Deciphering enzymatic potential in metagenomic reads through DNA language model

R Prabakaran[1,2,*], Yana Bromberg[1,2,*]

[1]Department of Biology, Emory University, Atlanta, GA 30322, USA
[2]Department of Computer Science, Emory University, Atlanta, GA 30322, USA
prabakaran@emory.edu, yana.bromberg@emory.edu

Abstract

The microbial world plays a fundamental role in shaping Earth's biosphere, steering global processes such as carbon and nitrogen cycling, soil rejuvenation, and ecological fortification. An overwhelming majority of microbial entities, however, remain unstudied. Metagenomics stands to elucidate this microbial "dark matter" by directly sequencing the microbial community DNA from environmental samples. Yet, our ability to explore these metagenomic sequences is limited to establishing their similarity to curated datasets of organisms or genes/proteins. Aside from the difficulties in establishing such similarity, the reference-based approaches, by definition, forgo discovery of any entities sufficiently unlike the reference collection.

Presenting a paradigm shift, language model-based methods, offer promising avenues for reference-free analysis of metagenomic reads. Here, we introduce two language models, a pretrained foundation model REMME, aimed at understanding the DNA context of metagenomic reads, and the fine-tuned REBEAN model for predicting the enzymatic potential encoded within the read-corresponding genes. By emphasizing function over gene identification, REBEAN is able to label known functions carried both by previously explored genes and by new (orphan) sequences. Furthermore, even though it is not explicitly trained to do so, REBEAN identifies the functionally relevant parts of a gene. Our comprehensive analysis highlights our models' potential for metagenomic read annotation and unearthing of novel enzymes, thus enriching our understanding of microbial communities.

Introduction

The world we live in is estimated to host a million to a trillion bacterial species spread wide and deep across the globe (Locey and Lennon 2016, Louca, Mazel et al. 2019). These minuscule creatures are the foundation of Earth's biosphere and are responsible for much of life as we know it. We learned of the existence of microbes with the discovery of the microscope in the 1600's (Lane 2015). Centuries later, however, we are only capable of isolating, growing, and studying a tiny fraction of these organisms in the lab (Staley and Konopka 1985, Solden, Lloyd et al. 2016, Steen, Crits-Christoph et al. 2019, Ruscheweyh, Milanese et al. 2022).

An alternate to the exploration of individual microbes is studying them directly within their communities via analysis of metagenomes, i.e. the genetic material of complete environmental samples. Current applications of metagenomic studies are diverse, including but not limited to surveillance of antimicrobial resistance (AMR), exploring microbial adaptation to environmental changes, studying the influence of host microbiome on host health, identification of new microbial species, and discovery of novel genes and gene products (Handelsman 2004, Zhu, Miller et al. 2018, Ko, Chng et al. 2022). Significant advances in sequencing techniques in the past two decades have largely overcome the technical challenges associated with metagenomic data extraction, shifting the focus towards downstream analyses, such as gene prediction, taxonomic classification, and functional annotation (Bharti and Grimm 2021).

Metagenomic (and metatranscriptomic) functional profiling is an important step in deciphering the contents of the microbial sample (Pushkarev, Inoue et al. 2018, Zhu, Miller et al. 2020, Ko, Chng et al. 2022). It provides a way to track molecular functions known to be important and to identify novel genes of known or yet-undescribed functionality. Functional annotation of a metagenomic sample may (1) involve annotation of genes identified from assembled stretches of sequence (transcripts, contigs, or genomes) or (2) be inferred directly from sequencing reads (Bharti and Grimm 2021).

Functional labels are predefined terms describing molecular functions and/or associated biological processes and pathways or cellular locations. They are part of extant ontologies and domain/family collections, e.g. Enzyme Commission numbers (EC), Gene Ontology (GO) terms, Pfam protein families and InterPro signatures, Clusters of Orthologous Genes (COG), and KEGG genes and pathways (International Union of and Webb 1992, Tatusov, Koonin et al. 1997, Ashburner, Ball et al. 2000, Mistry, Chuguransky et al. 2021, Kanehisa, Sato et al. 2022, Paysan-Lafosse, Blum et al. 2023).

Most current metagenome annotation methods rely on gene or read mapping to related reference sequences of annotated genes or proteins. Reference-based function assignments use sequence alignment, k-mer indexing and matching (Tatusov, Galperin et al. 2000, Overbeek, Begley et al. 2005, Aziz, Bartels et al. 2008, Mistry, Chuguransky et al. 2021, Sayers, Beck et al. 2024) Hidden Markov Models (HMMs) of families, or structural comparisons. Requiring an existing reference, however, limits the range of discovery possible with these methods (Prabakaran and Bromberg 2023). Recently, deep learning models including language models (LM) have shown promising results in predicting protein structure and making domain and function annotations (Jumper, Evans et al. 2021, Bileschi, Belanger et al. 2022, Lin, Akin et al. 2023, Sanderson, Bileschi et al. 2023). Deep learning models have been experimented with in the metagenomic world as well (Ryu, Kim et al. 2019, Hoarfrost, Aptekmann et al. 2022, Pan, Zhu et al. 2022). We thus suggest that a machine learning approaches can be developed to generalize the information encoded in sequence for the purposes of predicting function reference-free.

In this study, we describe a DNA language model (dLM) for reference- and assembly- independent annotation of enzymatic activities encoded by microbial genes that give rise to metagenomic sequencing reads. Our pretrained dLM REMME (Read Embedder for Metagenomic exploration) learns the "language" of sequencing reads and is adaptable to various downstream tasks. One example is REBEAN (Read Embedding Based Enzyme Annotator), a functional classifier demonstrating robust predictive performance, by leveraging the understanding on the context of reads within their "parent" enzymes. Our analyses describe REBEAN's extensive applicability to metagenome annotations, particularly highlighting its ability to forgo sequence-defined homology in favor of discovering novel enzymes.

We note that much of the recent emphasis in the protein and nucleotide work has been on creating novel sequences that perform desired functions (Madani, Krause et al. 2023, Nguyen, Poli et al. 2024). This approach, however, discards the well-optimized natural results of billions of years of evolution in favor of synthetic, often questionable, biology. Here we suggest that annotation of the existing rich diversity of sequences, rather than de novo synthesis, should be the first step in harnessing the power of biological molecules.

RESULTS AND DISCUSSION

Building a DNA-language model (dLM) to decipher metagenomic reads. We trained a transformer-based language model to annotate the enzymatic activity of genes giving rise to sequenced reads. That is, given a particular read, we asked: "What would be the function of the gene that it came from?"

Our model was pretrained as a generalized read embedder to understand the DNA language using 53.6 million reads from 1,496 prokaryotic genomes (**Methods**). It was then fine-tuned on 19 million non-redundant reads from 19,316 metagenomic samples to annotate enzymatic function in terms of the seven first level EC classes. For clarity of the discussion, we named the pretrained dLM REMME (Read Embedder for Metagenomic exploration) and the fine-tuned model REBEAN (Read Embedding Based Enzyme Annotator).

REMME was trained to embed DNA reads through mask-token prediction, i.e. predicting nucleotides that are masked in each read. It attained over 98.5% accuracy (**Eq. 1**) in each: training, validation, and testing (**Figure S1**). This performance highlights its solid understanding of the DNA "spelling", i.e. the nucleotide order of occurrence. To enforce further understanding of biological context, we gave REMME two additional learning tasks: (1) estimate the fraction of protein-coding nucleotides in each read and (2) if the read is part of a coding DNA sequence (CDS), predict the reading frame; as part of the latter task, REMME was trained to predict a given read class: 1, 2, or 3 for each of the

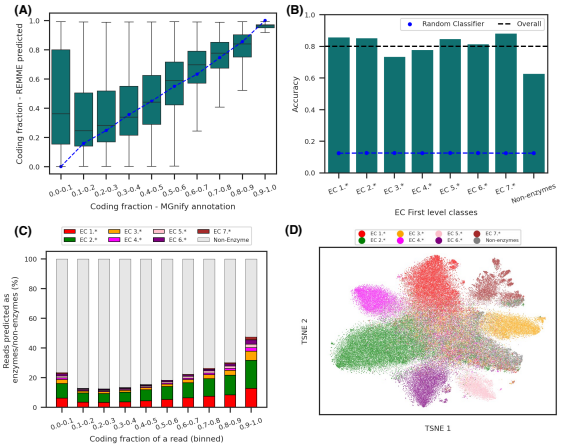

Figure 1: Training language models for metagenomic reads. (A) Distribution of REMME-predicted protein-coding nucleotide fraction across reads in REMME's test set (Y-axis) is similar to that of MGnify annotations (X-axis). (B) REBEAN's test set prediction accuracy (**Eqn. 1**) highlights non-random performance across all enzyme classes (bars higher than the blue line), although some classes are deemed easier to label than others. (C) Distribution of reads in REMME's test set across different fractions of protein-coding residues (X-Axis) vs. the number of these reads assigned to REBEAN-predicted EC classes (colors). Note that there is a trend from left to right, relating the coding fraction to the number of reads identified as coming from enzymes. For example, there are many more enzyme-like reads (~47.4%) among those with a higher fraction of coding residues (0.9-1) than among those within the 0.1-0.2 coding fraction range (12.9%). (D) t-SNE projection of REBEAN embeddings of REMME's test set, annotated with mifaser-predicted EC classes (colors), illustrates REBEAN's ability to differentiate read classes.

three-reading frames (with respect to the first position of the read in the 5' to 3' direction) and 4 if the read was not transcribed.

REMME learned to identify coding regions within reads well. In both validation and testing, the REMME-predicted fraction of coding nucleotides per read was correlated with the "ground truth" fraction of coding nucleotides (**Figure 1A;** Pearson's r = 0.73, *p-value* < 1E-32; MSE=0.04 and MAE=0.11, **Eqn. 7-8**). Furthermore, REMME predicted reading frames with an average accuracy of 48.6% and AUC of 0.789, i.e. not perfectly but better than a random classifier (25% accuracy). Note that the annotation of CDSs and reading frames in this study was produced by computational annotation pipelines (Richardson, Allen et al. 2023) and, as such, could be incomplete or erroneous. Also note that the reading frame is not well defined for reads, i.e. sequence fragments, possibly shared by multiple CDSs with different reading frames, on either forward or reverse strands. Despite these ambiguities, REMME distinguished reads which are part of CDSs (classes 1, 2 & 3) from non-transcribed reads with an accuracy of 88.5% (recall=94.2%, precision=92.7%; **Eqn. 2-3**).

To summarize, REMME was pretrained by enforcing relevant objectives to understand the biological context of reads and did well in addressing the tasks it was assigned.

Can we identify enzyme-coding fragments in metagenomic data? To answer this question, we fine-tuned REMME to predict the EC classes (level 1) of the proteins encoded by genes that gave rise to each of the reads. This task is more challenging than predicting deeper EC levels, as deeper levels provide a more specific characterization of protein catalytic activity and typically correspond to only a handful of enzymes." For example, EC number 1.2.3.4 represents Oxalate oxidase activity. Here, level 1 (1.*) of the number represents the type of catalysis – an oxidoreductase reaction. Level 2 (1.2.*) denotes the fact that the substrate contains an aldehyde or oxo group and level 3 (1.2.3.*) indicates the involvement of oxygen as accepter. Finally, level 4 (1.2.3.4) describes the exact enzyme/catalyzed reaction, i.e. oxalate oxidase.

Read functional annotations were generated by mi-faser (**See Methods**), an alignment-based method that attained ~90% precision and ~50% recall in testing (Zhu, Miller et al. 2018). The decision to rely on mi-faser annotations as the ground truth was made to expose the model to a broader, real metagenome sequence space, instead of what can be gleaned from simulating reads from the commonly used annotated reference databases. Note that this decision allowed us to train on all genome regions (coding, non-coding, annotated or not) with no explicit biases.

The resulting REBEAN model is a multi-class classifier that predicts a score for each read as belonging to one of the seven level 1 EC classes (1: Oxidoreductases, 2: Transferases, 3: Hydrolases, 4: Lyases, 5: Isomerases, 6: Ligases & 7: Translocases) or a non-enzyme class. To accomplish this task, REBEAN had to learn broad but unique signals corresponding to each enzyme class despite the diversity of sequences within the class. We suggest that managing this broad scope is far harder than the much more focused task of the complete EC annotation (at all four levels). REBEAN, however, attained consistently high accuracy in classifying non-redundant reads (<80% sequence identity, (Hauser, Steinegger et al. 2016)) across training, validation and test sets (81.7%, 81.0% and 80.6%, respectively; **Table S1** & **Figure 1C**). In testing, REBEAN attained an average multiclass AUC of 0.969 and recall of 80.1% with a precision of 83.6%. REBEAN's test performance is far better than any random classifier which would score 12.5% accuracy. Does pretraining helps finetuning for EC prediction? To answer this question, we trained the REBEAN's base model initiated with randomly assigned weights instead of REMME weights, using same training protocol as REBEAN. Interestingly, the base model reached a maximum accuracy of only. 30% at the end of 200 epochs. We also fine-tuned DNABert pretrained model on REBEAN's training dataset (Ji, Zhou et al. 2021). At the end of 200,000 steps, the model reached a peak accuracy of 63.8 accuracy.

Despite the overall strong performance, REBEAN's ability to identify non-enzymatic reads was relatively low (recall=50.6%; **Figure 1D,** where a portion of non-enzymes exists in the same space as enzymes). That is, REBEAN labelled as enzymatic half (176,254) of the test set reads,

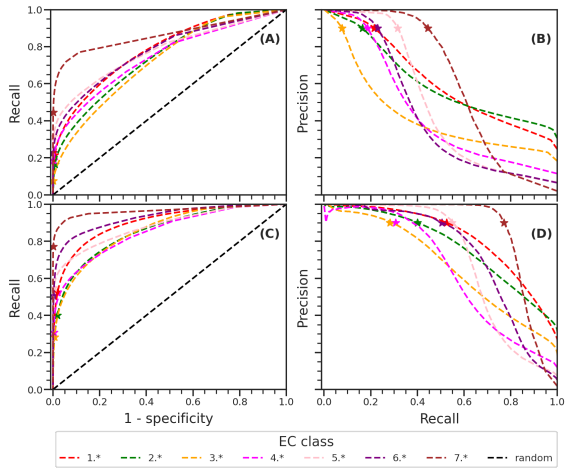

Figure 2: REBEAN synthetic read annotation. We extracted read-sized gene fragments from 4,295 prokaryotic enzymes and labelled them using REBEAN. A) ROC and B) PR curves demonstrate the performance of REBEAN in annotating reads with level 1 EC classes. C) ROC and D) PR curves represent performance of REBEAN in annotating complete proteins as an as average score of multiple reads with level 1 EC classes. The performance at chosen threshold corresponding to 90% precision is marked with '*' on the curve.

which mi-faser had not identified as such. We note, however, that mi-faser is a very precise (90% precision at 4th EC level) method that itself has low recall (50%). We thus further examined these 176K reads for possibly missed enzymatic character by aligning them against SwissProt (Hauser, Steinegger et al. 2016, UniProt 2023). Of these, only a third (50,756 reads) aligned to SwissProt proteins with >30% sequence identity, suggesting that the rest could represent yet-unseen genes, which may have enzymatic activity or not. Notably, three quarters (38,241 reads) of these 50K reads mapped to enzymes. These results demonstrate that REBEAN could identify enzymatic reads that were not captured through alignment-based mi-faser.

Table 1: List of sequence datasets used in analysis

Dataset	Type	Number of entries	Number of reads	Synthetic reads	Read sampling rate (reads/kbp)
SwissProt Exp. Enzymes I	Gene	4,295	525,775	Yes	100
SwissProt Exp. Enzymes II	Gene	4,295	54,775	Yes	10
SwissProt	Gene	473,854	5,305,894	Yes	10
Ortholog pairs	Gene	24,938	707,947	Yes	10
Synthetic Metagenome	Genomes	3,820	124,269,382	Yes	10
Extremophile metagenomes	Metagenome	27	8,030,102	No	-

REBEAN is robust for labeling enzymatic functions. To evaluate REBEAN performance against experimental annotations, we generated a set of 525,775 fragments (length 200bp) from 4,295 enzyme-coding genes with experimental evidence of enzymatic activity (**Methods, Table 1**). Each fragment, i.e. synthetic read, was labelled with the 1st level enzyme class of the parent gene. We evaluated REBEAN's performance in annotating these reads' enzyme classes (**Figure 2A, B**). At a threshold of 0.5, REBEAN attained an average recall=33.4%, precision=71.5%, and accuracy =88.1%; it attained an average Area Under the Receiver Operating Characteristics curve (AUROC) of 0.79, ranging from 0.73-0.90 across the seven enzyme classes (**Figure 2B** and **Table S1**).

Note that at 90% precision, REBEAN recovered, on average, only 23% (7.6% to 44.5%) of the reads from each enzyme class correctly. That is, REBEAN attributed high prediction confidence to only a fraction of reads. Since each of the reads in this set is derived from an enzyme, REBEAN's inability to assign high confidence to most reads suggests that it only learned to associate a specific fraction of each gene with its molecular activity.

Three possible, non-exclusive, protein biology-driven explanations for this observation come to mind. (1) One is based on the fact that specific functional activities, e.g. catalysis and ligand binding, tend to occupy only a small portion of the protein, with the rest of the sequence providing structural support. This may stem from the fact that enzyme evolution is incremental in nature, with functional specificity tuned within a small subsection of the protein and some things restricted to just a single residue (Babbitt, Hasson et al. 1996, Baier, Copp et al. 2016, Allen and Whitman 2021). In contrast, similarity between independently evolved enzymes of same function proteins is also often limited to a subset of the residues, i.e. active-site convergence (Davidi, Longo et al. 2018). These functionally active protein sections are likely ubiquitous across enzyme classes (Bromberg, Aptekmann et al. 2022). (2) Another possibility is that proteins are often multifunctional, e.g. moonlighting, or evolved to carry out different functions by tuning of a small portion the protein (Babbitt, Hasson et al. 1996, Singh and Bhalla 2020). In this case, a fraction of the protein could be specific to a different activity. (3) In the same vein, a read could be part of multiple genes encoding different regions

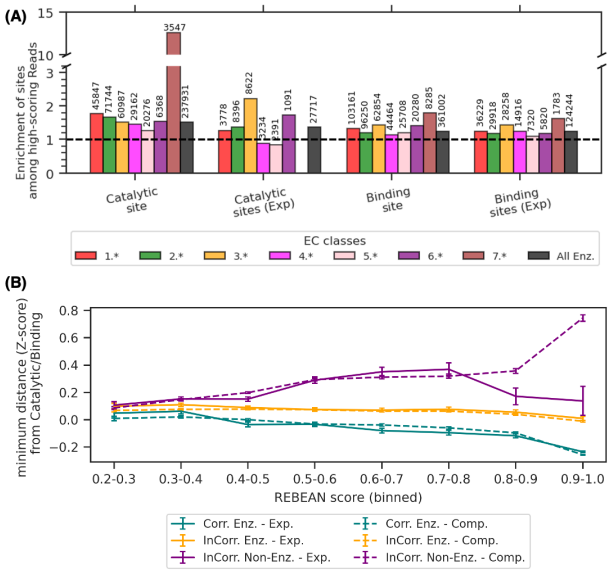

Figure 3: REBEAN high scoring enzymatic reads capture specifics of function. A) Enrichment in residues annotated as active/catalytic and binding residues among reads predicted with high confidence across enzyme classes (colors). All bars are higher than 1 (dashed line), indicating enrichment vs. numbers (bar heigh) expected for all reads. Numbers on top of the bars indicate the number of reads for which the enrichment was calculated, B) normalized spatial (3D) distance of the reads from functional residues (active and binding) computed using experimental (continuous lines) and Alphafold2 predicted protein structures (dotted lines). Reads correctly predicted by REBEAN to its gene's enzyme class (teal line) are structurally closer to functional residues in comparison to reads incorrectly labelled as non-enzymatic (magenta line) or assigned a different enzyme class (orange line).

of multiple proteins. The prokaryotic genome is densely packed, with an average of 88% of the genome encoding proteins (Chaumeil, Mussig et al. 2019). To elucidate the possible overlap of genes from the context of reads, we generated 124 million random reads of 200bps from 3,822 complete genomes from MGnify (**See Methods**). About 7% of these reads were shared among more than one predicted gene, suggesting multiple functions encoded by the same read.

REBEAN labels complete genes better than reads. In light of the above possibilities, we tested REBEAN's ability to label enzyme class based on predictions of all reads covering a complete gene. That is, we assigned each enzyme in our SwissProt enzyme dataset (**Table 1**) the average score across the corresponding reads. Compared to read-based annotations, our model attained a much higher average AUROC of 0.89 (range: 0.86-0.97) and was able to recover 48% of the enzymes at 90% precision (**Figure 2C & D**).
We observed a strong positive correlation between the number of enzymes correctly classified and the number of reads sampled per gene. For example, the number of genes labeled correctly at 90% precision shot up from 30% to 65% when the number of sampled reads increased from one to 50 reads per 1000 base pairs of a gene. This finding strongly suggests that REBEAN would perform better for samples at higher depths of the metagenomic sequencing (**Figure S2**).
We also assessed REBEAN's ability to predict SwissProt putative non-enzymes (**Table 1**), i.e. proteins without any experimental or computational annotation of catalytic activity. A protein sequence was predicted to be non-enzymatic only if all its corresponding reads were predicted non-enzymatic. We collected gene sequences for 473,854 proteins from SwissProt, fragmenting them into 5,305,894 reads, each 200 bp in length. Among these 473,854 genes, 231,203 (2,965,689 reads) are enzymes based on experimental and computational annotations in UniProt. But only 12,581 (165,766 reads) were with experimentally confirmed enzymatic activity. We evaluated REBEAN's ability to identify the 242,651 unannotated proteins from 231,203 annotated enzymes. REBEAN attained an AUROC and AUPRC of 0.826, which is somewhat lower than its performance on the enzymes (**Table S1**). There are two potential explanations for this result: (1) some of the unannotated proteins may eventually be labelled as enzymes and (2) unlike enzyme classes, non-enzymatic proteins do not belong to a single class—meaning that the diversity within the non-enzymatic group is likely similar to that of proteins in different enzyme classes.

REBEAN captures reads of functional importance. We investigated the biological relevance of REBEAN read prediction scores by exploring the distribution of catalytic and binding residues covered by these reads. We extracted both experimentally and computationally annotated functional site (catalytic and binding) residues from SwissProt proteins covered by our synthetic reads (UniProt 2023). We observed an enrichment of functional residues in reads with higher REBEAN scores. Reads predicted with a high prediction score (i.e. achieving 90% per-class precision; **Figure 2**) are 50% more likely to contain catalytic site residues than others (**Figure 3A**).
As function site annotations are often incomplete, we further investigated the spatial distribution of read-encoded amino acid residues around functional sites in protein structures. We calculated the minimum distance from the read-covered amino acids to the closest functional site residues. The calculated distance for each read was normalized across all reads from a given gene to avoid the misinterpretation due to the variable gene lengths. Again, we observed that translations of high scoring reads were significantly closer (in 3D space) to functional residues (**Figure 3B**).

Identifying orthologs without relying on sequence similarity. We extended our analysis of REBEAN embeddings to compare 707,947 reads sampled from 65,129 genes representing 50,000 orthologous and 50,000 non-orthologous gene pairs from OrthoDB (Kuznetsov, Tegenfeldt et al. 2023) (**Methods**). For each of the 100,000 gene pairs, we calculated the average cosine similarity between read embeddings from the two genes (**Eq. 10**). As expected from our earlier work (Hoarfrost, Aptekmann et al. 2022), we observed a significant distinction in embedding similarity between reads from orthologous and non-orthologous gene pairs at all the five taxonomic levels: genus, family, order, class and phylum (p-value < 1E-32; **Figure 4**). We also noted a small but significant drop in embedding similarity between reads from orthologous gene pairs in successive taxonomic levels from genus to phylum among (p-value < 4E-4 - 1E-102).

We further explored the performance of the average similarity scores in distinguishing reads from orthologous vs. non-orthologous gene pairs. That is, given a particular embedding similarity, is the gene pair more likely to be from one

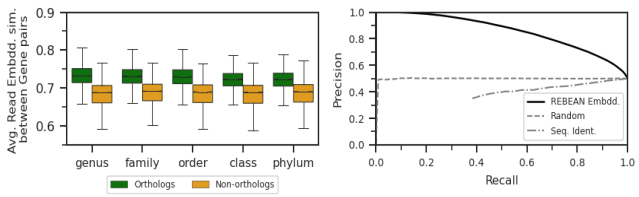

Figure 4: Fragments from Ortholog gene pairs share embedding space. A) Distribution of average embedding similarity between reads from orthologous and non-orthologous gene pairs, B) PR curve for the prediction of orthologous vs. non-orthologous gene pairs using embedding similarity.

or another type of gene pairs? We obtained an overall accuracy of 74.6%, with an AUROC and AUPRC of 0.843 and 0.858 (**Table 2**) – a performance significantly better than random (50%) or sequence identity (50%) computed using MMseqs (Hauser, Steinegger et al. 2016). As expected, we again observed better performance for lower taxonomic levels; e.g. for gene pairs at genus level, the accuracy was 79.6% (AUROC=0.88, AUPRC=0.9), while at phylum level the accuracy dropped to 70% (AUROC=0.81, AUPRC=0.82). Note that REMME (instead of REBEAN) embeddings attained a similar performance at this task.

Table 2: Predicting orthologs gene pairs across taxa

Taxa	Method	AUC ROC	AUC PR	F1 max	thresh-old	Re-call	Pre-cision	Spec-ificity	Ac-cu-racy
All	REBEAN	0.843	0.858	0.766	0.701	0.837	0.705	0.650	0.744
Genus	REBEAN	0.880	0.895	0.799	0.707	0.830	0.770	0.752	0.791
Family	REBEAN	0.851	0.866	0.772	0.703	0.846	0.709	0.654	0.750
Order	REBEAN	0.854	0.869	0.773	0.706	0.810	0.739	0.713	0.762
Class	REBEAN	0.824	0.835	0.753	0.699	0.820	0.697	0.644	0.732
Phylum	REBEAN	0.808	0.82	0.742	0.693	0.861	0.651	0.539	0.700
All	Seq. Sim-ilarity	0.319	0.526	0.667	0	1	0.5	0	0.5

To further explore whether the REBEAN embeddings identified reads from orthologous genes by relying on more precise estimates of sequence similarity, we chose a subset of 7,311 genes from 100 OrthoDB ortholog groups (OG; 79,783 reads; **Methods**) in our set. Among ~6.4B read pairs, correlation between sequence identity and embedding similarity of reads was comparably low for read pairs of (1) same gene (Pearson ρ=0.32), (2) same OGs (0.28), or (3) different OGs (0.25). In other words, REBEAN embeddings of the reads carried substantially more information than sequence identity alone.

We also observed that fine-tuning our models to predict enzymatic function reduced the correlation between embedding similarity and sequence identity. That is, REMME (pretrained LM) read embedding similarities were more sequence-driven than REBEAN (fine-tuned for EC prediction) read embedding similarities for both the enzyme and ortholog datasets (**Figure S5**). REMME is a DNA language model trained to encode information embedded in sequence and hence is expected to capture sequence identity. However, tuning the model to predict function loosens the need for explicit sequence representation. From the application point of view, REMME is useful in sequence encoding, identifying coding regions, and predicting likelihood of a DNA reads in the context of others, e.g. in genome assembly (**Figure S3**). In contrast, REBEAN is useful for exploring enzymatic functions encoded in read collections, e.g. metagenomes.

REBEAN can be used to discover novel enzymes from metagenomes. REBEAN's ability to capture function without requiring sequence similarity can help in identifying proteins that carry out known functions in novel ways. Here, we aimed to evaluate REBEAN's ability to mine potential novel oxidoreductases from a metagenomic set of reads. Oxidoreductases are broad class of enzymes that catalyze redox reactions by facilitating transfer of electrons. Oxidoreductases are ubiquitous as they are part of every biological energy production mechanism (Kim, Senn et al. 2013, Hay Mele, Monticelli et al. 2023) and are likely to have appeared on the scene early in history of life on Earth (Bromberg, Aptekmann et al. 2022).

For this analysis, we compiled a synthetic metagenome dataset comprising 124 million reads (200bp length), randomly sampled from the assembled sequences of 3,820 metagenome-assembled genomes (MAGs) in MGnify (Gurbich, Almeida et al. 2023). REBEAN predicted that approximately 10% (12,861,390) of these reads were enzymatic with a high confidence score of >0.9; of these 3,019,892 (23%) were deemed oxidoreductases (**Figure S4**). Almost all (98.7%) of these predicted oxidoreductase reads mapped to MGnify labelled MAG genes (1,126,995 genes) (Richardson, Allen et al. 2023).

In search for truly novel enzymes, we applied stringent filtering, retaining only reads that mapped to genes lacking any existing annotations (COG, EC, KEGG & Pfam). This analysis retained 407,428 genes (670,637 reads) distributed across all 3,820 MAGs. This result (36% of the genes in extracted MAGs) underscores the prevalence of currently unannotatable proteins in the microbiome spaces. We clustered (Hauser, Steinegger et al. 2016) these proteins at a 30% sequence identity, retaining 51,326 representative protein sequences. We then aligned these to SwissProt, retaining matches with ≤30% sequence identity. We ultimately retained 39,617 putatively novel proteins.

To examine whether these REBEAN-identified proteins could be oxidoreductases, we further selected 32,030 sequences of 100 to 800 residues in length and predicted their structures using ESMFold (Lin, Akin et al. 2023). We aligned these structures against the PDB database using FoldSeek (van Kempen, Kim et al. 2022) to find that 464 proteins (1.4%) matched 1,698 PDB chains (TM ≥ 0.9). Notably, 938 (55%) of these PDBs lacked an EC annotation and were designated as proteins of unknown function. Among the remaining 760 proteins with EC annotations, 373 were classified as oxidoreductases (49%; EC 1), confirming our predictions and REBEAN's ability to label enzyme classes without reference.

Note that the remaining matching PDB structures comprised 143 transferases (19%; EC 2), 128 hydrolases (17%; EC 3), 28 lyases (4%; EC4), 86 isomerases (11%; EC 5) and 2 ligases (EC 6). The top two non-oxidoreductase matches were

2.7.7.7 (42 structures, DNA-directed DNA polymerase) and 5.3.1.5 (31). Enzyme subclass 5.3.1.5 is an isomerase subclass of Intramolecular oxidoreductases (EC 5.3) that carries out oxidation and reduction within a molecule. We thus suggest that REBEAN errors may be driven by functional similarity of enzyme regions due to evolutionary and/or molecular activity constraints.

Interpreting REBEAN predictions through Embeddings. What did the model learn? To answer this question, we computed read-pair Euclidean similarity (**Eqn. 9**) and cosine similarity (**Eqn. 10**) among confidently predicted reads with REBEAN score above 0.90. For both metrics, we computed the average similarity (**Eqn. 11-12**) of each read with other reads in our dataset of 54,115 reads generated from 4,295 enzyme genes (**Table 1**). Note that both cosine and Euclidean similarity metrics returned similar trends, so from here on we only describe Euclidean similarity results. We observed average embedding similarity between reads belonging to same gene or enzyme class were higher than reads from different genes or enzyme classes (**Figure 5A**). The average similarity of a read with other reads sampled from same gene was 0.64±0.06 and varied widely from 0.37 to 1.0 (**Figure 5B**). On the other hand, read embeddings from different genes were only slightly lower (0.59±0.03) irrespective of their EC annotation (range: 0.30 to 0.64). These observations reinforce the conclusion that much of any given gene sequence is likely shared between enzymes for e.g. structural or evolutionary reasons, with only a small fraction being specific to a particular function.

Similarity between overlapping reads within a gene was higher than that of non-overlapping reads, i.e. as expected, reads that share the same exact sequence within a gene tend to have higher embedding similarity. As we had demonstrated in ortholog dataset, we did not observe any correlation between embedding similarity and sequence identity of read pairs from different genes [(Hauser, Steinegger et al. 2016), **Figure 5B**]. Together, these observations suggest that the model indeed learned functional, rather than solely sequence-based, signatures encoded in genomic data. We thus probed deeper into the question: how much function did the model learn?

We computed the embedding similarity between all pairs of reads within same EC class at first, second and third EC levels. At all three levels, reads from same enzyme classes were more similar than reads from different classes. Interestingly, we observed that read similarity (0.64±0.03, and 0.65±0.03) within same enzyme classes at level 2nd and 3rd (e.g. 1.2 and 1.2.3.*) were higher than level 1st (0.62±0.02), even though REBEAN was trained to only recognize 1st EC level classes. On the other side, the average similarity of reads of different, level one EC classes was 0.61 but similarity varied across enzyme classes (EC1-7) (**Fig. 5C**). For example, translocase (EC 7.*) reads tended to be least similar, in line

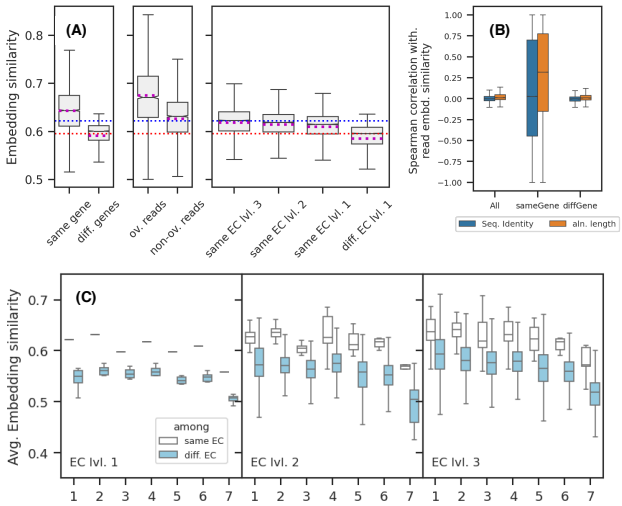

Figure 5: Deeper EC levels are captured by REBEAN embeddings. A) Distribution of average embedding similarities of each read vs. reads from the same gene, different gene, same EC, or different ECs for reads predicted with score above 0.90. The blue and red dotted lines indicate the median embedding similarity between reads of same EC level 3 classes and reads of different EC level 1 classes, respectively, i.e. similarity between reads of similar enzymatic function and different enzymatic functions. The mean embedding similarity for each distribution is marked by the magenta line. B) Distribution of Spearman correlations between the embedding similarity and sequence alignment for each read pair. Sequence identity is represented by the identity over the aligned length and the aligned length; C) Distribution of average embedding similarities of each read across the seven EC level 1 classes with reads from other enzyme classes at different EC levels (first, second and third).

with the fact that this class comprises subgroups that were once part of oxidoreductases, hydrolases, and lyases (McDonald and Tipton 2023).

Addressing the need to step up tool development. Accurate functional annotation of metagenomes is an important step in understanding microbial communities across diverse research areas. However, discovery and identification of novel functions in metagenomic data requires innovative approaches that transcend traditional homology-based methods (Prabakaran and Bromberg 2023). For metagenomic tools to effectively aid in discovery, they must be both robust and highly accurate. Recent advancements in language models have accelerated the field of protein structure and function prediction and we believe that a similar leap in metagenomic functional annotation is essential.

In this study, we introduce REMME, a foundational DNA language model designed to capture and interpret contextual information encoded within metagenomic reads. REMME's versatility makes it valuable for diverse research applications, including generating read embedding for machine

learning tasks, clustering reads for comprehensive ecological and evolutionary studies, and also, building fine-tuned models like REBEAN for specific downstream tasks.

In training REBEAN, we leveraged millions of diverse metagenomic reads, annotated with EC classes using the precise mi-faser tool, rather than relying solely on the limited datasets of experimentally annotates genes/proteins. This approach mitigated the risk of overfitting to specific enzyme sequences and allowed the model to generalize, capturing biologically relevant insights. We showed that REBEAN's predictions correlate with the presence of functionally significant residues. Remarkably, it can even identify related enzymes across different EC classes, such as oxidoreductases categorized under both EC 1 and EC 7, despite being trained to label them as distinct entities. We believe REBEAN holds considerable potential for the discovery of novel enzymes which includes novel proteins of know enzyme function (annotation) and enzymes of novel function (discovery). While the former task of annotation is easier than the later, had been the goal of tool development so far, the growing metagenomic exploration has created sufficient necessity to step up our focus.

Conclusion

The increasing accumulation of metagenomic samples presents both new challenges and opportunities for tool development. To address this, we have developed REMME, a robust DNA-language model tailored to capture biological context for metagenomic reads and read-based downstream tasks. We demonstrate REMME's adaptability and utility by fine-tuning it for enzyme class prediction from metagenomic reads. The extensive analysis highlights the fine-tuned model, REBEAN's potential for metagenomic read annotation and the discovery of novel enzymes. We believe REMME and REBEAN are valuable tools for metagenomic exploration.

MATERIALS AND METHODS

Model Development: REMME - Read Embedder for Metagenomic exploration.

Data. We obtained 1,496 genomic assemblies representing 1,496 prokaryotic species from the marine microbiome samples in MGnify (Richardson, Allen et al. 2023). From these assemblies, we extracted 72.9 million reads (72,966,774) by randomly sampling 20 reads per 1Kbp with an average length of 136 bp (60-300 bps), as described in (Hoarfrost, Aptekmann et al. 2022). The fraction of coding region residues in each read, along with the read position relative to the start codon were noted. Of these ~73 million reads, 65.1M and 7.8M reads were from coding and non-coding region, respectively. We randomly selected 20% (14.6M) of these reads for testing (13M and 1.6M of coding and non-coding reads, respectively) and used the remaining 58 million for training.

We clustered training and test datasets at 80% sequence identity using MMseqs (Hauser, Steinegger et al. 2016) to obtain 45.2M and 8.4M sequence-dissimilar representatives, respectively. Further, the training representatives were split into training (70%) and validation (30%) sets for pre-training.

Training. REMME is an encoder-only transformer model, holding six transform layers with eight multi-attention heads. Each input DNA sequence was transformed into a sequence of overlapping nucleotide triplets (tokens) with a stride of one nucleotide. The tokens were fed into a token embedder that encodes each token in the sequence as a numeric vector of length 128 and a position embedder that encodes the sequence position of each token as a vector of length 128. The combined encoding from token embedder and position embedder was fed into the encoder module containing six encoder layers with eight multi-head attention heads. We had used GELU activation along with a drop out of 0.10 throughout the model everywhere, unless specified otherwise. The total number of trainable parameters was 3,520,561.

The encoder model was trained to perform masked-token prediction. Here, 15% of the nucleotides to the encoder input were perturbed, similar to BERT (Devlin 2018) model training (80% masked and 10% random). The transformed read embeddings from the encoder module were then fed into three modules: 1) a *decoder* module, of three linear layers reconstructing the original nucleotide triplet sequence, 2) a *regression* module, comprising a 2d Fractional Max pooling layer, combined with two dense layers to predict the fraction of coding/non-coding residues in the read, and 3) a four-class *classifier* module, predicting the reading frame of the coding DNA sequence (CDS) that overlaps with a given read. The four classes represent three reading frames (classes 1, 2 & 3) and the non-transcribed reads (class 4). The losses from all three modules were summed and back propagated during training. The model was trained for 53 epochs until no significant decrease in total loss was observed.

Model Development: REBEAN - Read Embedding Based Enzyme Annotation.

Data. We collected 306 studies containing 19,316 diverse metagenomic samples of non-viral origin from the recent update to the SRA database (20230627) (Katz, Shutov et al. 2022). Among these, 3,136 metagenomic samples belonging to microbiomes associated with soil, water, coastal water, sea water, oligotrophic water, deep marine sediment, phyllosphere, algal, and anthropogenic material, we randomly selected 40 SRA experiment runs from each of the nine environments, yielding 360 runs representing 332 samples and containing 267.3M reads (**Appendix A**).

The mi-faser method (Zhu, Miller et al. 2018), using 41,640 enzyme sequences as reference, was used to annotate these metagenomic reads with an enzymatic activity assigned to their putative 'parent' genes, i.e. gene sequences from which the read is taken. The enzyme reference set comprised genes (extracted from NCBI RefSeq (O'Leary, Wright et al. 2016)) encoding Swiss-Prot proteins with Enzyme Commission (EC) number annotations and experimental evidence of protein existence (UniProt 2023, Sayers, Beck et al. 2024). Note that only a third (14,229) of these enzymes had experimental evidence of their enzymatic activity. Mi-faser annotated 59.5M of 267.3M reads (22%) as belonging to enzymes. We extracted reads of length 60 to 300 bps and clustered them, at 80% sequence identity using MMseqs, to retain 16.6M (16,624,341) and 112.3M (112,334,039) sequence-dissimilar reads of enzymatic and non-enzymatic origin, respectively. To create a balanced class distribution for training, we randomly sampled 2.4 million non-enzymatic reads – a number representative of the average number of reads in each EC class. Final composition of the 19 million read dataset was 17.7% of oxidoreductase (EC 1), 25.9% of transferases (EC 2), 11.3% of hydrolases (EC 3), 8.8% of lyases (EC 4), 5.8% of isomerases (EC 5), 10.5% of ligases (EC 6), 7.6% of translocases (EC 7), 12.5% of non-enzymes. We held out 3.8M (20%) of the reads as a test dataset and split the rest of the data (15.2M) into training (13.7M, 90%) and validation (1.5M, 10%) without altering the proportion of each enzyme class.

Training: REBEAN is a version of REMME that was fine-tuned to label reads that come from 'parent' enzyme-coding genes. REBEAN consists of the six encoder layers from REMME coupled with one classifier module comprising three dense layers. The classifier annotates a given read as being class one through eight, representing seven first level EC classes and non-enzymes, annotated by mi-faser as described above. We trained REBEAN using an ADAMW optimizer and cosine restarts scheduler. The model was trained for 188 epochs until no significant loss decrease was observed.

Model Evaluation

Performance measures. We used multiple standard metrics such as accuracy, recall, precision, specificity, F_1 score, and balanced accuracy (**Eqn. 1-6**) to evaluate predictive performance of REMME and REBEAN in training and several independent analyses. We also computed the Area Under the Curve (AUC) for Precision-recall (PR-AUC) and Receiver operating characteristic (ROC-AUC) curves (Fabian, Gaël et al. 2011). In addition, we calculated the mean squared error (**Eqn. 7**) and the mean absolute error (**Eqn. 8**) to quantify the difference between "ground truth" (f) and predicted coding fraction (f^p) of reads.

$$Accuracy = \frac{Correctly\ predicted\ positives\ and\ negatives}{Total\ number\ of\ positives\ and\ negatives} \quad \textbf{(Eqn. 1)}$$

$$Recall = \frac{Correctly\ predicted\ positives}{Total\ number\ of\ positives} \quad \textbf{(Eqn. 2)}$$

$$Precision = \frac{Correctly\ predicted\ positives}{Total\ number\ of\ predicted\ positives} \quad \textbf{(Eqn. 3)}$$

$$Specificity = \frac{Correctly\ predicted\ negatives}{Total\ number\ of\ negatives} \quad \textbf{(Eqn. 4)}$$

$$F_1\ score = \frac{2\,(Precision \times Recall)}{Precision + Recall} \quad \textbf{(Eqn. 5)}$$

$$Balanced\ accuracy = \frac{Recall + Specificity}{2} \quad \textbf{(Eqn. 6)}$$

$$MSE(f_i, f^p{}_i) = \frac{1}{no.\ of\ reads} \sum_{i=1}^{no.\ of\ reads} (f_i - f^p{}_i)^2 \quad \textbf{(Eqn. 7)}$$

$$MAE(f_i, f^p{}_i) = \frac{1}{no.\ of\ reads} \sum_{i=1}^{no.\ of\ reads} |f_i - f^p{}_i| \quad \textbf{(Eqn. 8)}$$

Embeddings similarity between reads. We used cosine similarity and Euclidean similarity for comparing read embeddings (**Eqn. 9-10**). Apart from the inherent differences in similarity distributions, our inferences were consistent across these metrics. For clarity, we only show the results of Euclidean similarity unless specified otherwise.

$$S_{Euclidean}(E1,\ E2) = \frac{0.5}{0.5 + \sqrt[2]{\sum_{i=1}^{l} (E_1{}^i - E_2{}^i)^2}} \quad \textbf{(Eqn. 9)}$$

$$S_{Cosine}(E1,\ E2) = \frac{E_1.E_2 + 1}{2} \quad \textbf{(Eqn. 10)}$$

In several comparison, we computed embedding similarity of a read against a set of reads representing a gene, gene set, genome or metagenome. In these cases, we represent the multiple read-pair similarities as an aggregate score using mean and maximum statistics. For a read *k* compared against a set of *n* reads, the aggregate score is defined using **Eqn. 11 and 12**. **Eqn. 11** depicts the average embedding similarity of the read *k* against the read set representing a gene or genome. **Eqn. 12** represents the embedding similarity with the best-aligned read in the read set.

$$\langle CS\ (R_k) \rangle = \frac{1}{n} \sum_{i=1}^{n} S_{Cosine}(E_k,\ E_i) \quad \textbf{(Eqn. 11)}$$

$$CS_{max}\ (R_k) = max\ \{S_{Cosine}(E_k,\ E_i)\ : i = 1, \ldots, n\} \quad \textbf{(Eqn. 12)}$$

Genes-to-synthetic reads. For multiple analysis, we generated synthetic reads from gene sequences to simulate metagenomic samples (**Table 1**). Unless otherwise specified, we randomly sampled, with no restrictions on overlap or coverage, ten gene fragments of length 200 nucleotides per 1Kbp of a gene. For example, for a gene of length 2Kbp, the number of reads sampled would be 20. This process of reads generation was also employed to generate reads from genomes.

UniProt annotated enzyme read dataset. From UniProt, we extracted 12,277 manually curated enzymes annotated with one unique EC number and experimental evidence of enzymatic activity (ECO:0000269) and mapped these to gene sequences using NCBI Entrez (UniProt 2023, Sayers, Beck et al. 2024). Of these, we selected the 4,295 enzymes belonging to prokaryotes and archaea. These included 1,096 (EC 1), 1,300 (EC 2), 785 (EC 3), 509 (EC 4), 315 (EC 5), 216 (EC 6), and 74 (EC 7) enzyme sequences. We further generated a dataset of 525,775 reads by randomly sampling 100 reads of length 200 bp per 1Kbp of a gene (as described in section **Genes-to-synthetic reads).** Each read was labelled with the 1st level enzyme class of the parent gene. Among the half million reads, 129,111 (24.6%), 160,065 (30.5%), 95,638 (18.2%), 60,325 (11.5%), 35,476 (6.7%), 34,365 (6.5%) and 10,795 (2.1%) were annotated as EC class 1 to 7, respectively, based on the parent gene. This dataset of 545,775 reads was used to access REBEAN's prediction performance. The performance statistics were computed by sampling 10% of the reads over 100 independent iterations to simulate a read sample of 10 reads per 1Kbp. We also created additional read datasets by varying the sampling rate (1, 2, 5, 10, 20, 50 and 100 reads per 1Kbp) and read length (50, 100, 150, 200, 250, 30 and 400 bp) to assess the influence of sequencing depth and other parameters.

Ortholog dataset. We randomly selected 1,000 different ortholog groups (OGs) from OrthoDB (Kuznetsov, Tegenfeldt et al. 2023) at each of the five taxonomic levels: genus, family, order, class and phylum. For each OG, we randomly chose five genes and identified ten random ortholog gene pairs between them. We then formed ten non-ortholog pairs by pairing genes across different OGs of the same taxon. The final dataset comprised 65,169 unique genes which includes 24,938 genes in 50,000 orthologous gene pairs and 65,169 genes in 50,000 non-orthologous gene pairs. We mapped these 65,129 genes to EMBL CDS to obtain respective gene sequences. 707,947 reads were sampled from these 65,129 genes as described above (as described in section **Genes-to-synthetic reads)**.

Synthetic Metagenome Read dataset. We collected 3,820 marine metagenome-assembled genomes (MAGs) from MGnify (Gurbich, Almeida et al. 2023) and generated 124 million genome fragments of 200bp length resembling metagenome reads. We sampled ten gene fragments per 1Kbp of a genome without restrictions on overlap. In addition to the genomes, we also retrieved annotations derived using EggNOG-mapper, HUMANN, hmmer, and KEGG mappings from MGnify (Eddy 2011, Seemann 2014, Huerta-Cepas, Szklarczyk et al. 2019, Gurbich, Almeida et al. 2023, Richardson, Allen et al. 2023).

Embedding visualization and statistical tests. We used TSNE for dimensionality reduction of the embedding space. TSNE 2D-projections were computed using (Poličar, Stražar et al. 2019) with following parameters: n_neighbors = 15, min_dist = 0.1 using *cosine* metric. Statistical significances were assessed through the Wilcoxon rank-sum test and Student's t-test using SciPy (Virtanen, Gommers et al. 2020).

Acknowledgements
This work was supported by NAI Grant Number: 80NSSC18M0093. Y.B. was also supported by the NSF (National Science Foundation) awards number 2310114 and 2310114 to Y.B.

References

Allen, K. N. and C. P. Whitman (2021). "The Birth of Genomic Enzymology: Discovery of the Mechanistically Diverse Enolase Superfamily." Biochemistry **60**(46): 3515-3528.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke and O. Zagnitko (2008). "The RAST Server: rapid annotations using subsystems technology." BMC Genomics **9**: 75.

Babbitt, P. C., M. S. Hasson, J. E. Wedekind, D. R. Palmer, W. C. Barrett, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon and J. A. Gerlt (1996). "The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids." Biochemistry **35**(51): 16489-16501.

Baier, F., J. N. Copp and N. Tokuriki (2016). "Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence-Function Relationships." Biochemistry **55**(46): 6375-6388.

Bharti, R. and D. G. Grimm (2021). "Current challenges and best-practice protocols for microbiome analysis." Brief Bioinform **22**(1): 178-193.

Bileschi, M. L., D. Belanger, D. H. Bryant, T. Sanderson, B. Carter, D. Sculley, A. Bateman, M. A. DePristo and L. J. Colwell (2022). "Using deep learning to annotate the protein universe." Nat Biotechnol **40**(6): 932-937.

Bromberg, Y., A. A. Aptekmann, Y. Mahlich, L. Cook, S. Senn, M. Miller, V. Nanda, D. U. Ferreiro and P. G. Falkowski (2022). "Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer." Sci Adv **8**(2): eabj3984.

Chaumeil, P. A., A. J. Mussig, P. Hugenholtz and D. H. Parks (2019). "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." Bioinformatics **36**(6): 1925-1927.

Davidi, D., L. M. Longo, J. Jabłońska, R. Milo and D. S. Tawfik (2018). "A Bird's-Eye View of Enzyme Evolution: Chemical, Physicochemical, and Physiological Considerations." Chem. Rev. **118**(18): 8786-8797.

Devlin, J. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Eddy, S. R. (2011). "Accelerated Profile HMM Searches." PLoS Comput Biol **7**(10): e1002195.

Fabian, P., V. Gaël, G. Alexandre, M. Vincent, T. Bertrand, G. Olivier, B. Mathieu, P. Peter, W. Ron, D. Vincent, V. Jake, P. Alexandre, C. David, B. Matthieu, P. Matthieu and D. Édouard (2011). "Scikit-learn: Machine Learning in Python." J. Mach. Learn. Res. **12**: 2825–2830.

Gurbich, T. A., A. Almeida, M. Beracochea, T. Burdett, J. Burgin, G. Cochrane, S. Raj, L. Richardson, A. B. Rogers, E. Sakharova, G. A. Salazar and R. D. Finn (2023). "MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues." J Mol Biol: 168016.

Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." Microbiol Mol Biol Rev **68**(4): 669-685.

Hauser, M., M. Steinegger and J. Soding (2016). "MMseqs software suite for fast and deep clustering and searching of large protein sequence sets." Bioinformatics **32**(9): 1323-1330.

Hay Mele, B., M. Monticelli, S. Leone, D. Bastoni, B. Barosa, M. Cascone, F. Migliaccio, F. Montemagno, A. Ricciardelli, L. Tonietti, A. Rotundi, A. Cordone and D. Giovannelli (2023). "Oxidoreductases and metal cofactors in the functioning of the earth." Essays Biochem **67**(4): 653-670.

Hoarfrost, A., A. Aptekmann, G. Farfanuk and Y. Bromberg (2022). "Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter." Nat Commun **13**(1): 2606.

Huerta-Cepas, J., D. Szklarczyk, D. Heller, A. Hernandez-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering and P. Bork (2019). "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." Nucleic Acids Res **47**(D1): D309-D314.

International Union of, B. and E. C. Webb (1992). Enzyme nomenclature, 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular biology on the nomenclature and classification of enzymes. San Diego, California, Published for the International Union of Biochemistry and Molecular Biology by Academic Press, Inc.

Ji, Y., Z. Zhou, H. Liu and R. V. Davuluri (2021). "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." Bioinformatics.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis (2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.

Kanehisa, M., Y. Sato and M. Kawashima (2022). "KEGG mapping tools for uncovering hidden features in biological data." Protein Sci **31**(1): 47-53.

Katz, K., O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister and C. O'Sullivan (2022). "The Sequence Read Archive: a decade more of explosive growth." Nucleic Acids Res **50**(D1): D387-D390.

Kim, J. D., S. Senn, A. Harel, B. I. Jelen and P. G. Falkowski (2013). "Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases." Philos Trans R Soc Lond B Biol Sci **368**(1622): 20120257.

Ko, K. K. K., K. R. Chng and N. Nagarajan (2022). "Metagenomics-enabled microbial surveillance." Nat Microbiol **7**(4): 486-496.

Kuznetsov, D., F. Tegenfeldt, M. Manni, M. Seppey, M. Berkeley, E. V. Kriventseva and E. M. Zdobnov (2023). "OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity." Nucleic Acids Res **51**(D1): D445-D451.

Lane, N. (2015). "The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'." Philos Trans R Soc Lond B Biol Sci **370**(1666).

Lin, Z., H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science **379**(6637): 1123-1130.

Locey, K. J. and J. T. Lennon (2016). "Scaling laws predict global microbial diversity." Proc Natl Acad Sci U S A **113**(21): 5970-5975.

Louca, S., F. Mazel, M. Doebeli and L. W. Parfrey (2019). "A census-based estimate of Earth's bacterial and archaeal diversity." PLoS Biol **17**(2): e3000106.

Madani, A., B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, Jr., C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser and N. Naik (2023). "Large language models generate functional protein sequences across diverse families." Nat Biotechnol.

McDonald, A. G. and K. F. Tipton (2023). "Enzyme nomenclature and classification: the state of the art." FEBS J **290**(9): 2214-2231.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman (2021). "Pfam: The protein families database in 2021." Nucleic Acids Res **49**(D1): D412-D419.

Nguyen, E., M. Poli, M. G. Durrant, A. W. Thomas, B. Kang, J. Sullivan, M. Y. Ng, A. Lewis, A. Patel, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu and B. L. Hie (2024). "Sequence modeling and design from molecular to genome scale with Evo." bioRxiv: 2024.2002.2027.582234.

O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy and K. D. Pruitt (2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." Nucleic Acids Res **44**(D1): D733-745.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko and V. Vonstein (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." Nucleic Acids Res **33**(17): 5691-5702.

Pan, S., C. Zhu, X. M. Zhao and L. P. Coelho (2022). "A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments." Nat Commun **13**(1): 2326.

Paysan-Lafosse, T., M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu and A. Bateman (2023). "InterPro in 2022." Nucleic Acids Res **51**(D1): D418-D427.

Poličar, P. G., M. Stražar and B. Zupan (2019). "openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding." bioRxiv: 731877.

Prabakaran, R. and Y. Bromberg (2023). "Functional profiling of the sequence stockpile: a review and assessment of in silico prediction tools." bioRxiv.

Pushkarev, A., K. Inoue, S. Larom, J. Flores-Uribe, M. Singh, M. Konno, S. Tomida, S. Ito, R. Nakamura, S. P. Tsunoda, A. Philosof, I. Sharon, N. Yutin, E. V. Koonin, H. Kandori and O. Beja (2018). "A distinct abundant group of microbial rhodopsins discovered using functional metagenomics." Nature **558**(7711): 595-599.

Richardson, L., B. Allen, G. Baldi, M. Beracochea, M. L. Bileschi, T. Burdett, J. Burgin, J. Caballero-Pérez, G. Cochrane, L. J. Colwell, T. Curtis, A. Escobar-Zepeda, T. A. Gurbich, V. Kale, A. Korobeynikov, S. Raj, A. B. Rogers, E. Sakharova, S. Sanchez, D. J. Wilkinson and R. D. Finn (2023). "MGnify: the microbiome sequence data analysis resource in 2023." Nucleic Acids Res. **51**(D1): D753-D759.

Ruscheweyh, H. J., A. Milanese, L. Paoli, N. Karcher, Q. Clayssen, M. I. Keller, J. Wirbel, P. Bork, D. R. Mende, G. Zeller and S. Sunagawa (2022). "Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments." Microbiome **10**(1): 212.

Ryu, J. Y., H. U. Kim and S. Y. Lee (2019). "Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers." Proc Natl Acad Sci U S A **116**(28): 13996-14001.

Sanderson, T., M. L. Bileschi, D. Belanger and L. J. Colwell (2023). "ProteInfer, deep neural networks for protein functional inference." Elife **12**.

Sayers, E. W., J. Beck, E. E. Bolton, J. R. Brister, J. Chan, D. C. Comeau, R. Connor, M. DiCuccio, C. M. Farrell, M. Feldgarden, A. M. Fine, K. Funk, E. Hatcher, M. Hoeppner, M. Kane, S. Kannan, K. S. Katz, C. Kelly, W. Klimke, S. Kim, A. Kimchi, M. Landrum, S. Lathrop, Z. Lu, A. Malheiro, A. Marchler-Bauer, T. D. Murphy, L. Phan, A. B. Prasad, S. Pujar, A. Sawyer, E. Schmieder, V. A. Schneider, C. L. Schoch, S. Sharma, F. Thibaud-Nissen, B. W. Trawick, T. Venkatapathi, J. Wang, K. D. Pruitt and S. T. Sherry (2024). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **52**(D1): D33-D43.

Seemann, T. (2014). "Prokka: rapid prokaryotic genome annotation." Bioinformatics **30**(14): 2068-2069.

Singh, N. and N. Bhalla (2020). "Moonlighting Proteins." Annu Rev Genet **54**: 265-285.

Solden, L., K. Lloyd and K. Wrighton (2016). "The bright side of microbial dark matter: lessons learned from the uncultivated majority." Curr Opin Microbiol **31**: 217-226.

Staley, J. T. and A. Konopka (1985). "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats." Annu Rev Microbiol **39**: 321-346.

Steen, A. D., A. Crits-Christoph, P. Carini, K. M. DeAngelis, N. Fierer, K. G. Lloyd and J. Cameron Thrash (2019). "High proportions of bacteria and archaea across most biomes remain uncultured." ISME J **13**(12): 3126-3130.

Tatusov, R. L., M. Y. Galperin, D. A. Natale and E. V. Koonin (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Res **28**(1): 33-36.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.

UniProt, C. (2023). "UniProt: the Universal Protein Knowledgebase in 2023." Nucleic Acids Res. **51**(D1): D523-D531.

van Kempen, M., S. S. Kim, C. Tumescheit, M. Mirdita, C. L. M. Gilchrist, J. Söding and M. Steinegger (2022). "Foldseek: fast and accurate protein structure search." bioRxiv.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and C. SciPy (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python." Nat Methods **17**(3): 261-272.

Zhu, C., M. Miller, S. Marpaka, P. Vaysberg, M. C. Ruhlemann, G. Wu, F. A. Heinsen, M. Tempel, L. Zhao, W. Lieb, A. Franke and Y. Bromberg (2018). "Functional sequencing read annotation for high precision microbiome analysis." Nucleic Acids Res **46**(4): e23.

Zhu, C., M. Miller, Z. Zeng, Y. Wang, Y. Mahlich, A. Aptekmann and Y. Bromberg (2020). "Computational Approaches for Unraveling the Effects of Variation in the Human Genome and Microbiome." Annual Review of Biomedical Data Science **3**(Volume 3, 2020): 411-432.