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)[4–7] 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[8–10]. These processes are particularly pronounced in bacterial and archaeal genomes where frequent horizontal gene transfers (HGT) shape genomic organization and diversity[11,12]. Thus, there exists an inherent evolutionary linkage between genes, their genomic context, and gene function[13–15], which can be explored by characterizing patterns that emerge from large metagenomic datasets.

Recent efforts to model genomic information have shown predictive power of genomic context in gene function[16] and metabolic trait evolution[17] in bacterial and archaeal genomes. However, these methods represent genes as categorical entities, despite these genes existing in continuous space where multidimensional properties such as phylogeny, structure, and function are abstracted in their sequences. On the other end of the spectrum of representations, there have been efforts to use unsupervised learning on nucleotide sequences to predict gene expression level[18] and detect regulatory motifs[19–21]. These models are largely trained and benchmarked on the human genome and focus on predicting gene regulation rather than function. Most recent efforts to leverage diverse microbial sequences to model genome-scale information include GenSLMs[22], which is pretrained on codon-level representations of diverse bacterial and viral gene sequences and later fine-tuned on SARS-CoV-2 genomes. In order to learn generalizable gene-to-gene-context interactions across biology, a model needs to be pretrained on 1) diverse lineages of organisms, 2) rich and continuous representation of genes and 3) longer segments of genomes with multiple genes. To our knowledge, there has been no method that combines all three aspects of pretraining to learn genomic information across diverse lineages of biology (see summary of previous studies in **Extended Data 1**).

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[23]. Our model is based on the transformer[24] architecture and is trained using millions of unlabelled metagenomic sequences via the masked language modeling[25] 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[25]), 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[25], where some fraction of the words are masked. Significant advances in language modeling performance was achieved by adopting the transformer[24] 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)[26] that sequentially processes tokens. To model genomic sequences, we trained a 19-layer transformer model (**Figure 1A;** for a detailed figure see **Extended Data 2**) on seven million metagenomic contig fragments consisting of 15 to 30 genes from the MGnify[27] database. Each gene in a genomic sequence is represented by a 1280 feature vector (context-

free protein embeddings) generated by using ESM2 pLM[23], 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 also predict their associated probabilities (**Extended Data 2**). 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[28] 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 3)**. 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 3 and 4,** note that biLSTM is smaller because it failed to converge when scaling the number of layers). We ablate the use of pLM representations as input to gLM by replacing them with one-hot amino acid representations (**Extended Data 5**) and report performance equivalent to random predictions (3% pseudo-accuracy and 0.02% absolute accuracy).

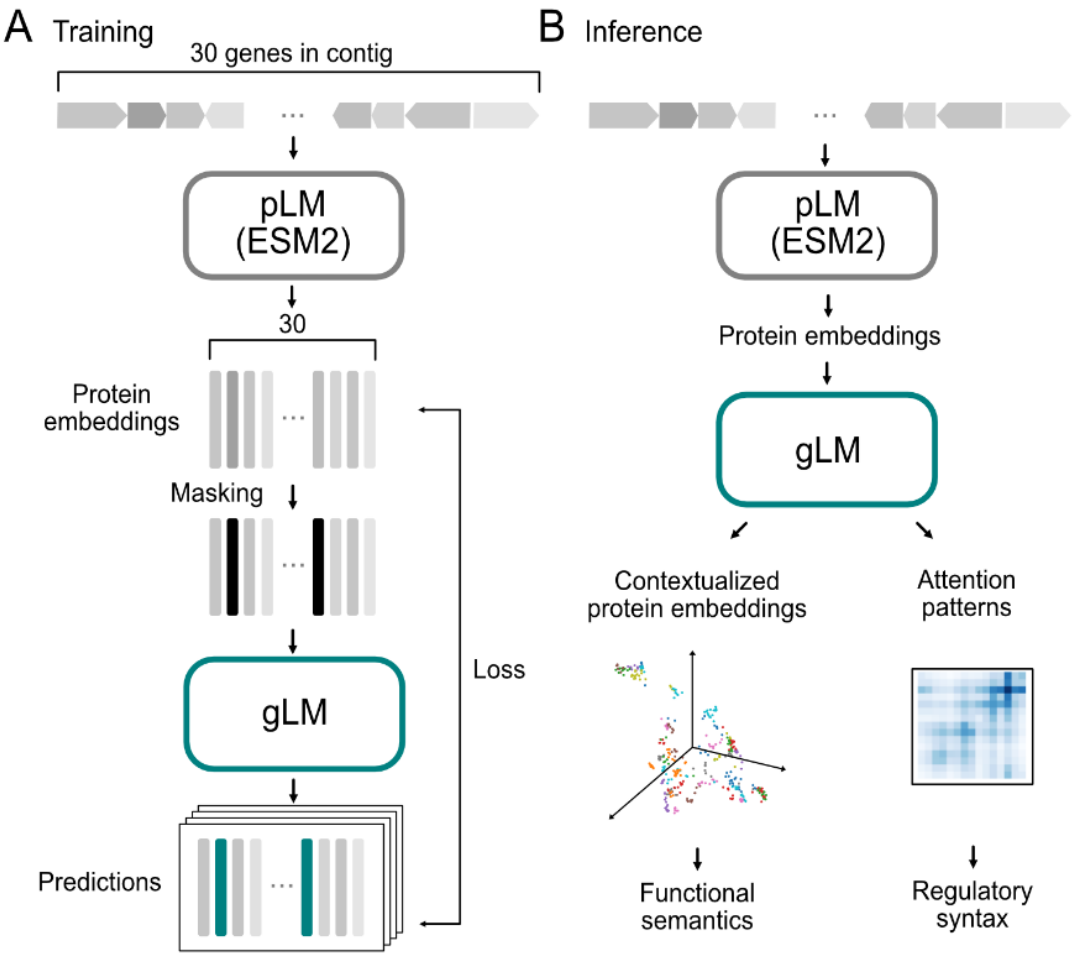


Figure 1. gLM training and inference schematics. A) For training, contigs (contiguous genomic sequences) containing up to 30 genes are first translated into proteins, which are subsequently embedded using a pLM encoder (ESM2). Masked inputs are generated by random masking at 15% probability and gLM (a transformer encoder) 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 (hidden layers of gLM) and attention patterns are used for various downstream tasks. See **Extended Data 2** for detailed schematics.

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[29] depending on its context[30], and many genes confer similar functions (i.e. convergent evolution[31], remote homology[32]). 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 representation) is identical across all occurrences, gLM's output (contextualized protein embeddings) cluster with biome types (**Figure 2B**; silhouette score = 0.17, see the other 30 multi-biome MGYP visualizations in **Extended Data 6**). 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 is able to 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, similar 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[33] and methanogens conducting methane oxidation in specific growth conditions[34]. 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 with methanogens (silhouette score after contextualization: 0.24; before contextualization: 0.027). 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[35]. As proposed by Shao et al.[35], 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)**. We quantitate this pattern with a higher silhouette score measuring phage defense and biosynthetic gene separation (gLM representation: 0.105±0.012, pLM representation: 0.078±0.011; paired t-test, t-statistic: 4.6, p-value = 0.001, n=10). Contextualized protein embeddings are therefore able to capture relational properties akin to semantic information[36], where genes encoding proteins that are more similar in their function are found in similar genomic contexts.

In order to quantify the information gain as a result of training a transformer on genomic contexts, we compare clustering results in 2B, F, and I with clustering conducted on (sub)contig-averaged pLM embeddings (**Extended Data 7**). By mean-pooling pLM embeddings across a given subcontig, we can summarize the context information as a naive baseline. We report a most consistent clustering (higher silhouette scores) of gLM embeddings compared to contig-averaged pLM for all three analyses (see **Extended Data 7** figure captions for values). We demonstrate that the gLM transformer model learns representations that correlate with biological function, which are not captured by the naive baseline.

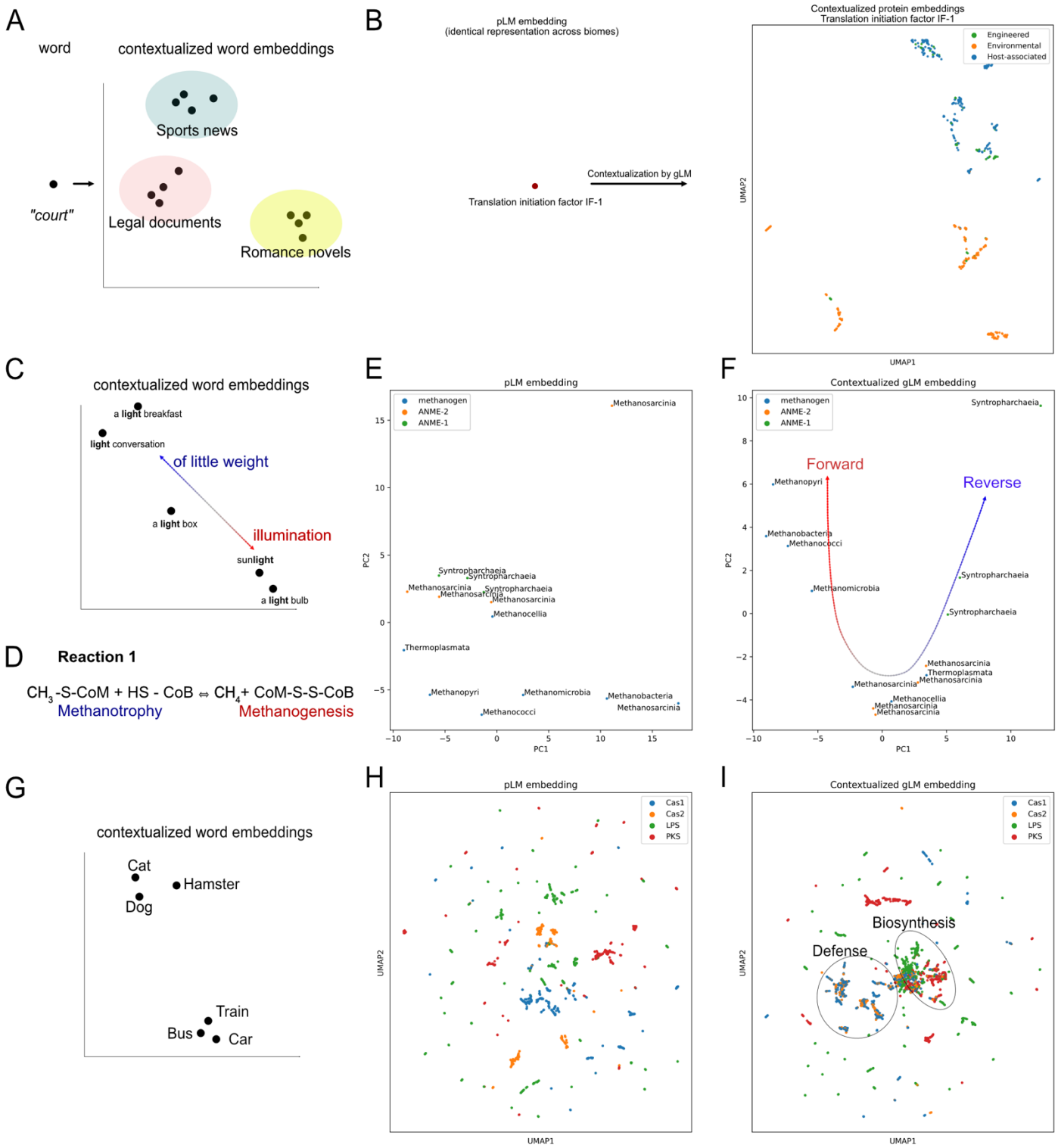


Figure 2. Contextualized protein embedding analysis and comparison with concepts in natural language modeling. A) A word upon contextualization can be mapped to embedding space. For many words, the semantic meaning varies in different types of literature, and therefore their contextualized embeddings 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 (A). Contextualization of 30 other multi-biome MGYPs can be found in **Extended Data 6**. **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)** Principal Component Analysis (PCA) of context-free protein embeddings of McrA sequences in genomes (total explained variances = 0.56), colored by metabolic classification of the organism (ANME, methanogen) based on previous studies and labeled by class-level taxonomy. **F)** PCA of contextualized McrA embeddings (total explained variance = 0.68), where gLM embeddings 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[37]. 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[38]). 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. In microbial genomes, proteins conferring 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. Based on this observation, we posited that contextualization would provide richer information that pushes the distribution of unannotated genes closer to the distribution of annotated genes. We compared the distributions of unannotated and annotated fractions of proteins in our dataset using context-free pLM embeddings and contextualized gLM embeddings. We found a statistically significant lower divergence between distributions of unannotated and annotated genes in gLM embeddings compared to pLM embeddings (paired t-test of Kullback-Leibler divergences, t-test statistic = 7.61, p-value < 1e-4, n = 10; see **Methods** for sampling and metric calculation). This suggests a greater potential for using gLM embeddings to transfer validated knowledge in cultivable and well-studied strains (e.g. *E. coli* K-12) 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 proteins can be used to aid function prediction, we evaluated how contextualization can improve the expressiveness of protein representations for enzyme function prediction. First, we generated a custom MGYP-EC dataset where the train and test data were split at 30% sequence identity for each EC class (see **Methods**). Second, we apply a linear probe (LP) to compare the expressiveness of representations at each gLM layer, with and without masking the queried protein (**Extended Data 8**). By masking the queried protein, we can assess gLM's ability to learn functional information of a given protein, only from its genomic context, without the propagation of information from the protein's pLM embeddings. We observed that a large fraction of contextual information pertaining to enzymatic function is learned in the first six layers of gLM. We also demonstrate that context information alone can be predictive of protein function, reaching up to 24.4 ± 0.8% accuracy. In contrast, without masking, gLM can incorporate information present in the context with the original pLM information for each queried protein. We observed an increase in expressivity of gLM embeddings also in the shallower layers, with accuracy reaching up to 51.6 ± 0.5% in the first hidden layer. This marks a 4.6 ± 0.5% increase from context-free pLM prediction accuracy (**Figure 3A)** and 5.5 ± 1.0% increase in mean average precision (**Figure 3C**) Thus, we demonstrate that information that gLM learns from the context is orthogonal to information captured in pLM embedding. We also observed diminishing expressivity in enzyme function information with deeper layers of gLM; this is consistent with previous examinations of LLMs, where deeper layers are specialized to the pretraining task (masked token prediction), and is consistent with previous examinations of LLMs, where the best-performing layer depends on the specific downstream tasks[39]. Finally, to further examine the expressiveness of these representations, we compared per-class F1 score gains (**Figure 3B**). We observe statistically significant differences in F1 scores (t-test, Benjamini/Hochberg corrected p-value < 0.05) between the two models in 36 out of 73 EC classes with more than ten samples in the test set. Majority (27 out of 36) of the statistical differences resulted in improved F1 score in LP trained on gLM representations.

Horizontal transfer frequency corresponds to genomic context embedding variance

A key process that shapes microbial genome organization and evolution is horizontal gene transfer (HGT). The taxonomic range in which genes are distributed across the tree of life depends on their function and the

selective advantage they incur in different environments. Relatively little is known about the specificity in the genomic region into which a gene gets transferred across phylogenetic distances. We examined the variance of gLM embeddings for proteins that occur at least one hundred times in the database. Variance of gLM-learned genomic contexts are calculated by taking a random sample of 100 occurrences and then calculating the mean pairwise distances between the hundred gLM embeddings. We conduct such independent random sampling and distance calculation ten times per gene and then calculate the mean value. As a baseline, we calculate variance of subcontig-averaged pLM embeddings using the same sampling method, to compare the information learned from training gLM. Our results show that gLM-learned genomic context variances have a longer right-hand tail (kurtosis = 1.02, skew = 1.08) compared to the contig-averaged pLM baseline that is more peaked (kurtosis = 2.2, skew = 1.05) **(Figure 3D)**. Notably, the most context-variant genes in the right tail of gLM-learned context variance distribution (orange) included phage genes and transposases, reflecting their ability to self-mobilize. Interestingly, we did not find any phage genes in the right-most tail of contig-averaged pLM embedding variance distribution (blue), although we did find genes involved in transposition (**Extended Data 9**). gLM-learned genomic context variances can be used as a proxy for horizontal transfer frequencies and can be used to compare the fitness effects of the genomic context on the evolutionary trajectory (e.g. gene flow) of genes, as well as to identify novel and functional transposable elements.

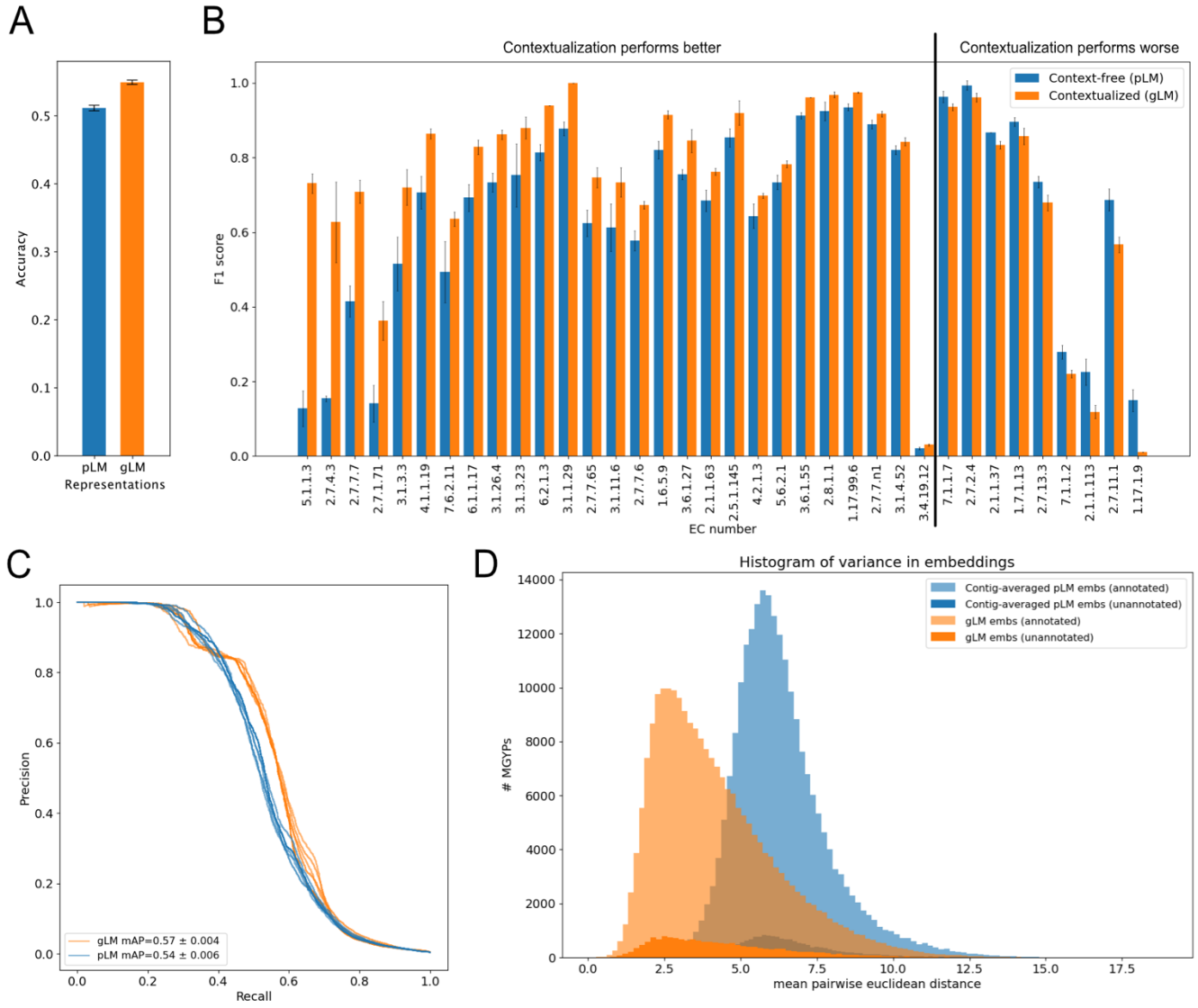


Figure 3. Contextualization of gene function. A) Linear probe EC classification accuracy for pLM (ESM2) representations and gLM (1st hidden layer) representations. **B)** F1-score comparisons of statistically significant (Benjamini/Hochberg corrected p-value < 0.05) differences in performance of pLM- and gLM-based EC number linear probes. EC classes are ordered with the largest gain with contextualization on the left to the largest loss with contextualization on the right. **C)** Precision-Recall curves of pLM- and gLM-based EC number linear probes. **D)** Histogram of variance (# bins = 100) calculated using contextualized embeddings (gLM; orange) and contig-averaged pLM (blue) 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 (see annotations of top tens most variant genes in **Extended Data 9**).

Transformer's attention captures operons

The transformer attention mechanism[24] models pairwise interaction between different tokens in the input sequence. Previous examinations of the attention patterns of transformer models in natural language processing (NLP)[39] have suggested that different heads appear to specialize in syntactic functions. Subsequently, different attention heads in pLMs[40] 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 pronounced in microbial genomes where multiple genes of related function are expressed as single polycistronic transcripts. Operons are prevalent in bacterial, archaeal and their viral genomes, while rare in eukaryotic genomes. We used the *E.coli* K-12 operon database[41] 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 10)**, with raw attention scores in the 7th head of the 2th 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 (operon predictor) 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 = 0.775±0.28, five-fold cross-validation) (**Figure 4C**). We baseline this performance by training an operon predictor on the one-hot amino acid representation-based gLM ablation (mean average precision = 0.426 ± 0.015, five-fold cross-validation; **Extended Data 5**), that learns from the orientation and co-occurrence information but cannot fully leverage rich representation of genes.

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[42]. 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[43] in Tn7-like transposon systems. Multiple bioinformatic efforts focused on discovery of novel transposons through metagenome search[44] and sequence searches of assembled genomes[45] aimed at identifying suitable homologs for genome-editing applications[46]. 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 (KL divergence ratio = 1.03; see **Methods** for calculation of this metric, **Figure 4E**). However, with added contextualization, previously identified TnsC (orange) and manually inspected TnsC-like structural homolog (red, labeled "TnsC-like") cluster together (KL divergence ratio = 0.38; **Figure 4F**; see **Extended Data 11BC** for comparison with gLM-only and contig-averaged pLM baselines). We further validate this visualization using embedding distance-based clustering (**Extended Data 12)**. 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 11A**). 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 11A, 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.

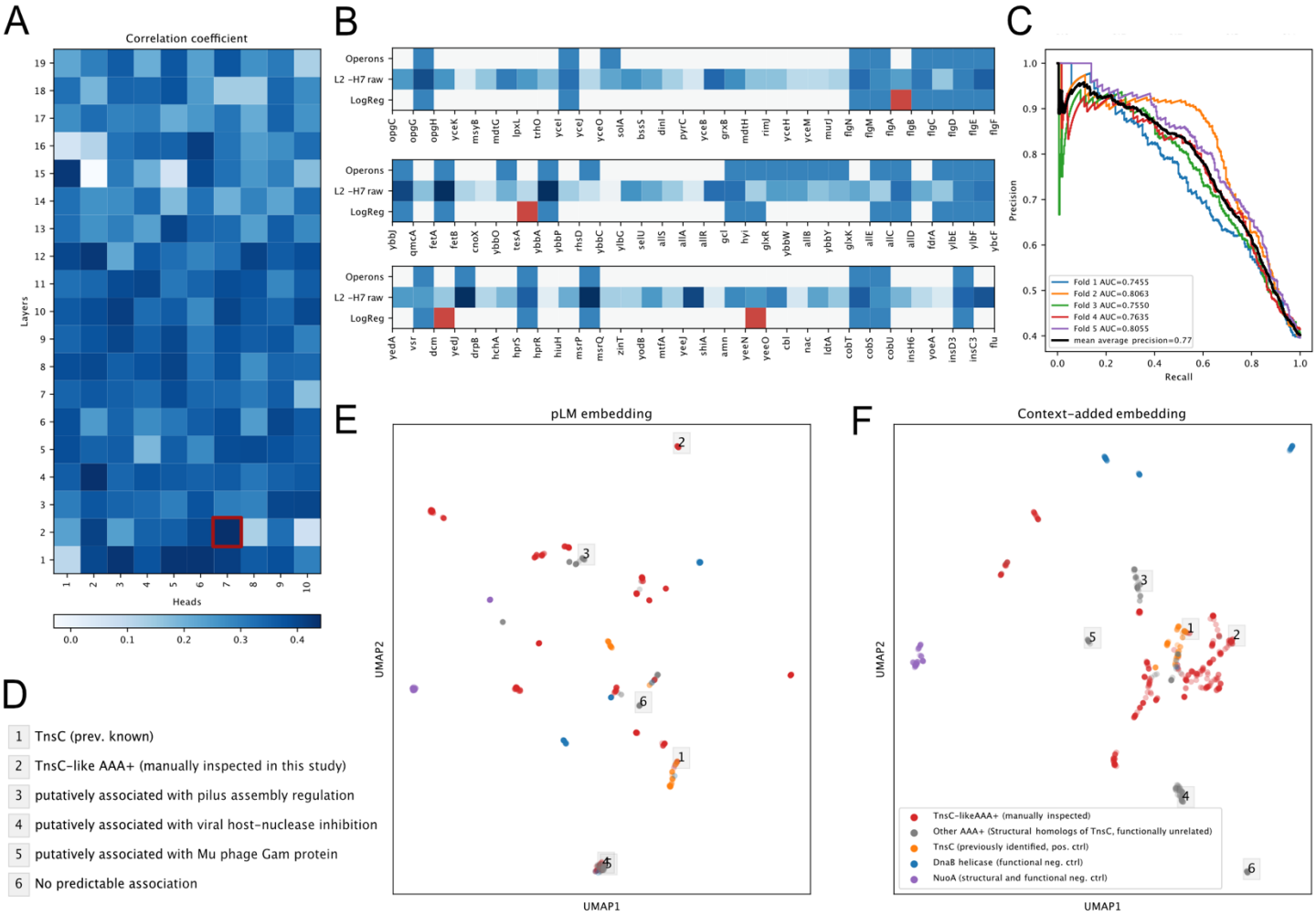


Figure 4. Attention analysis. A) Correlation coefficients (Pearson's rho) between attention heads across layers and operons. Darker color corresponds to stronger correlation with previously identified operons. Attention patterns of the second layer-seventh head [L2-H7] is most strongly correlated with the operons. **B)** Three random examples of contigs and predicted operonic relationship between neighboring proteins. Proteins are listed in the order they are encoded in the contig. Ground truth E.coli K-12 operons (top row), raw attention scores in the attention head [L2-H7] most correlated with operons (middle row) and logistic regression prediction using all attention heads (last row) where 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 11A**) 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[47]). 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 genes 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 11A.** Raw distance based clustering of the embeddings are shown in **Extended Data 12**.

354 ***gLM predicts paralogy in protein-protein interactions***
355 Proteins in an organism are found in complexes and interact physically with each other. Recent advances
356 in protein-protein interaction (PPI) prediction and structural complex research has largely been guided by
357 identifying interologs (conserved PPI across organisms) and co-evolutionary signals between residues[48].
358 However, distinguishing paralogs from orthologs (otherwise known as the "Paralog matching" problem)
359 in the expanding sequence dataset remains a computational challenge requiring queries across the entire
360 database and/or phylogenetic profiling. In cases where multiple interacting pairs are found within an
361 organism (e.g. histidine kinases (HK) and response regulators (RR)), prediction of interacting pairs is
362 particularly difficult[49]. We reasoned that gLM, although not directly trained for this task, may have learned
363 the relationships between paralogs versus orthologs. In order to test this capability, we used a well studied
364 example of interacting paralogs (ModC and ModA, **Figure 5A**) which form an ABC transporter complex.
365 We queried gLM to predict the embedding of an interacting pair given no context except the protein
366 sequence of either ModA or ModC. We find that without any fine-tuning gLM performs at least an order
367 of magnitude better than what is expected by random chance (see **Methods**). Specifically, for 398 out of
368 2700 interacting pairs, gLM makes predictions that belong to the same cluster (50% sequence identity,
369 n=2100 clusters) as the true label, and in 73 pairs, the gLM predicts a label that is closest to the exact
370 interacting pair (simulated random chance expected match=1.6 ± 1.01, n=10) (**Figure 5B)**. Importantly,
371 when considering only very high confidence predictions (prediction likelihood > 0.9, n=466), gLM is able
372 to match paralogs with an increased accuracy of 25.1%. When paralogs are correctly paired, gLM is more
373 confident about the prediction (average confidence for correct prediction = 0.79, average confidence across
374 all predictions = 0.53), while less certain predictions are either out of distribution, or closer to the mean of
375 labels (**Figure 5C**). We attribute part of the inaccuracies in prediction due to the fact that gLM was not
376 trained on the task of predicting a masked gene given only a single gene as genomic context, though we
377 expect the performance to improve with expanding the training sequence length range and fine-tuning the
378 model specifically for the "paralog matching" problem.
379
380 ***Contextualized contig embeddings and potential for transfer learning***
381 Contextualized protein embeddings encode the relationship between a specific protein and its genomic
382 context, retaining the sequential information within a contig. We hypothesized that this contextualization
383 adds biologically meaningful information that can be utilized for further characterization of the multi-gene
384 genomic contigs. Here, we define a contextualized contig embedding as a mean-pooled hidden layer across
385 all proteins in the subcontig, and a context-free contig embedding as mean-pooled ESM2 protein
386 embeddings across the sequence (see methods). Both embeddings consist of 1280 features. We test our
387 hypothesis by examining each of these embeddings' ability to linearly distinguish viral sequences from
388 bacterial and archaeal subcontigs. In metagenomic datasets, the taxonomic identity of assembled sequences
389 must be inferred post-hoc, therefore the identification of viral sequences is conducted based on the presence
390 of viral genes and viral genomic signatures[50]. However, such classification task remains a challenge
391 particularly for smaller contig fragments and less characterized viral sequences. Here, we sampled random
392 30-protein subcontigs from the representative bacterial and archaeal genome database and reference viral
393 genomes in the NCBI and visualized their context-free contig embeddings (**Figure 5D)** and contextualized
394 contig embeddings (**Figure 5E**). We observed more separation and taxonomic clusters at both domain- and
395 class-levels (**Extended Data 10)**, suggesting that taxonomic signature is enhanced by encoding the latent
396 relationships between proteins. This is further validated by training a logistic regression classifier on
397 context-free and contextualized contig embeddings for class-level taxonomy (**Extended Data 13AB**),
398 where we see a statistically significant improvement in average precision (**Figures 5F,** see **Extended Data**
399 **13CD** for confusion matrices). This emphasizes the biological importance of a protein's relative position
400 in the genome and its relationship with the genomic context, and further indicates that this information can
401 be effectively encoded using gLM. Contextualized contig embeddings present opportunities for transfer
402 learning beyond viral sequence prediction, such as improved metagenomically-assembled genome (MAG)
403 binning and assembly correction.

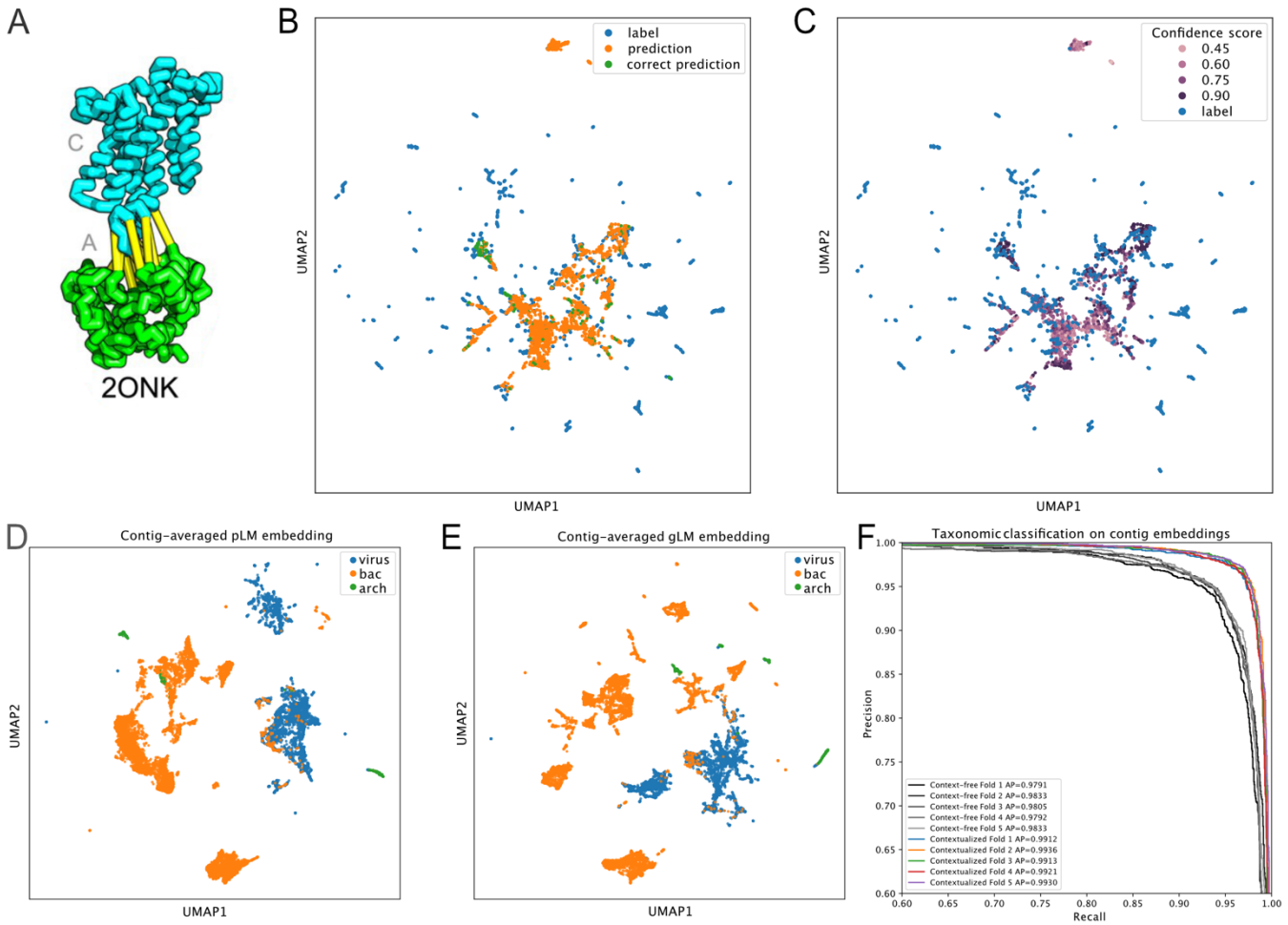


Figure 5. Potential for transfer learning. A) ModA and ModC interaction (adapted with permission from Ovchinnikov et al.[48]) **B)** UMAP projection of predictions (orange) and labels (blues) of paralogs (ModAC shown in A), where correct predictions are colored in green. **C)** Predicted embeddings are 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 (colored 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 13AB** and the confusion matrices for logistic regression classifiers are shown in **Extended Data 13CD.**

Discussion

The unprecedented amount and diversity of metagenomic data, coupled with advances in deep learning presents exciting opportunities for building models that can learn hidden patterns and structures of biological systems. Such models build upon the conceptual and statistical frameworks that evolutionary biologists have developed for the past century. With capabilities of abstracting much larger amounts of data, these models can disentangle the extraordinary complexity of organismal genomes and their encoded functions; this is a key step in furthering our understanding of biological processes. The work presented here demonstrates and validates the concept of genomic language modeling. Our implementation of the masked genomic language modeling illustrates the feasibility of training such a model, and provides evidence that biological information is being captured in learned contextualized embeddings and yielding meaningful interpretations of the attention patterns. We show that gLM can be used for diverse downstream tasks including enzyme function prediction, operon prediction, paralog matching and contig taxonomy prediction. Furthermore, we demonstrate gLM's ability to illuminate context dependency in functions across structural and sequence homology through the example of AAA+ regulators. Taken together, gLM presents a highly promising direction for interpreting biology and we propose key areas for further development: First, the transformer architecture has shown to be successful in efficient scaling; in both natural language[51] and protein language processing[23], increasing the number of parameters in the model along with the training dataset size have been shown to lead to vastly improved performance and generalizability. Our model consists of ~1B parameters which is at least a magnitude smaller compared to state-of-the-art pLMs. With further hyperparameter tuning and scaling, we expect better performance of the model. Second, our model currently uses pLM embeddings to represent proteins in the input. These embeddings are generated by mean-pooling the amino acid residue-level hidden states across the protein sequence, and therefore the residue specific information and synonymous mutation effects are likely obscured. Future iterations of the model could use raw residue-level or codon-level embeddings as input to allow modeling of residue-to-residue co-evolutionary interactions between proteins and synonymous mutation effects on gene function. Third, the task of reconstructing masked protein embeddings requires modeling a distribution over possible embeddings; our method approximates this distribution using a fixed number of predictions. Future work could improve upon this by using a generative approach, such as a diffusion or GAN model. This may allow for better prediction accuracy and greater generalizability for unseen datasets. Fourth, adding non-protein modalities (e.g. non-coding regulatory elements) as input to gLM may also greatly improve gLM's representation of biological sequence data, and can learn protein function and regulation conditioned upon other modalities[52]. Finally, our model was trained largely on bacterial, archaeal and viral genomes, therefore, how this method can be adapted for eukaryotic genomes, especially those with extensive intergenic regions, remains to be further explored.

One of the most powerful aspects of the transformer-based language models is their potential for transfer learning and fine-tuning. We tested some of the capabilities of gLM and successfully showed that higher order biological information including gene function and regulation can be learned using genomic sequences. Our results highlight the importance of contextualization of biological data, particularly as we scale our modeling efforts from biomolecules to whole organisms. We propose the following promising future directions for applying gLM for advancing biological research. 1) Feature-based transfer learning for predicting protein function (e.g. Gene Ontology [GO] term), particularly those with limited sequence and structural homology. 2) Fine-tuning gLM for the protein-protein-interactome prediction task. 3) Using gLM features to encode genomic contexts as additional input for improved and contextualized protein structure predictions. In conclusion, genomic language modeling is a powerful tool to unbiasedly condense important biological information from full metagenomic sequences. Coupled with the advances in long-read sequencing, we expect a drastic increase in the input data quality, quantity and diversity. Genomic language modeling presents an avenue to bridge the gap between atomic structure and organismal function, and thereby brings us closer to modeling biological systems, discovering novel biology, and ultimately, manipulating biology with precision (e.g. genome editing, synthetic biology).

Methods

Sequence database

The genomic corpus was generated using the MGnify[27] 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). We chose 30 as maximum context length because 67% of the raw MGnify contigs with > 15 genes were of =< 30 genes in length (**Extended Data 14**), and therefore increasing the context length beyond 30 would have resulted in many examples with padding (reduced computational efficiency). Each gene in the subcontig was mapped to a representative protein sequence (representative MGYP) using mmseqs/linclust[53], 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[54] "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 normalized (by subtracting the mean of each feature and dividing by its standard deviation) and clipped such that all features range from -10 to 10, to improve training stability. 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[55] transformer architecture. gLM consisted of 19 layers with hidden size 1280 and ten attention heads per layer, with relative position embedding ("relative_key_query")[56]. 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 (explained variance = 89.7%. **Extended Data 15**) of the corresponding ESM2 protein embedding concatenated with its orientation feature. Reduced dimensionality of labels using PCA increased the stability of training. 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[57] optimizer.

Performance metric and validation

In order 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[58]. 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[53]) 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: $absolute\ accuracy = \frac{\#count(argmin(dist(prediction, all\ genes\ in\ E.coli\ K12)) == index(masked\ gene))}{\#genes\ in\ E.coli\ K12}$

Contextualized embedding calculation and visualization

Contextualized protein embedding of a gene is calculated by first inputting a 15-30 gene subcontig containing the gene of interest, and then running inference on the subcontig using the trained gLM without masking. We then use the last hidden layer of the model corresponding to the gene as the embedding consisting of 1280 features.

Gene annotation

Genes were annotated using Diamond v2.0.7.145[59] against the UniRef90 database[60] with an e-value cut-off 1E-5. Genes were labeled as "unannotated" if either 1) no match was found in the UniRef90 database, or 2) the match was annotated with following keywords: "unannotated", "uncharacterized", "hypothetical", "DUF"(domain of unknown function).

McrA protein analysis

McrA protein encoding Methanogens and ANME genomes were selected from the accession ID list found in the supplement of Shao et al[35]. subcontigs containing *mcrA* were extracted with at most 15 genes before and after *mcrA.* The context-free and contextualized embeddings of McrA were calculated using the ESM2 and gLM respectively.

Distributions of unannotated and annotated embeddings

Distributions of unannotated and annotated embeddings in the database were compared using Kullback-Leibler (KL) divergence analysis. First, ten random samples of 10,000 subcontigs from the MGnify corpus. pLM and gLM embeddings of the genes were calculated using mean-pooled last hidden layer of ESM2 embeddings and mean-pooled last hidden layer of gLM respectively. Outliers were removed using Mahalanobis distance and a chi-squared threshold of 0.975. pLM and gLM embedding dimensions were reduced to 256 principal components (91.9 ± 1.72% and 80.1 ± 6.89% total variances explained respectively). KL divergence was calculated using the following equation:

$$D_{KL}(P||Q) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) - k + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) + ln(\frac{det\ \Sigma_1}{det\ 0}))$$

where P corresponds to the distribution of unannotated genes and Q corresponds to the distribution of unannotated genes, with μ_0, μ_1 respectively as means and Σ_0, Σ_1 respectively as covariance matrices. The significance of the KL divergence differences between pLM and gLM embeddings is calculated using a paired t-test across the ten samples.

Enzyme Commission number prediction

Custom MGYP-Enzyme Commission (MGYP-EC) dataset was created by first searching (mmseqs2[53] with default setting) MGYPs against the "split30.csv" dataset previously used to train CLEAN[61]. "split30.csv" dataset consists of EC numbers assigned to UniProt sequences clustered at 30% identity. Only MGYP hits with >70% sequences to "split30.csv" were considered and MGYPs with multiple hits with >70% similarity were removed. Test split was selected by randomly selecting 10% of "split30.csv" UniProt IDs in each EC category that map to MGYPs. EC categories with less than four distinct UniProt IDs with MGYP mapping were removed from the dataset, resulting in 253 EC categories. The train set consisted of MGnify subcontigs in the corpus that contained at least one the 27936 MGYPs mapping to 1878 UniProt IDs. The test set consisted of randomly selected MGnify subcontig containing each of 4441 MGYPs mapping to 344 UniProt IDs. pLM (context-free) embeddings were calculated for each of MGYP with EC number assignment by mean-pooling the last hidden layer of its ESM2 embedding. Masked (context-only) gLM embeddings were calculated for each of the 19 layers by running inference on subcontigs with masks at the positions of MGYPs with EC number assignment and subsequently extracting per-layer hidden representations for

masked positions. gLM (contextualized) embeddings were calculated also for each layer by running inference without masking and subsequently extracting per-layer hidden representations for MGYPs with EC number assignments. Linear probing was conducted for these embeddings with a single linear layer. Linear probes were trained with early stopping (patience = 10, github.com/Bjarten/early-stopping-pytorch/blob/master/pytorchtools.py) and batch size = 5000, and training results were replicated five times with random seeds to calculate error ranges.

Variance of contextualized protein embedding analysis

Contextualized protein embeddings are generated at inference time. Variances of contextualized protein embeddings were calculated for MGYPs that occur at least 100 times in the dataset, excluding the occurrences at the edges of the subcontig (first or last token). For each such MGYP, we take 10 random independent samples consisting of 100 occurrences and calculate the mean pairwise euclidean distances between the contextualized embeddings. To assess the role gLM plays in contextualization, we used the above sampling method to calculate the variance of contig-averaged pLM embeddings (pLM embeddings mean-pooled across the contig) for each MGYP that occurs at least 100 times in the dataset.

Attention analysis

Attention heads (n = 190) were extracted by running inference on unmasked subcontigs, and the raw attention weights were subsequently symmetrized. *E.coli* K12 RegulonDB[58] was used to probe heads with attention patterns that correspond the most with operons. Pearson's correlation between symmetrized raw attentions and operons were calculated for each head. We trained a logistic regression classifier that predicts whether two neighboring genes belong to the same operon based on the attention weights across all attention heads corresponding to the gene pair.

TnsC structural homolog analysis

TnsC structural homologs were identified by searching ShCAST TnsC (PDB 7M99 chain H) against the MGYP database using Foldseek on ESM Atlas (https://esmatlas.com/). The contigs containing these homologs in the MGnify database were used to calculate the contextualized protein embeddings of the identified structural homologs. Contigs with less than 15 genes were excluded from the analysis. Contigs encoding proteins that were previously identified as "TnsC" using the UniRef90 database (see **Gene annotation** methods section above) were included in the database. "TnsC-like" contigs were manually annotated based on the presence of transposase genes (TnsB) and TniQ. Fifty random examples of MGnify contigs containing MGYPs annotated as NuoA and DnaB were added as negative controls for the UMAP visualization. We calculated KL divergence ratios using the following formula $\frac{D_{KL}(B||A)}{D_{KL}(C||A)}$ where A is the distribution of representations of known TnsC, B is the distribution of representations of manually curated TnsC-like AAA+ regulators, C is the distribution of representations of other AAA+ regulators that are functionally unrelated structural homologs of known TnsC. Therefore, this metric ranges from 0 to 1, where a lower ratio represents increased ability to functionally discriminate distribution of B from C relative to A. KL divergence was calculated using the same formula as in the methods section **Distributions of unannotated and annotated embeddings**, except with 20 principal components that explained >85% of variances across all embeddings.

Paralogy and orthology analysis

UniProt IDs from ABC transporter ModA and ModC protein interacting paralog pairs (n = 4823) were previously identified by Ovchinnikov et al[48] and were downloaded from https://gremlin.bakerlab.org/cplx.php?uni_a=2ONK_A&uni_b=2ONK_C and subsequently used to download raw protein sequences from the UniProt server. Only pairs (n = 2700) where both raw sequences were available for download, and where the UniProt ID differed by one (indicating adjacent positioning in the reference genome) were selected for subsequent analyses. We constructed test contigs consisting of three genes, where first and third genes are masked, and the second gene encodes one of the pair in forward

direction. We then queried gLM to predict the two neighboring masked genes, and considered the prediction to be correct if either of the proteins closest to masked genes's highest confidence prediction in embedding space belongs to the same sequence cluster as the interacting protein (50% amino acid sequence identity, calculated using CD-HIT v4.6[62]). Random chance correct prediction rate (1.6±1.0 was simulated using 1000 iterations of random predictions generated within the standard normal distribution and performing the same operation as above to compute the rate of correct predictions.

Taxonomic analysis and visualization

4551 bacterial and archeal representative genomes and 11660 reference viral genomes were downloaded from the RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq) on 12 Feb 2023. A random 30-gene subcontig is chosen and encoded using ESM2, which then were subsequently concatenated with an orientation vector and then used as input for the trained gLM. The last hidden layer was mean-pooled across the sequence to retrieve 1280-feature contextualized contig embeddings. The ESM2 protein embeddings were also mean-pooled across the sequence to retrieve 1280-feature context-free contig embeddings. We trained a logistic regression classifier to predict the class-level taxonomy of subcontigs and evaluated the performance using stratified k-fold cross-validation (k=5).

UMAP visualization and statistical tests

All UMAP dimensionality reductions calculated with following parameters: n_neighbors = 15, min_dist = 0.1. Silhouette scores were calculated using the *sklearn* package using the default setting with euclidean distance metric.

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; YH wrote the manuscript with input from all authors; All authors read and approved the final manuscript.

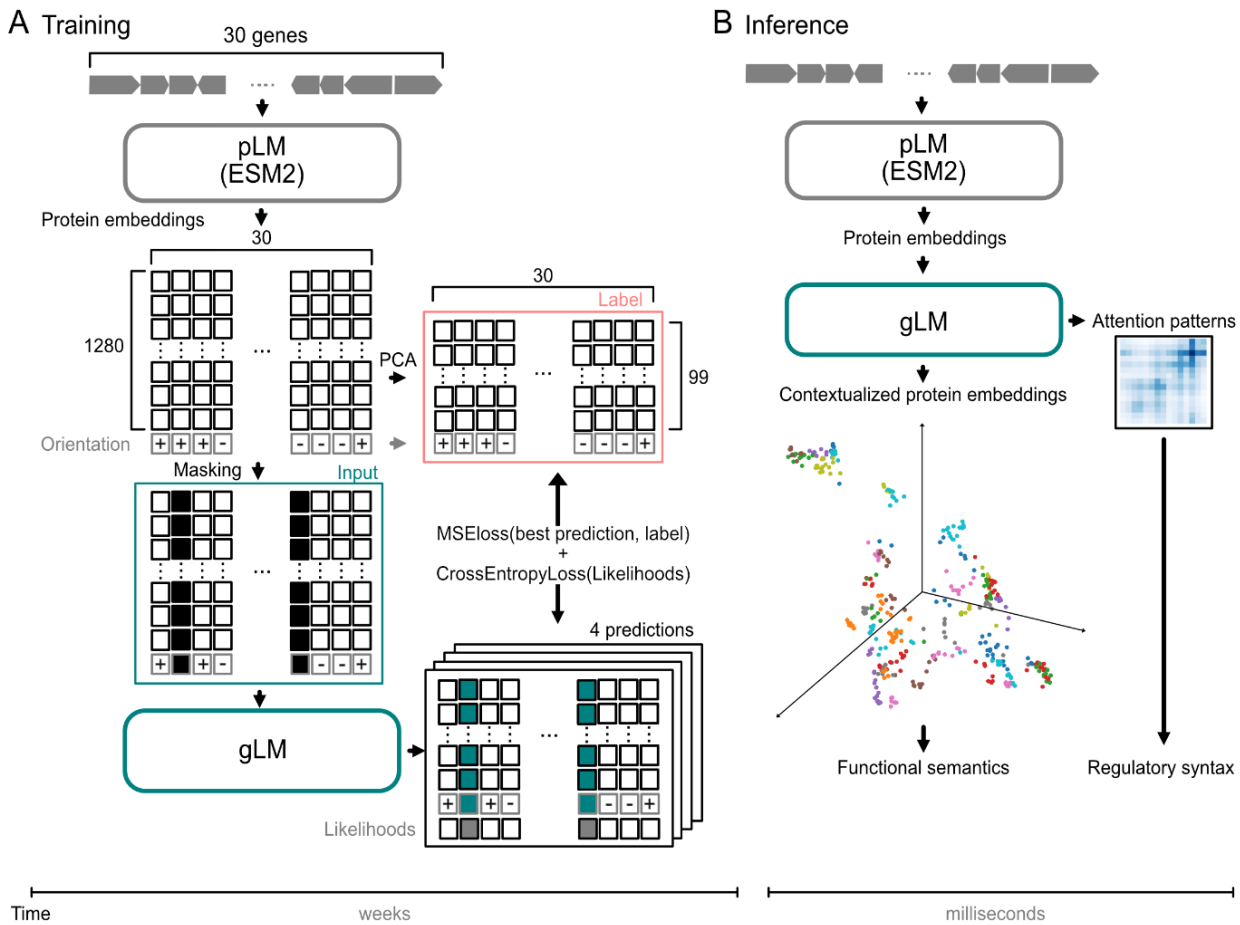
Funding information. This work was supported by the Gordon and Betty Moore Foundation grant #9208 to P.R.G., NSF OCE-1635365 to P.R.G, and by the National Aeronautics and Space Administration under grant no. 80NSSC18K1140 and 80NSSC19K1427 issued through the NASA Network for Life Detection program to 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.

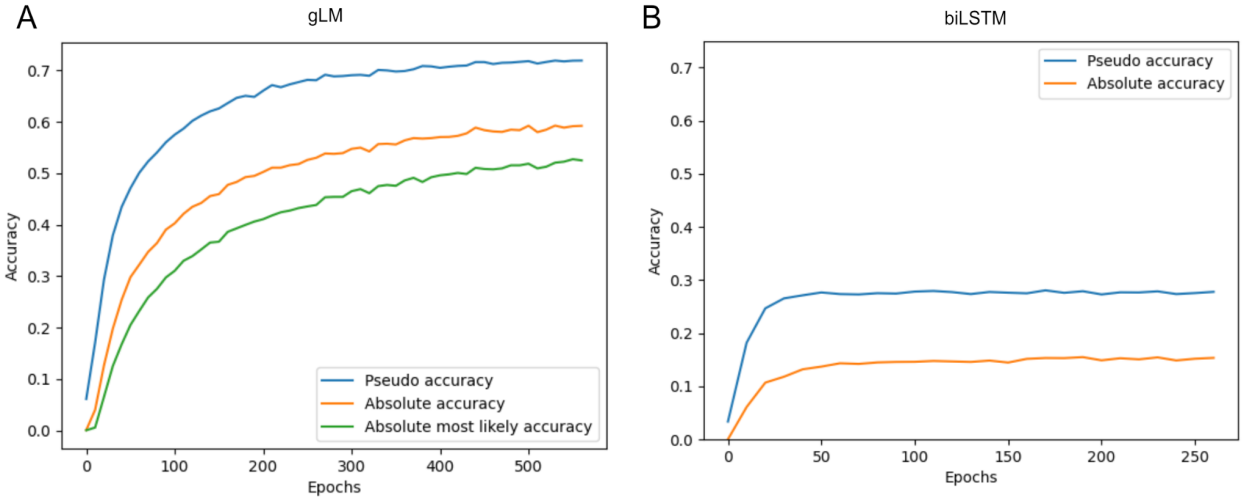
Extended data

Extended Data 1. Comparison of gLM to previous efforts in modeling various aspects of biological sequences.

	Multi-gene input	Continuous representation of genes	Generalizable across organisms (Organism-agnostic pretraining)	Self-supervised language model
gLM (this study)	✓	✓	✓ (Metagenomic sequences with bias towards bacteria, archaea and viruses)	✓
pLMs[4–7](e.g. ESM2, ProtBert, ProtT5, ProGen, CaLM)	×	✓	✓	✓
Miller et al[16]	✓	×	✓	×
Enformer[18]	✓	✓	× (Pretrained on human and mouse genomes only)	×
DNABERT[19]	× (Max context length of DNABERT-6 is 3072 bp, which is not sufficient to include a median length (26,288 bp) human protein coding gene[63])	✓	× (Pretrained on human genome)	✓
Nucleotide Transformer[20]	× (Max context length is 6000 bp, which is not sufficient to include a median length (26,288 bp) human protein coding gene)	✓	× (Heavily biased towards human genome)	✓
HyenaDNA[21]	✓	✓	× (Pretrained on Human genome)	✓
GenSLM-foundation model[22]	× (Single genes used for pretraining)	✓	✓	✓
GenSLM-SARS-CoV2 genome model[22]	✓	✓	× (fine-tuned on SARS-CoV2 genomes only)	✓



Extended Data 2. 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.



Extended data 3. Validation accuracy curves for gLM (A) and biLSTM baseline (B).

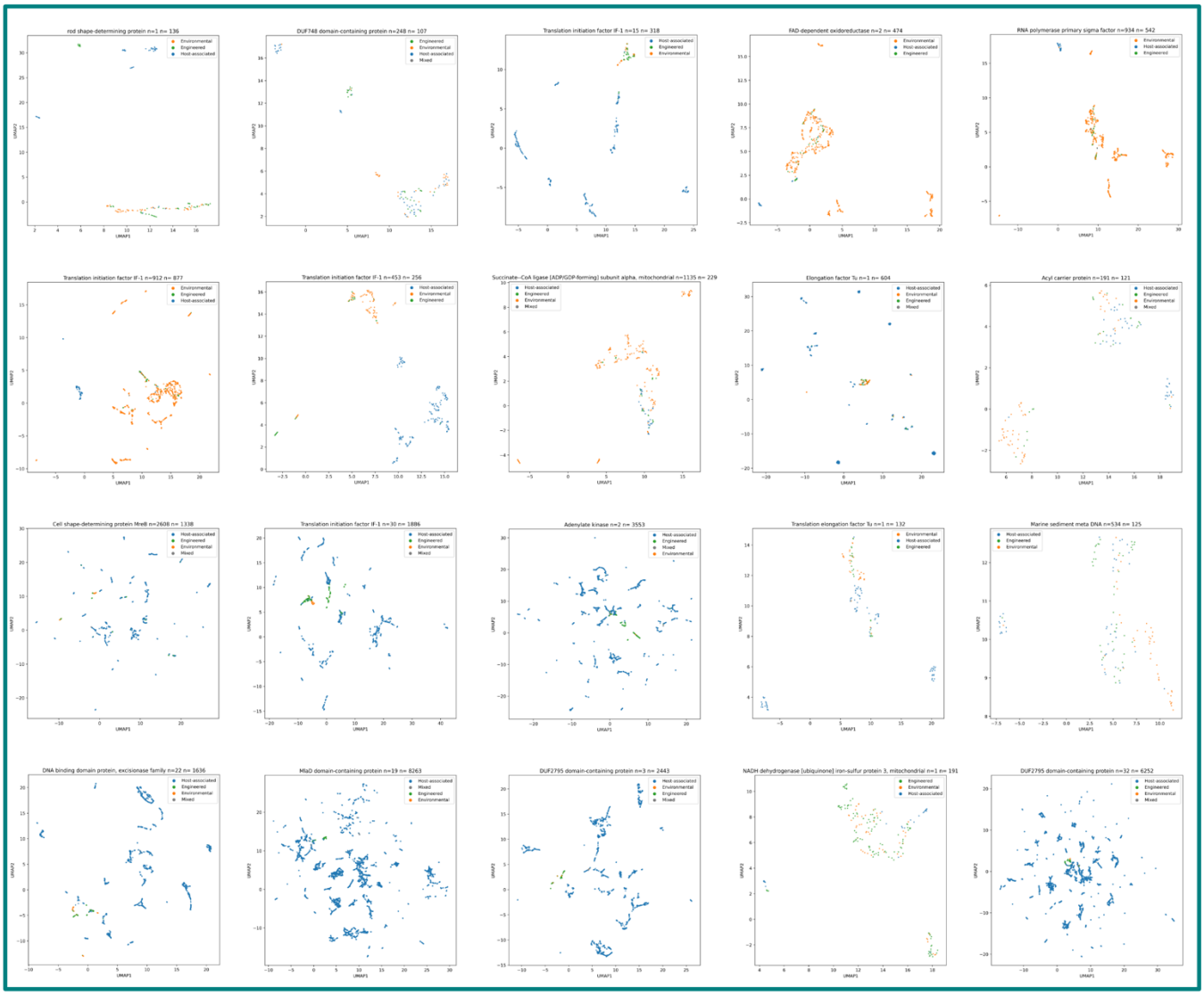
Extended data 4. Comparison of biLSTM baseline model with transformer-based gLM architecture and validation performances. Note that the biLSTM baseline is smaller in model size than gLM. While we attempted to scale this model by increasing the number of layers, the model failed to converge.

	biLSTM	**gLM**
Number of layers	5	19
Attention heads	N/A	10
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	467,253	1,296,960
Number of predictions	1	4
Number of parameters	27,811,840	954,736,916
% Pseudo-accuracy (validation)	27.9	71.9
% Absolute accuracy (validation)	14.78	59.2

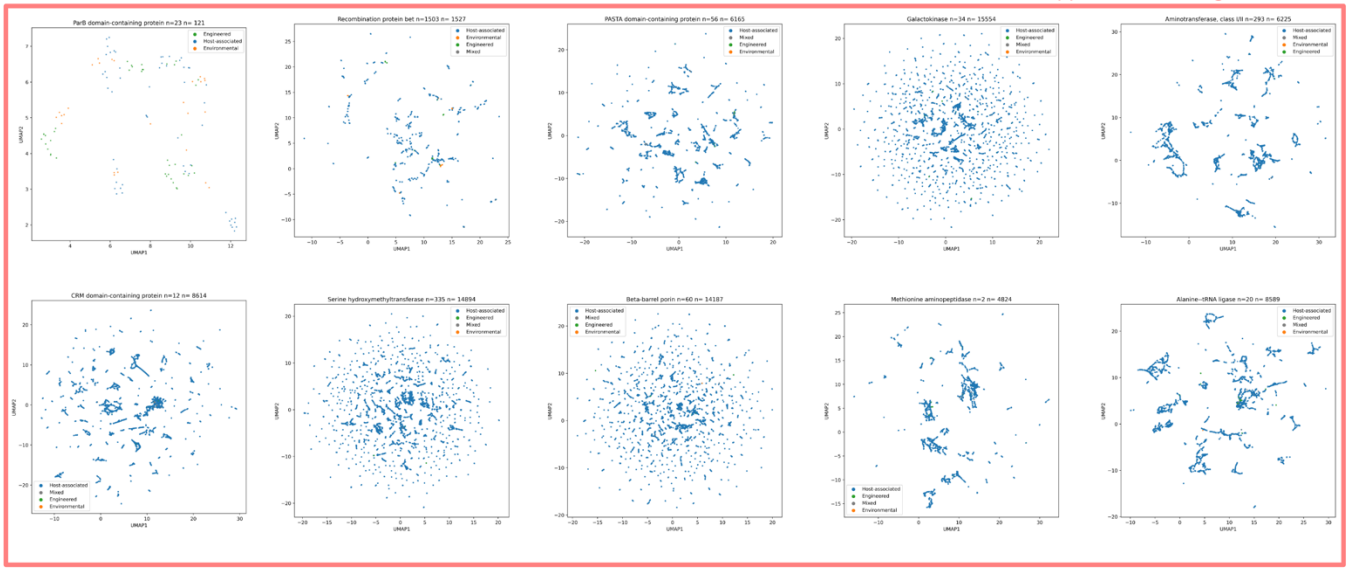
Extended Data 5. Ablation of pLM representations. Ablated gLM was trained on one-hot representations until convergence (<0.1% decrease in loss over 40k iterations).

	gLM	**gLM one-hot**
Representations	ESM2 embedding	One-hot amino acid encoding
Pooling	mean	mean
Number of layers	19	19
Attention heads	10	10
Input embedding dimension	1281	34
Hidden size	1280	1280
Batch size	3000	3000
Learning rate	1e-4	1e-4
Warm up steps	5000	5000
Training steps	1296960	132050
Number of predictions	4	4
Number of parameters	954736916	945338764
% Pseudo-accuracy (validation)	71.9	3.29
% Absolute accuracy (validation)	59.2	0.002
Operon prediction mAP	0.775 ± 0.29	0.426 ± 0.015

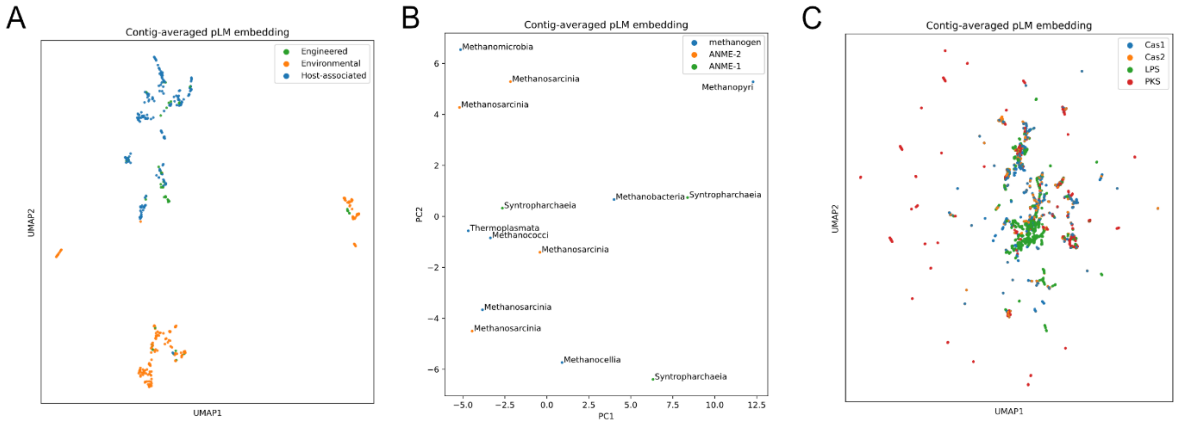
Clustering with biomes



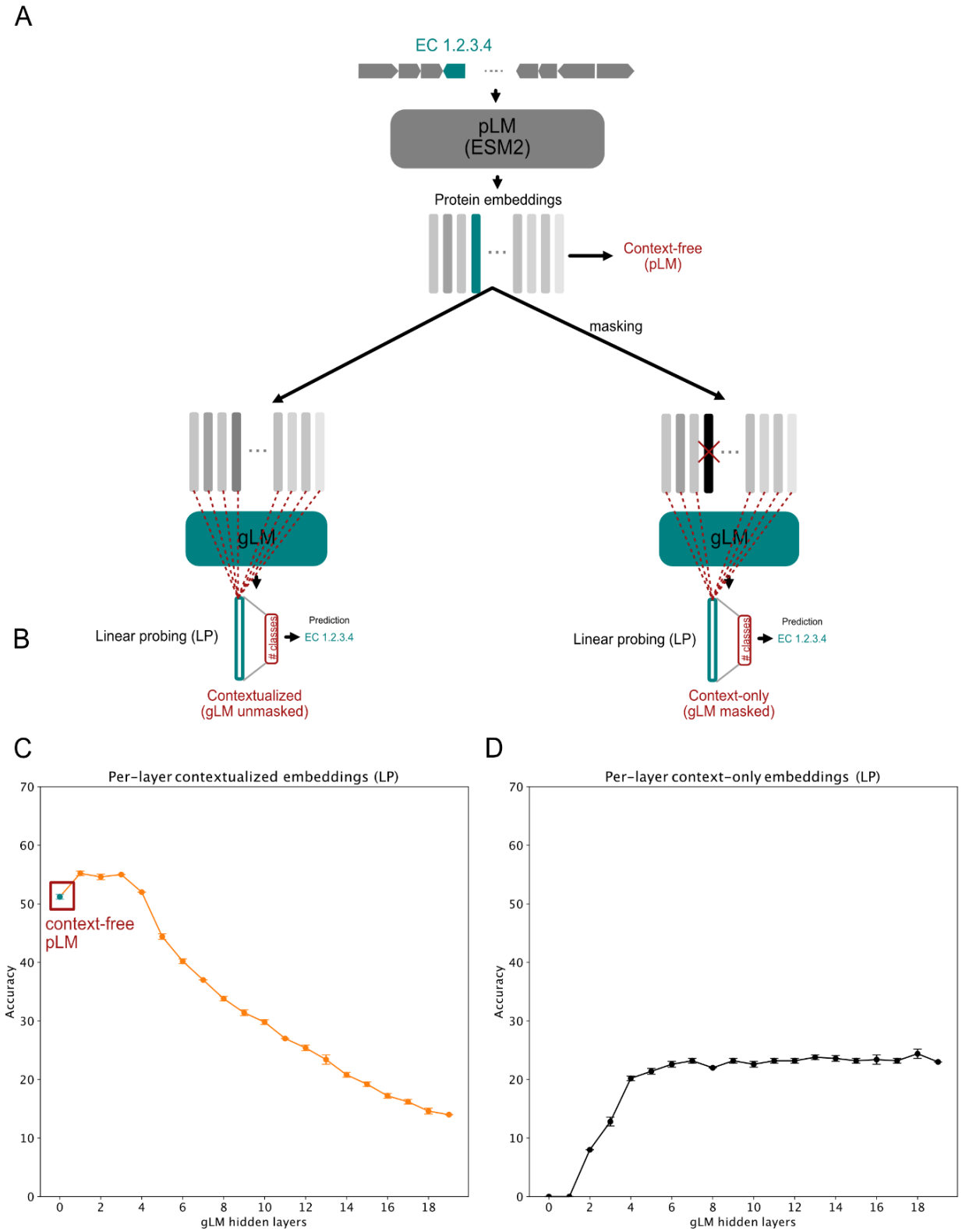
No apparent clustering with biomes



Extended Data 6. Contextualized protein embeddings for all 30 other multi-biome MGYPs. Biome color legend is identical to **Figure 2B**. Blue: Host-associated, Green: Engineered, Orange: Environmental, Grey: Mixed.



Extended Data 7. Contig-averaged pLM embeddings for figures 2B, F, and I respectively. Contig-averaged pLM embeddings were calculated by mean-pooling pLM embeddings across a given subcontig. Silhouette scores for clustering based on biomes (A), preferred direction of enzymatic reaction (B) and functional category (C) are 0.08, 0.067, and 0.05±0.17 (n = 10 independent samples) respectively.



Extended Data 8. Linear probing of context-free, context-only and contextualized gene embeddings.

A) Schematics of how context-free, context-only, and contextualized gene embeddings are extracted for linear probing. Context-only gene embeddings are extracted by masking the queried gene, therefore the original pLM signal cannot propagate through to gLM representation. This can be used to quantify how gLM-learned context informs EC number prediction independent of pLM signal. B) Schematics of how LP is used for EC number classification. C) Per-layer linear probing accuracy of gLM contextualized embeddings, where no-masking was performed at the time of inference. The 0th layer of gLM is equivalent to context-free pLM embedding. D) Per-layer linear probing accuracies of gLM context-only embeddings, where the queried genes are masked at the time of inference.

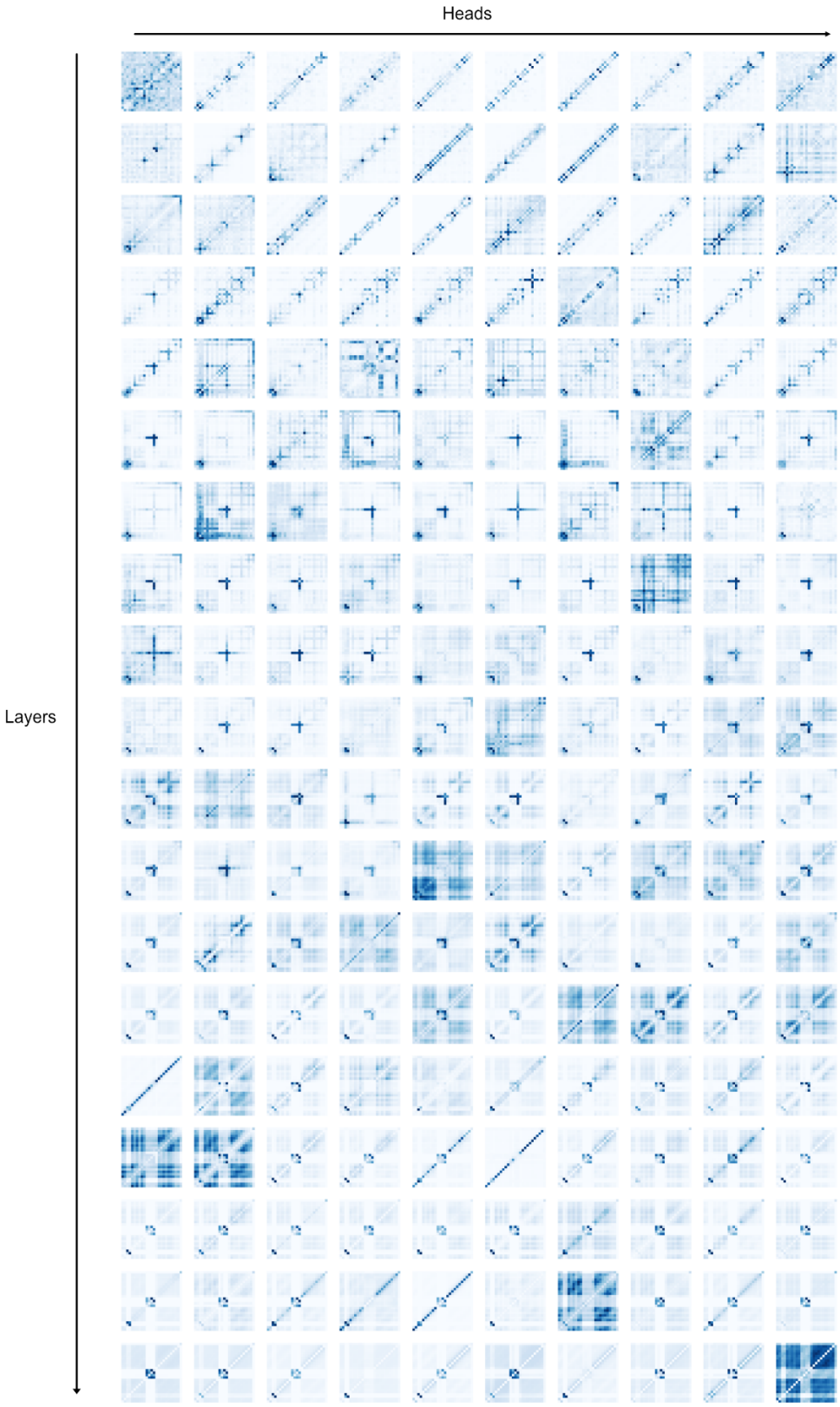
721 **Extended Data 9. Context-variant gene annotations.**
722
723 A. List of top ten most context variant genes where variance is calculated using gLM contextualized
724 embeddings.
725

MGYP ID	Mean embedding variance	UniRef annotation
815530568	18.87014775	Uncharacterized protein n=1
3385096772	17.70851124	Tail fiber protein n=20
2546832411	17.70269541	Capsular exopolysaccharide family n=1
767058056	17.08193362	Type I restriction enzyme endonuclease subunit
3381572716	16.44215407	Ferrous iron transport protein B n=11
2534762610	16.38703203	None
2733868861	16.21466745	30S ribosomal protein S18 n=4
551790786	16.06301055	30S ribosomal protein S18 n=285
3383952544	15.74879686	CRISPR-associated endoribonuclease Cas2 n=88
832772145	15.62986097	Ferrous iron transport protein A n=6

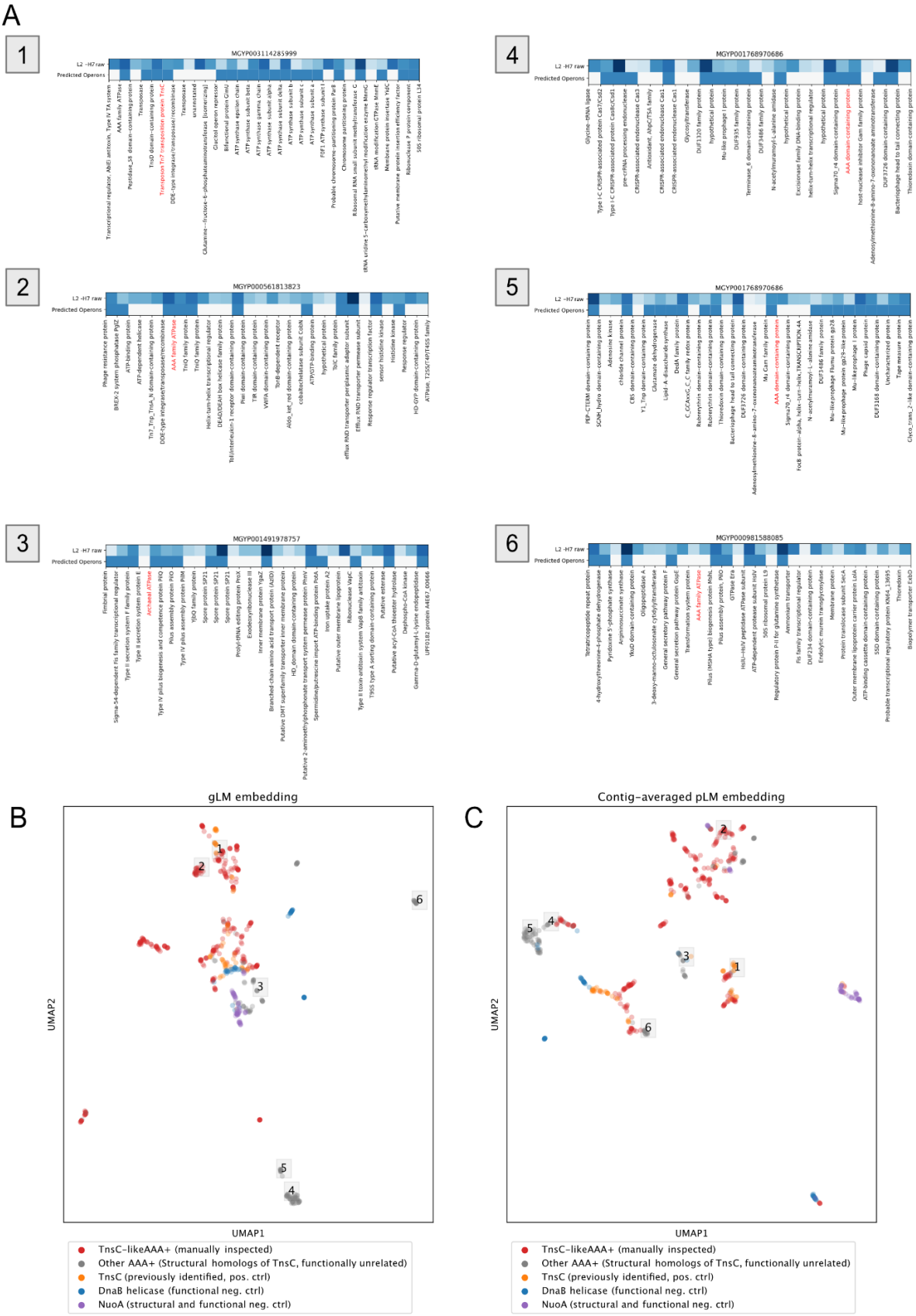
726
727 B. List of top most context variant genes where variance is calculated using contig-Averaged .
728

MGYP ID	Mean embedding variance	UniRef annotation
3384243084	18.43362238	DDE_Tnp_1 domain-containing protein n=137
943740416	17.77505976	Phosphoribosylamine--glycine ligase n=7
120256991	16.91703339	Asparaginase n=7
209907419	16.75686986	DNA mismatch repair protein MutS n=2
3380534673	16.67205026	IS66-like element ISBf10 family transposase
3384295730	16.44076785	IS66-like element ISBf10 family transposase
3381576767	16.23692972	VOC domain-containing protein n=96
3385319961	16.10454523	Peptidase E n=538
87924424	15.97918443	3-oxoacyl-[acyl-carrier-protein] synthase 3
997418675	15.91416004	HNH homing endonuclease n=2

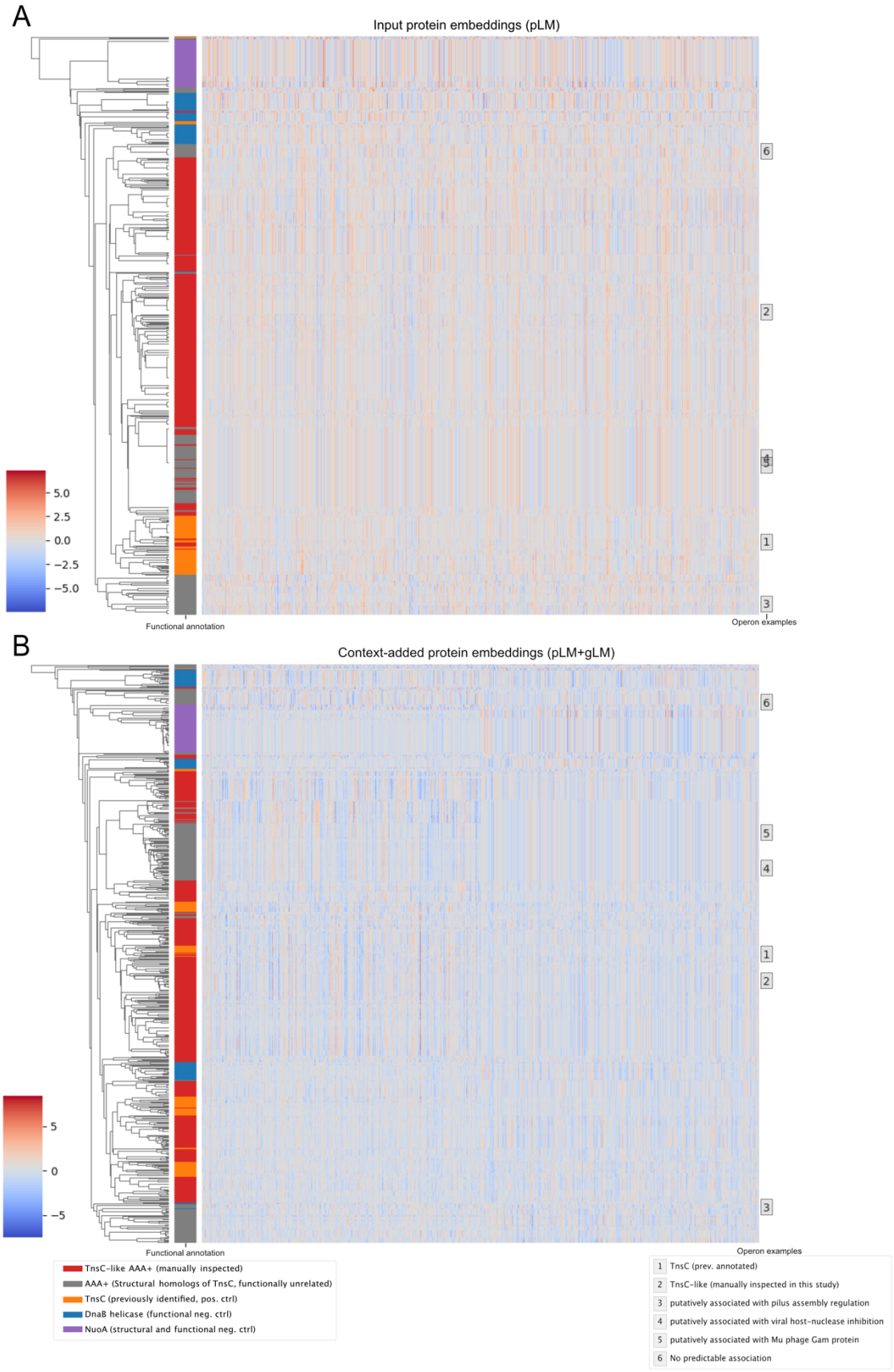
729
730



Extended data 10. Visualization of attention heads for a randomly chosen sequence. Increasing layer depth down the figure. Darker color represents larger attention scores.

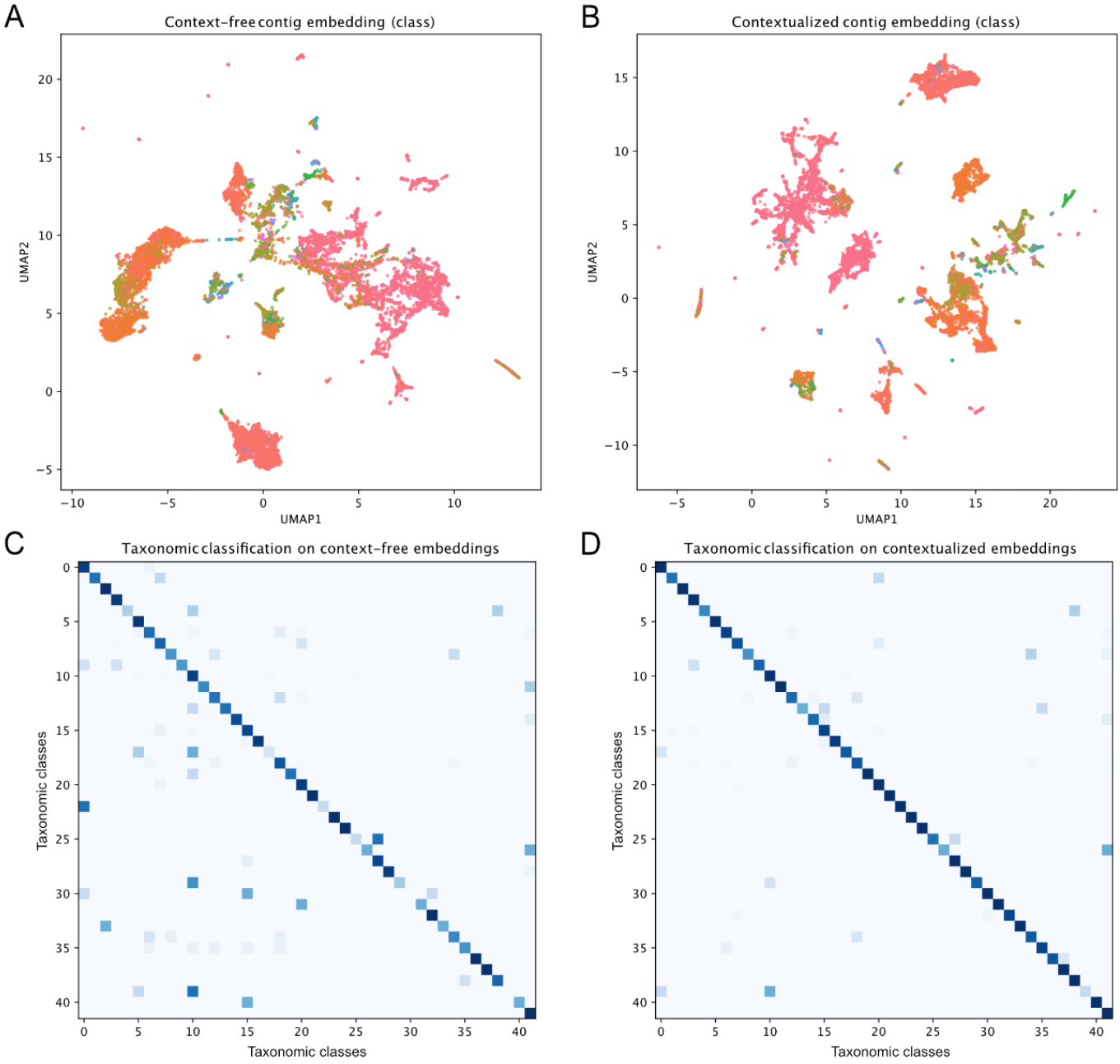

A
1
MGYP003114285999
L2 -H7 raw
Predicted Operons
Transcriptional regulator, AbiEi antitoxin, Type IV TA system
AAA family ATPase
Peptidase_S8 domain-containing protein
Transposase
TnsD domain-containing protein
Transposon Tn7 transposition protein TnsC
DDE-type integrase/transposase/recombinase
Transposase
uncharacterized
Glutamine--fructose-6-phosphate aminotransferase [isomerizing]
Cluster operon repressor
Bifunctional protein GmhD
ATP synthase epsilon chain
ATP synthase subunit beta
ATP synthase gamma chain
ATP synthase subunit alpha
ATP synthase subunit delta
ATP synthase subunit b
ATP synthase subunit c
ATP synthase subunit a
F0F1 ATP synthase subunit 9
Probable chemosome partitioning protein ParB
Chemosome partitioning protein
Ribosomal RNA small subunit methyltransferase G
tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG
RNA modification GTPase MnmE
Membrane protease subunit HflC
Putative membrane protease subunit HflC
Ribonuclease P protein component
50S ribosome protein L34
2
MGYP000561813823
L2 -H7 raw
Predicted Operons
Phage resistance protein
BNR/Asx 2 system phosphatase PgbZ
ATP-binding protein
ATP-dependent helicase
TnJ_Trp_TnsA_N domain-containing protein
DDE-type integrase/transposase/recombinase
AAA family ATPase
TnsQ family protein
TnsQ family protein
Helix-turn-helix transcriptional regulator
DEAD/DEAH box helicase family protein
Tollinterleukin-1 receptor domain-containing protein
Pleii domain-containing protein
TIR domain-containing protein
VWFA domain-containing protein
TonB-dependent receptor
Aldo_ket_red domain-containing protein
cobaltochelatase subunit CobN
ATPGTP-binding protein
hypothetical protein
TolC family protein
efflux RND transporter periplasmic adaptor subunit
Efflux RND transporter permease subunit
Response regulator transcription factor
sensor histidine kinase
Histidine kinase
Response regulator
HD-GYP domain-containing protein
ATPase, T2SS/T4P/T4SS family
3
MGYP001491978757
L2 -H7 raw
Predicted Operons
Fimbrial protein
Sigma-54-dependent Fis family transcriptional regulator
Type II secretion system F family protein
Type II secretion system protein E
Archaeal ATPase
Type IV pilus biogenesis and competence protein PilQ
Pilus assembly protein PilO
Pilus assembly protein PilM
Type IV pilus assembly protein
YjbQ family protein
Spore protein SP21
Spore protein SP21
Spore protein SP21
Prolyl-tRNA editing protein ProX
Exodeoxyribonuclease III
Inner membrane protein YgaZ
Branched-chain amino acid transport protein (AAT)
Putative DMT superfamily transporter inner membrane protein
HD_domain domain-containing protein
Putative 2-aminoethylphosphonate transport system permease protein PhnV
Spermidine/putrescine import ATP-binding protein PotA
Iron uptake protein A2
Putative outer membrane lipoprotein
Ribonuclease VapC
Type II toxin-antitoxin system VapB family antitoxin
TSSS type A sorting domain-containing protein
Putative esterase
Putative acyl-CoA thioester hydrolase
Dephospho-CoA kinase
Gamma-D-glutamyl-L-lysine endopeptidase
UPF0182 protein A4E87_00466
4
MGYP001768970686
L2 -H7 raw
Predicted Operons
Glycine--tRNA ligase
Type I-C CRISPR-associated protein Cas7/Csd2
Type I-C CRISPR-associated protein Cas8c/Csd1
pre-crRNA processing endonuclease
CRISPR-associated endonuclease Cas3
Antioxidant, AbpC/TSA family
CRISPR-associated endonuclease Cas1
CRISPR-associated endonuclease Cas1
Glycosyltransferase
DUF1320 family protein
hypothetical protein
Mu-like prophage I protein
DUF935 family protein
Terminase_6 domain-containing protein
DUF3486 family protein
N-acetylmuramoyl-L-alanine amidase
hypothetical protein
Excisionase family DNA-binding protein
helix-turn-helix transcriptional regulator
hypothetical protein
Sigma70_r4 domain-containing protein
AAA domain-containing protein
host-nuclease inhibitor Gam family protein
Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
DUF3726 domain-containing protein
Bacteriophage head to tail connecting protein
Thioredoxin domain-containing protein
5
MGYP001768970686
L2 -H7 raw
Predicted Operons
PEP-CTERM domain-containing protein
SGNH_hydro domain-containing protein
Adenosine kinase
chloride channel protein
CBS domain-containing protein
YL_Trp domain-containing protein
Glutamate dehydrogenase
Lipid A disaccharide synthase
DsdA family protein
C_GCAxxG_C_C family redox protein
Rubrerythrin domain-containing protein
Rubrerythrin domain-containing protein
Thioredoxin domain-containing protein
Bacteriophage head to tail connecting protein
DUF3726 domain-containing protein
Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
MA_Cam family protein
AAA domain-containing protein
Sigma70_r4 domain-containing protein
FocB protein-alpha, helix-turn-helix, TRANSCRIPTION-4A
N-acetylmuramoyl-L-alanine amidase
DUF3486 family protein
Mu-like prophage I Pilmu protein gp18
Mu-like prophage protein gp15-like protein
Mu-like prophage I protein
Phage capsid protein
DUF3168 domain-containing protein
Uncharacterized protein
Tape measure protein
Chyo_trans_2-like domain-containing protein
6
MGYP000981588085
L2 -H7 raw
Predicted Operons
Tetratricopeptide repeat protein
4-hydroxythreonine-4-phosphate dehydrogenase
Pyridoxine 5-phosphate synthase
Arginosuccinate synthase
YkuD domain-containing protein
Oligopeptidase A
3-deoxy-manno-octulosonate cytidylyltransferase
General secretory pathway protein F
General secretion pathway protein GspE
Transformation system protein
AAA family ATPase
Pilus (MSHA type) biogenesis protein MshL
Pilus assembly protein, PilO
GTPase Era
HslU--HslV peptidase ATPase subunit
ATP-dependent protease subunit HslV
50S ribosomal protein L9
Regulatory protein PsiE for glutamine synthetase
Ammonium transporter
Fis family transcriptional regulator
DUF734 domain-containing protein
Endolytic murein transglycosylase
Membrane protein
Protein translocase subunit SecA
Outer membrane lipopeptide carrier protein LolA
ATP-binding cassette domain-containing protein
SSD domain-containing protein
Probable transcriptional regulatory protein KN4_13955
Thioredoxin
Biopolymer transporter ExbD
B
gLM embedding
UMAP2
UMAP1
1
2
3
4
5
6
TnsC-likeAAA+ (manually inspected)
Other AAA+ (Structural homologs of TnsC, functionally unrelated)
TnsC (previously identified, pos. ctrl)
DnaB helicase (functional neg. ctrl)
NuoA (structural and functional neg. ctrl)
C
Contig-averaged pLM embedding
UMAP2
UMAP1
1
2
3
4
5
6
TnsC-likeAAA+ (manually inspected)
Other AAA+ (Structural homologs of TnsC, functionally unrelated)
TnsC (previously identified, pos. ctrl)
DnaB helicase (functional neg. ctrl)
NuoA (structural and functional neg. ctrl)

735 **Extended Data 11. Attention-based operon prediction of AAA+ regulators. A)** Annotations were
736 made by searching against the UniRef90 database. Protein annotation in Red indicates the location of the
737 AAA+ regulator identified based on a structural search of TsnC. Raw attention in the most correlated
738 head (2nd layer, 7nd head, "L2-H7") is shown in the first row of the heatmap, and the second row is the
739 predicted operon using the logistic regression predictor trained on *E.coli* K-12 operon dataset. **B)** gLM
740 representations (last hidden layer of gLM) for **Figure 4EF** (KL divergence ratio = 0.53). **C)** Contig-
741 averaged pLM baseline for **Figure 4EF**, where pLM embeddings across the subcontig were averaged to
742 represent the queried protein (KL divergence ratio = 0.57)

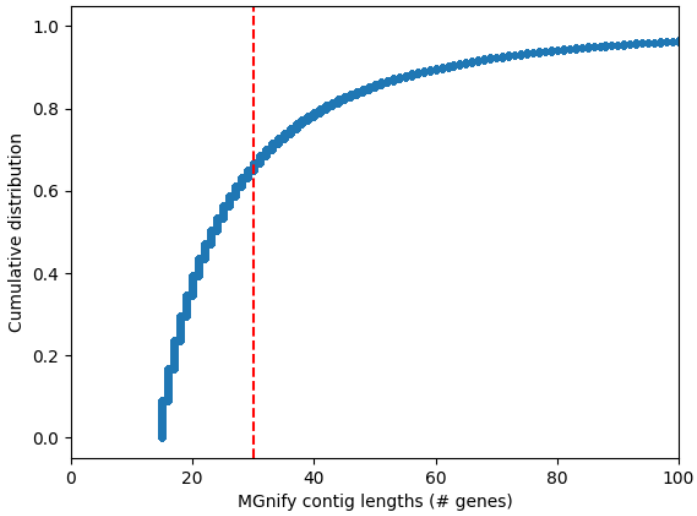

A
Input protein embeddings (pLM)
5.0
2.5
0.0
−2.5
−5.0
Functional annotation
Operon examples
6
2
4
5
1
3
B
Context-added protein embeddings (pLM+gLM)
5
0
−5
Functional annotation
Operon examples
6
5
4
1
2
3
TnsC-like AAA+ (manually inspected)
AAA+ (Structural homologs of TnsC, functionally unrelated)
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

743

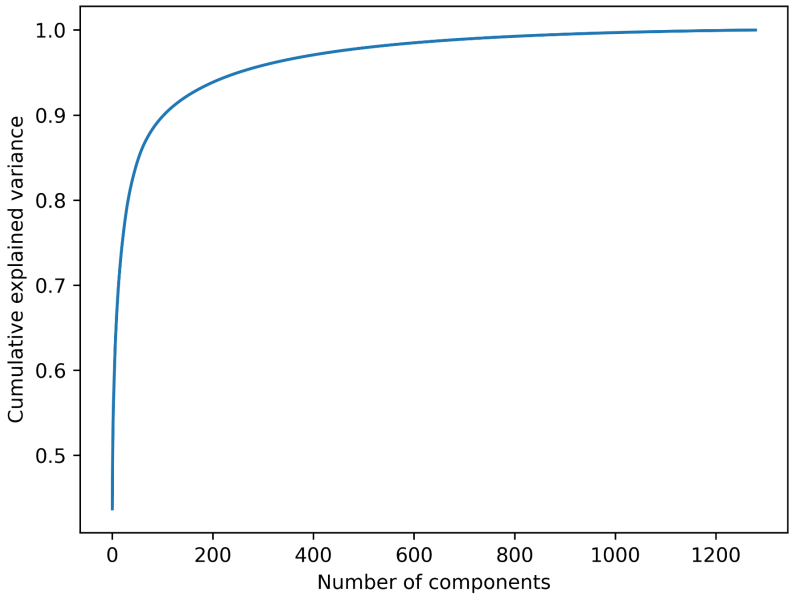
744 **Extended Data 12. Distance based clustering of raw embeddings of AAA+ regulators and controls**
745 **shown in Figure 4EF.** Row colors and labels correspond to the point colors and designation used in
746 **Figure 4EF** and **Extended Data 12**.



Extended Data 13. 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 5G**. Confusion matrices for context-free (C) and contextualized (D) logistic regression classifier as shown in **Figure 5F**.



Extended Data 14. Cumulative distribution of raw MGnify contig lengths (count of encoded genes), MGnify contigs < 15 genes were excluded from the corpus in this study. Red line is drawn at the contig length of 30 genes.



Extended Data 15. Cumulative explained variance of principal components of ESM2 embeddings calculated on 2.5 million randomly selected MGYPs.

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. Outeiral, C. & Deane, C. Codon language embeddings provide strong signals for protein engineering. *BioRxiv* (12/2022) doi:10.1101/2022.12.15.519894.
8. Wright, S. On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* **2**, 279–294 (1948).
9. Lynch, M. & Conery, J. S. The Origins of Genome Complexity. *Science* **302**, 1401–1404 (2003).
10. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).
11. Treangen, T. J. & Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
12. Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48–51 (2012).
13. Kountz, D. J. & Balskus, E. P. Leveraging Microbial Genomes and Genomic Context for Chemical

Discovery. *Acc. Chem. Res.* **54**, 2788–2797 (2021).

14. Shmakov, S. A. *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat. Protoc.* **14**, 3013–3031 (2019).
15. Yelton, A. P. *et al.* A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes. *PLoS Comput. Biol.* **7**, e1002230 (2011).
16. Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language processing. *Nat. Commun.* **13**, 5731 (2022).
17. Konno, N. & Iwasaki, W. Machine learning enables prediction of metabolic system evolution in bacteria. *Sci Adv* **9**, eadc9130 (2023).
18. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
19. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
20. Dalla-Torre, H. *et al.* The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv* 2023.01.11.523679 (2023) doi:10.1101/2023.01.11.523679.
21. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. (2023).
22. Zvyagin, M. *et al.* GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *bioRxiv* (2022) doi:10.1101/2022.10.10.511571.
23. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
24. Vaswani, Shazeer & Parmar. Attention is all you need. *Adv. Neural Inf. Process. Syst.*
25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).

26. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
27. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
28. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).
29. Jeffery, C. J. Protein moonlighting: what is it, and why is it important? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, (2018).
30. Miskei, M. *et al.* Fuzziness enables context dependence of protein interactions. *FEBS Lett.* **591**, 2682–2695 (2017).
31. 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).
32. Ben-Hur, A. & Brutlag, D. Remote homology detection: a motif based approach. *Bioinformatics* **19 Suppl 1**, i26–33 (2003).
33. 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).
34. 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 Preprint at https://doi.org/10.1029/2006jg000268 (2007).
35. Shao, N. *et al.* Expression of divergent methyl/alkyl coenzyme M reductases from uncultured archaea. *Commun Biol* **5**, 1113 (2022).
36. Reif, E. *et al.* Visualizing and measuring the geometry of BERT. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
37. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, (2022).
38. Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–

937 (2022).

39. Rogers, A., Kovaleva, O. & Rumshisky, A. A primer in BERTology: What we know about how BERT works. *Trans. Assoc. Comput. Linguist.* doi:10.1162/tacl_a_00349/96482.

40. Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models. *arXiv [cs.CL]* (2020).

41. Salgado, H. *et al.* Using RegulonDB, the Escherichia coli K-12 Gene Regulatory Transcriptional Network Database. *Curr. Protoc. Bioinformatics* **61**, 1.32.1–1.32.30 (2018).

42. White, S. R. & Lauring, B. AAA+ ATPases: achieving diversity of function with conserved machinery. *Traffic* **8**, 1657–1667 (2007).

43. Park, J.-U. *et al.* Structures of the holo CRISPR RNA-guided transposon integration complex. *Nature* **613**, 775–782 (2023).

44. Rybarski, J. R., Hu, K., Hill, A. M., Wilke, C. O. & Finkelstein, I. J. Metagenomic discovery of CRISPR-associated transposons. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

45. Benler, S. *et al.* Cargo Genes of Tn7-Like Transposons Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic Resistance Genes. *MBio* **12**, e0293821 (2021).

46. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).

47. van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *bioRxiv* 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.

48. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).

49. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Pairing interacting protein sequences using masked language modeling. *arXiv* (2023) doi:2308.07136.

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

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

52. 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).
53. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
54. 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.
55. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv [cs.CL]* (2019).
56. Huang, Z., Liang, D., Xu, P. & Xiang, B. Improve Transformer Models with Better Relative Position Embeddings. *arXiv [cs.CL]* (2020).
57. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv [cs.LG]* (2017).
58. Tierrafría, V. H. *et al.* RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb Genom* **8**, (2022).
59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
60. 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).
61. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
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).
63. Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. & Vitale, L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database* **2016**, (2016).