. Attention is all you need. *Adv. Neural Inf. Process. Syst.*
17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).
18. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
19. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
20. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).
21. Jeffery, C. J. Protein moonlighting: what is it, and why is it important? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, (2018).
22. Miskei, M. *et al.* Fuzziness enables context dependence of protein interactions. *FEBS Lett.* **591**, 2682–2695 (2017).
23. Gherardini, P. F., Wass, M. N., Helmer-Citterich, M. & Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **372**, 817–845 (2007).
24. Ben-Hur, A. & Brutlag, D. Remote homology detection: a motif based approach. *Bioinformatics* **19 Suppl 1**, i26–33 (2003).
25. Bertram, S. *et al.* Methanogenic capabilities of ANME-archaea deduced from^{13}C-labelling approaches. *Environmental Microbiology* vol. 15 2384–2393 Preprint at https://doi.org/10.1111/1462-2920.12112 (2013).
26. Moran, J. J., House, C. H., Thomas, B. & Freeman, K. H. Products of trace methane oxidation during nonmethyltrophic growth by*Methanosarcina*. *Journal of Geophysical Research* vol. 112

696 Preprint at https://doi.org/10.1029/2006jg000268 (2007).
697 27. Shao, N. *et al.* Expression of divergent methyl/alkyl coenzyme M reductases from uncultured
698 archaea. *Commun Biol* **5**, 1113 (2022).
699 28. Reif, E. *et al.* Visualizing and measuring the geometry of BERT. *Adv. Neural Inf. Process. Syst.* **32**,
700 (2019).
701 29. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**,
702 (2022).
703 30. Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–
704 937 (2022).
705 31. Heinzinger, M. *et al.* Contrastive learning on protein embeddings enlightens midnight zone. Preprint
706 at https://doi.org/10.1101/2021.11.14.468528.
707 32. Egelhofer, V., Schomburg, I. & Schomburg, D. Automatic Assignment of EC Numbers. *PLoS*
708 *Computational Biology* vol. 6 e1000661 Preprint at https://doi.org/10.1371/journal.pcbi.1000661
709 (2010).
710 33. Melo, A. M. P., Bandeiras, T. M. & Teixeira, M. New Insights into Type II NAD(P)H:Quinone
711 Oxidoreductases. *Microbiol. Mol. Biol. Rev.* **68**, 603 (2004).
712 34. Schirmer, T. C-di-GMP Synthesis: Structural Aspects of Evolution, Catalysis and Regulation. *J.*
713 *Mol. Biol.* **428**, 3683–3701 (2016).
714 35. Walsh, D. A., Glass, D. B. & Mitchell, R. D. Substrate diversity of the cAMP-dependent protein
715 kinase: regulation based upon multiple binding interactions. *Curr. Opin. Cell Biol.* **4**, 241–251
716 (1992).
717 36. Mittenhuber, G. Phylogenetic analyses and comparative genomics of vitamin B6 (pyridoxine) and
718 pyridoxal phosphate biosynthesis pathways. *J. Mol. Microbiol. Biotechnol.* **3**, 1–20 (2001).
719 37. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
720 38. Ryu, J. Y., Kim, H. U. & Lee, S. Y. Deep learning enables high-quality and high-throughput
721 prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 13996–14001

722 (2019).
723 39. Li, Y. *et al.* DEEPre: sequence-based enzyme EC number prediction by deep learning.
724 *Bioinformatics* **34**, 760–769 (2017).
725 40. Rogers, A., Kovaleva, O. & Rumshisky, A. A primer in BERTology: What we know about how
726 BERT works. *Trans. Assoc. Comput. Linguist.* doi:10.1162/tacl_a_00349/96482.
727 41. Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models. *arXiv*
728 *[cs.CL]* (2020).
729 42. Salgado, H. *et al.* Using RegulonDB, the Escherichia coli K-12 Gene Regulatory Transcriptional
730 Network Database. *Curr. Protoc. Bioinformatics* **61**, 1.32.1–1.32.30 (2018).
731 43. White, S. R. & Lauring, B. AAA+ ATPases: achieving diversity of function with conserved
732 machinery. *Traffic* **8**, 1657–1667 (2007).
733 44. Park, J.-U. *et al.* Structures of the holo CRISPR RNA-guided transposon integration complex.
734 *Nature* **613**, 775–782 (2023).
735 45. Rybarski, J. R., Hu, K., Hill, A. M., Wilke, C. O. & Finkelstein, I. J. Metagenomic discovery of
736 CRISPR-associated transposons. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
737 46. Benler, S. *et al.* Cargo Genes of Tn7-Like Transposons Comprise an Enormous Diversity of Defense
738 Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. *MBio* **12**, e0293821 (2021).
739 47. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–
740 Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
741 48. van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv*
742 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
743 49. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue
744 interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
745 50. View article.
746 https://scholar.google.ch/citations?view_op=view_citation&hl=en&user=BA0mI18AAAAJ&sortby
747 =pubdate&citation_for_view=BA0mI18AAAAJ:WF5omc3nYNoC.

51. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).

52. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. *arXiv [cs.LG]* (2020).

53. Kiros, R., Salakhutdinov, R. & Zemel, R. Multimodal Neural Language Models. in *Proceedings of the 31st International Conference on Machine Learning* (eds. Xing, E. P. & Jebara, T.) vol. 32 595–603 (PMLR, 22--24 Jun 2014).

54. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

55. Lin, Z. *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv* 2022.07.20.500902 (2022) doi:10.1101/2022.07.20.500902.

56. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv [cs.CL]* (2019).

57. Huang, Z., Liang, D., Xu, P. & Xiang, B. Improve Transformer Models with Better Relative Position Embeddings. *arXiv [cs.CL]* (2020).

58. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv [cs.LG]* (2017).

59. Tierrafría, V. H. *et al.* RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb Genom* **8**, (2022).

60. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

61. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

62. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).