Distribution Agreement

Weiyang Zheng April 10th, 2025

Finetuned DNA Language Model Based- Classifiers Captures Significant Enzymatic Activity from Metagenomic Datasets

by

Weiyang Zheng

Yana Bromberg

Adviser

Biology

Yana Bromberg

Adviser

Kyle F. Biegasiewicz

Committee Member

Carl J. Yang

Committee Member

2025

Finetuned DNA Language Model Based- Classifiers Captures Significant Enzymatic Activity from Metagenomic Datasets

By

Weiyang Zheng

Yana Bromberg

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2025

Abstract

Finetuned DNA Language Model Based- Classifiers Captures Significant Enzymatic Activity from Metagenomic Datasets

By Weiyang Zheng

The surge of metagenomic sequencing data demands functional annotation methods that move beyond traditional homology-based approaches. In this study, we utilize REBEAN (Read Embedding-Based Enzyme Annotator), a fine-tuned DNA language model designed to predict enzymatic activity directly from raw nucleotide sequences, and developed two classifiers, REBEAN-Halo and REBEAN-Nitro, targeting halogenase and nitrogenase functions, respectively. REBEAN-Halo identified functionally important regions within known halogenases and detected 92 candidates of novel halogenases from marine metagenomes. REBEAN-Nitro, though undertrained, successfully distinguished higher nitrogenase activity in unfertilized agricultural soils relative to fertilized ones, aligning with ecological expectations. Both models highlight REBEAN's potential to uncover functionally relevant but sequence-divergent enzymes in complex metagenomic datasets, offering a powerful tool for advancing enzyme discovery and microbiome functional profiling.

Finetuned DNA Language Model Based- Classifiers Captures Significant Enzymatic Activity from Metagenomic Datasets

By

Weiyang Zheng

Yana Bromberg
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2025

Table of Contents

Table of Figures

Finetuned DNA Language Model Based-Classifiers Captures Significant Enzymatic Activity from Metagenomic Datasets

Weiyang Zheng, R Prabakaran, Yana Bromberg

Introduction

Advances in next-generation sequencing technologies have resulted in an unprecedented influx of genomic data from environmental samples. Analysis of microbial community genomic data, i.e. metagenomes, has transformed microbial ecology, allowing researchers to explore microbial diversity and function in ecosystems that were previously inaccessible[1]. However, this surge in data has outpaced our ability to analyze it effectively. Challenges stemming from the sheer volume of data, un-assemblable sequence fragments, and lack of annotated reference sequences are becoming more and more pronounced, highlighting the need for advanced computational tools to extract meaningful information. That is, while millions of new genomic sequences can be generated from a single study, turning these sequences into useful biological insights remains a significant hurdle[2]. Additionally, environmental metagenomics data often lacks standardized metadata and contextual information, which is crucial for interpreting such data in light of environmental constraints.

[1] Wooley, Godzik, and Friedberg, "A Primer on Metagenomics."
[2] Navgire et al., "Analysis and Interpretation of Metagenomics Data: An Approach."

Deep learning is increasingly being employed in metagenomics analysis for the purposes of annotating molecular functions carried out by the corresponding microbiomes [3]. Leveraging Deep learning methods, particularly those that rely on neural networks, can also help identify patterns and make predictions about microbial taxonomy, as well as gene functions, and metabolic pathways. For example, tools like DeepARG use deep learning models to predict antibiotic resistance genes with high accuracy based on metagenomic data (Arango-Argoty et al., 2018). However, while machine learning has advanced metagenomic analysis, the complexity of environmental samples still presents a challenge in extracting functional insights, especially for proteins with few, if any, known homologs.

DNA Language models are adopted by more and more researchers to extract information from nucleotides sequences. These models, for example, DNABERT, transforms input nucleotides into information-rich numerical representation of input sequences and can be utilized for versatile downstream tasks, like promotor identification and binding site prediction[4]. Being able to learn and extract biologically meaning representation of nucleotides sequences, DNA language model plays an increasingly important role in metagenomic analysis[5].

Halogenases are enzymes that catalyze the incorporation of halogen atoms—such as chlorine, bromine, or iodine—into organic compounds. These enzymes play a crucial role in the biosynthesis of halogenated natural products, i.e. natural compounds that contain halogen atoms[6]. Halogenated compounds are of much interest as they carry out diverse biological

[3] Libbrecht and Noble, "Machine Learning Applications in Genetics and Genomics."

[4] Ji et al., "DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome."

[5] Yan et al., "Recent Advances in Deep Learning and Language Models for Studying the Microbiome."

[6] Xu and Wang, "Independent Evolution of Six Families of Halogenating Enzymes."

activities and are commonly used in pharmaceuticals, agrochemicals, and synthetic biology[7]. For example, halogenated natural products, such as antibiotics and anticancer agents, benefit from the unique chemical properties that halogens impart, enhancing their bioactivity, stability, and pharmacokinetics. The discovery and engineering of novel halogenases that aid production of novel halogenated compounds, have opened new avenues in drug development and, particularly, in the production of new therapeutic agents with improved efficacy[8].

Nitrogenases are a class of enzymes that catalyze the reduction of atmospheric nitrogen (N_2) to ammonia (NH_3), a process known as biological nitrogen fixation. This reaction is fundamental to the nitrogen cycle, converting inert atmospheric nitrogen into a bioavailable form essential for the synthesis of nucleotides and amino acids, thereby supporting all forms of life [9]. Industrial nitrogen fixation involves Haber–Bosch process, which synthesizes ammonia from atmospheric nitrogen and hydrogen gas under high temperature and pressure. While this process has significantly boosted agricultural productivity by providing synthetic fertilizers, it is energy-intensive and contributes substantially to global carbon emissions[10]. Consequently, understanding and harnessing the efficiency of biological nitrogen fixation through nitrogenases is of great interest for developing sustainable agricultural practices and reducing environmental impacts associated with synthetic fertilizer production.

Here we used REBEAN, a model developed in the Bromberg Lab[11] to annotate the enzymatic activity encoded in metagenome samples. Starting with REMME – a transformer-based

[7] Smith, Grüschow, and Goss, "Scope and Potential of Halogenases in Biosynthetic Applications."
[8] Dong et al., "Structural Biology: Tryptophan 7-Halogenase (PrnA) Structure Suggests a Mechanism for Regioselective Chlorination."
[9] Threatt and Rees, "Biological Nitrogen Fixation in Theory, Practice, and Reality: A Perspective on the Molybdenum Nitrogenase System."
[10] Gu et al., "The Role of Industrial Nitrogen in the Global Nitrogen Biogeochemical Cycle."
[11] Prabakaran and Bromberg, "Deciphering Enzymatic Potential in Metagenomic Reads through DNA Language Model."

foundational DNA language model, REBEAN was a fine-tuned with metagenomic samples to predict high level enzymatic function descriptors of genes/proteins giving rise to metagenomic reads. We further fine-runed REBEAN to identify halogenase activity. Combinding the model with structural alignment we identified novel halogenases, sequence-dissimilar (i.e. non-homologous or very remotely homologous) to known enzymes. Furthermore, we finetuned the model to identify nitrogenase, another class of enzyme, instead of halogenases, from environmental data. The model, though undertrained, captures significant nitrogenase activity in metagenomic dataset. Together, our finding provides insight of the potential of REBEAN as template for various downstream metagenomic analysis tasks.

Results & Discussion

Building neural network classifiers to identify enzymatic functional representing nucleotides from metagenomic datasets:

REBEAN (Read Embedding Based Enzyme Annotator) is a functional classifier demonstrating robust predictive performance, leveraging the understanding on the context of reads within their “parent” enzymes. Our previous work described REBEAN’s extensive applicability to metagenome annotations at first Enzyme Commission level[12]. Here, we adopt the encoder of REBEAN, which were pretrained on genomics sequences and finetuned on metagenomic sequences and trained two classifiers using CLS token’s embedding REBEAN encoder’s output as input(Figure 1). CLS token is a special token of which embedding captures overall contextual information of the whole sequence and is specifically used for classification tasks[13]. The

[12] Prabakaran and Bromberg.

[13] Ji et al., “DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome.”

resulting classifiers, REBEAN-Halo and REBEAN-Nitro, predict how likely an input nucleotides sequence is representing halogenases (REBEAN-Halo) or nitrogenases(REBEAN-Nitro), i.e. the read is taken from a halogenase/nitrogenase gene. REBEAN-Halo achieves accuracy of 78.96% on test dataset (See Methods). By assigning the highest scoring nucleotides' score to the mapped protein in the test data, the model achieve accuracy of 92.56%. Notice that when mapped to the protein space, the model achieves high recall (100%) but low precision (10.46%). The high recall suggests that the model can capture known halogenases very confidently, while the low precision suggests that the model produces many false positive in predicting metagenomic dataset. This is due to that after the mapping, the result dataset is imbalanced (554 known nitrogenases v.s. 62905 "negative protein") This is potentially acceptable given that metagenomics datasets tend to contain functional "dark matter", i.e. proteins of which functions haven't been characterized[14]. The high false positive rate would allow us to explore the functionally unannotated space of input metagenomic dataset and facilitate with finding novel halogenases.

REBEAN-Nitro achieves 78.17% accuracy on test dataset (See Methods). Notice that REBEAN-Nitro converges at a training loss of 0.4414, which suggests the model could achieve lower loss and higher accuracy if being trained further with hyperparameter finetuning or reducing the complexity of the model. Given the limited scope of this project, we adopt the current model for further task, but as we demonstrate in the following sections, even the undertrained classifier, REBEAN-Nitro, can captures significant enzymatic activities from metagenomic datasets, highlighting the potential of REBEAN's output embedding for versatile downstream tasks like classifications.

[14] Pavlopoulos et al., "Unraveling the Functional Dark Matter through Global Metagenomics."

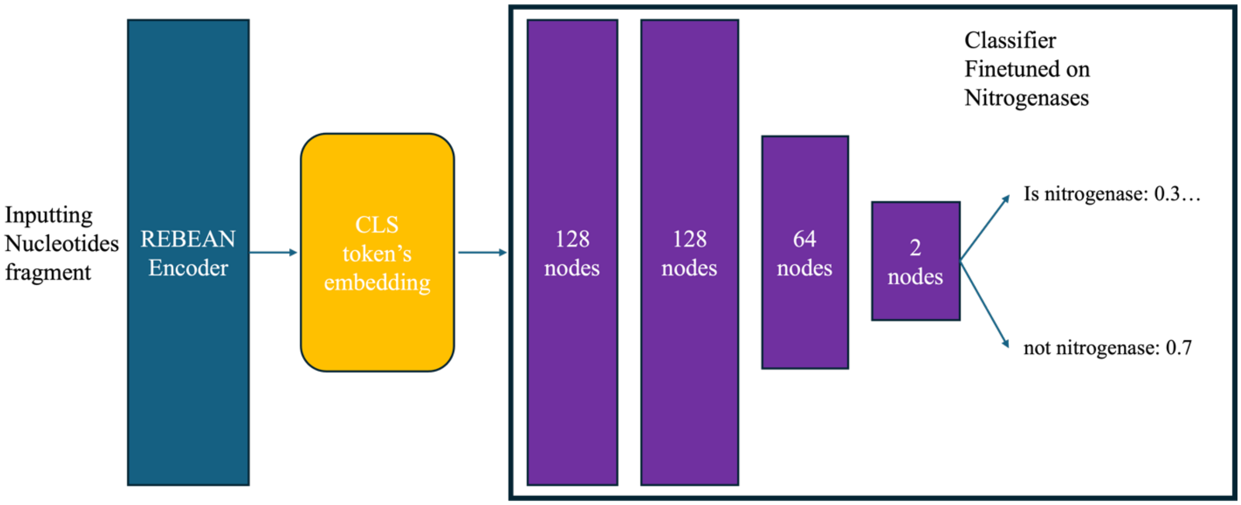

Figure 1 Illustration of REBEAN-Halo & REBEAN-Nitro's structure. The classifiers (in black bracket) are fully connected neural network that takes the embedding of CLS token encoded by REBEAN's encoder as input.

REBEAN-Nitro and REBEAN-Halo potentially captures functionally important region in enzymes

In order understand the underlying mechanism of how the classifiers distinguish enzyme v.s. non-enzyme, we investigates REBEAN-Halo's and REBEAN-Nitro's predictions on known halogenase and nitrogenase by calculates the average prediction scores per amino acid positions for rebH (halogenase) from *Lentzea aerocolonigenes,* and nifD (nitrogenase)from *Klebsiella pneumoniae* (Figure 1). We observe that the binding site of the enzyme, comparing to regions with less functional importance, tend to coincide with regions with higher average prediction scores. This implies that the classifiers potentially capture the functional information encoded in the input embedding of REBEAN's encoder and utilize it to identify enzyme origin reads from non-enzyme origin reads.

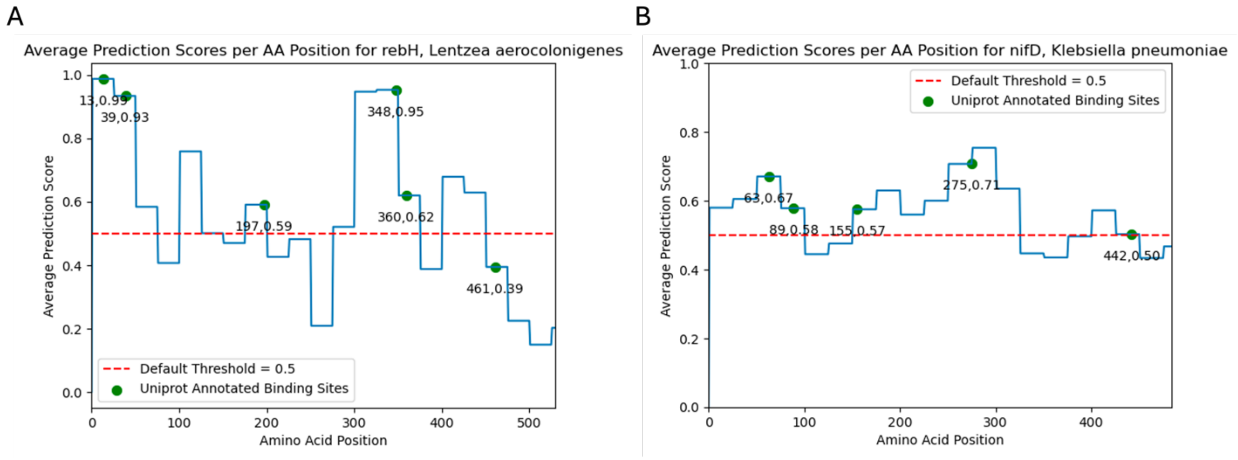

Figure 2 The average prediction scores per amino acid positions for rebH(A) and nifD(B). The x-axis is amino acid position along the length of the proteins, and y-axis is the average prediction scores of the position; the red horizontal line in both figures shows the default threshold for predicting whether a given nucleotides fragment is representing enzymatic function or not; the green dots are Uniprot annotated binding sites; labels under the green dots are amino acids position and score of the position.

REBEAN-Halo identify potential novel halogenases from metagenomics dataset:

In order to access the performance of the finetuned REBEAN-Nitro and discover novel halogenases, we use the classifier to retrieve nucleotides level prediction of how likely a given nucleotides fragment, from a metagenomics data, is representing halogenating function, i.e., the nucleotide is sampled from a halogenase gene. We select Sample MH127A083106DA1521 from Study MGYS00006577 in Malaspina Expedition 2010 Microbial Vertical Profiles Metagenomes [15]submitted to MGnify. [16] we selected the dataset given it is a marine metagenomic dataset collected at a depth of 20 meter, of which environment potentially contains halogenases that are not characterized previously. The original dataset contains 32031053 raw nucleotide reads and 1038326 predicted proteins. We align these nucleotides fragments to the predicted proteins to retrieve exact mapping (See Methods). 13456666 (42.01% of original 32031053 raw reads) of

[15] Duarte, “Seafaring in the 21st Century: The Malaspina 2010 Circumnavigation Expedition.”
[16] Richardson et al., “MGnify: The Microbiome Sequence Data Analysis Resource in 2023.”

the raw reads are mapped to 935315 (90.07% of 1038326 predicted proteins) of the predicted proteins. We further filter out those reads that are mapped to more than one predicted gene. The final input dataset consists of 7597466 raw nucleotides reads mapped to 830598 predicted proteins.

After retrieving nucleotides level prediction score, we assign the score of the highest scoring reads to it mapped protein to retrieve proteins level prediction score. We adopt such scheme given the assumption, discussed in above section, that the classifiers captures functional important region within an enzyme. We thus identify 18793 proteins (2.01% of 830598 prediction proteins) with very high confidence prediction (prediction score >0.9) by REBEAN-Halo.

In search for truly novel enzymes, we remove 37 sequences that aligns to known halogenases at 0.3 minimal sequence identity from these 18793 proteins. Then, to examine whether these remaining proteins could carry out halogenases' enzymatic function, we cluster the predicted structures of remaining proteins with the experimentally verified structures of 35 known halogenase at minimal threshold of 0.8 TMscore (See Methods) and retrieve 3 clusters that contains both known halogenases' structures and predicted halogenases' structures. These resulting 92 predicted halogenases are structurally similar to known halogenases but different in sequence identity and are potentially novel halogenases that were not reported before.

REBEAN-Nitro identifies more nitrogenase activity in untreated agricultural soil sample than in fertilized agricultural soil sample

In order to evaluate REBEAN-Nitro's ability to identify nitrogenases from metagenomic datasets, we run the prediction on two datasets with expected differences in nitrogenase enzymatic activities. It has been shown that the usage of inorganic fertilizer will drastically damper nitrogenases activities in agricultural soil environment[17]. Hence, we select two out of the five metagenomics samples that represent unfertilized agricultural soil sample (Control group) and highly fertilized, with inorganic fertilizer, agricultural soil sample(Treatment group), respectively[18]. The two datasets consist of ~9.9M (9938172) and ~15.7M (15743927) raw reads, respectively, after filtering out sequences that are shorter than 60bps or longer than 300bps, as required by REBEAN[19].

In order to confirm our expectation that control group contains more nitrogenase activity than treatment group, we align known nitrogenases to raw reads in both datasets, and result show control group contains significantly more (one tail p-score: 2.89e-09) reads that align to known nitrogenase (Figure 3A).

Then, we run REBEAN-Nitro to retrieve nucleotides level prediction for each raw read. The result of REBEAN-Nitro's prediction shows that it identifies significantly more nitrogenases activity at both high confidence (prediction score >0.85) and very high confidence (prediction score >0.9) (Figure 3B). At high confidence, REBEAN-Nitro identifies only 1456 reads (0.0147% of 9938172 raw reads) in treatment to represent nitrogenase gene while significantly more (one tail p-value ≈ 0), which is, 3793 reads (0.0241% of 15743927 raw reads) in control group to represent nitrogenase gene. Similarly, REBEAN-Nitro identifies significantly more (one tail p value = 5.21e-03) reads representing nitrogenases gene in control group compared to

[17] Shi et al., "Organic Manure Rather than Phosphorus Fertilization Primarily Determined Asymbiotic Nitrogen Fixation Rate and the Stability of Diazotrophic Community in an Upland Red Soil."
[18] Babalola and Enebe, "Metagenomes of Maize Rhizosphere Samples after Different Fertilization Treatments at Molelwane Farm, Located in North-West Province, South Africa."
[19] Prabakaran and Bromberg, "Deciphering Enzymatic Potential in Metagenomic Reads through DNA Language Model."

treatment group. Notice that none of the reads identified at 0.9 prediction score threshold align to known nitrogenase nor found significant alignment using BLAST searching against NCBI core nucleotides database[20]. These reads could potentially represent novel nitrogenases that were not reported before, while the scope of this project limit further investigation. More detail analysis on corresponding protein sequences and structure of these reads could reveal whether they are unidentified nitrogenases or not and suggests the need for extended evaluation on REBEAN-Nitro’s prediction results.

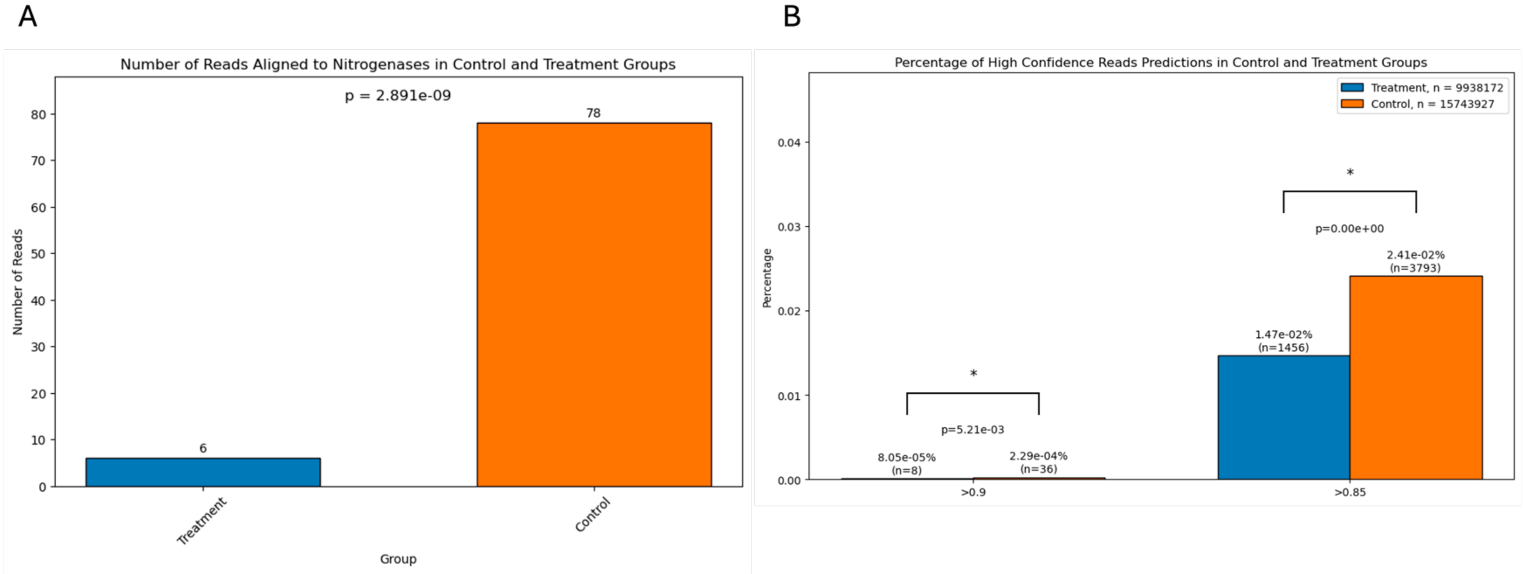

Figure 3 Nitrogenase Activity Identified via Alignment (A) and REBEAN-Nitro(B). (A) sequence alignment of raw reads to known nitrogenases via MMseqs at 0.8 minimum sequence identity identifies 78 aligned reads in control group and 6 aligned reads in Treatment group; y-axis is number of reads, x-axis is different group; two sample z test was performed (Z score = -5.937, one tail p value = 2.891e-09) and shows statistical significance; (B) REBEAN-Nitro identifies significantly more nitrogenases activity at both high confidence (prediction score >0.85) and very high confidence (prediction score >0.9). At high confidence, REBEAN-Nitro identifies 1456 reads in treatment group and 3793 reads in control group representing nitrogenase; At very high confidence, REBEAN-Nitro identifies 8 reads in treatment group and 36 reads in control group representing nitrogenases; y-axis is percentage of identified reads compare to total number of raw reads, and x-axis are different groups; two sample z test are performed on both confidence level(>0.9: Z score = -2.79, one tail p value = 5.21e-03; > 0.85: Z score = -16.3, one tail p value ≈0), both show statistic significance.

[20] ic Sa ers et al., “Database Resources of the National Center for Biot Ec Hnolog y Inf Ormation.”

Conclusion

Understanding functionalities presents in metagenomics dataset is crucial to advance our knowledge of microbial communities across different research fields, while effective function annotation of metagenomics dataset needs to go beyond traditional homology-based methods [21]. In this study, we build on top of the REBEAN described in Bromberg's Lab previous work to explore enzyme functions in metagenomic data with a focus on identifying two specific enzyme—halogenases and nitrogenases. Using REBEAN's DNA language model-derived embeddings, we fine-tuned two classifiers, REBEAN-Halo and REBEAN-Nitro, capable of capturing halogenase and nitrogenase activity from raw nucleotide reads.

REBEAN-Halo is able to detect functionally important regions within halogenase sequences, demonstrating signal enrichment of at known binding sites. Applying REBEAN-Halo to marine metagenomic data, we identified a significant fraction (2.01%) of proteins with high-confidence predictions. Further filtering with structural clustering revealed 92 candidate sequences that are structurally similar yet sequence-dissimilar to known halogenases, highlighting REBEAN's potential for discovery of novel enzyme.

In parallel, REBEAN-Nitro identifies significantly higher nitrogenase activity in unfertilized agricultural soil compared to fertilized soil, consistent with ecological expectations and supporting the model's biological relevance. REBEAN-Nitro also captured functional signal localized to binding site of known nitrogenases, reinforcing previous findings that REBEAN embeddings emphasize functional over sequence similarity.

[21] Ramakrishnan and Bromberg, "Functional Profiling of the Sequence Stockpile: A Review and Assessment of in Silico Prediction Tools."

Given that REBEAN-Nitro is undertrained, its capability of capturing significant functional signals further confirm the powerfulness of REBEAN for the novel enzymes' discovery from metagenomic datasets. We believe that REBEAN can be valuable tools in metagenomic annotations and analysis.

Methods

Training Datasets preparation

Halogenases are encoded by EC numbers 1.14.19.49, 1.14.19.58, 1.14.19.59, 1.14.19.60, 1.14.19.9. Using these EC numbers, Uniprot[22] was searched to find sequences that had the EC annotation corresponding to the curated set. We quered the NCBI database for genes that correspond to the curated dataset and retrieved 554 complete nucleotides sequences of known halogenases. Then, we randomly generate sequences in length range of 60 to 300 base pairs long, with mean of 136 base pairs long and a sampling rate of 100, i.e. we generate 100 sequences for each 1k base pair. Then, the resulting nucleotides fragments library were clustered at 0.8 minimum sequence identity with MMseqs[23]. The nitrogenase positive datasets are curated similarly. Nitrogenases are encoded by EC numbers 1.18.6.1, .18.6.2, 1.19.6.1, and we retrieved 192 complete nucleotides sequences of known nitrogenases.

We sampled from the annotated metagenomic dataset described in our previous work[24] to match the number of positive reads in training dataset, while retaining the proportion of each classes (EC1 to EC7 for 7 first level classes of Enzyme Commission, and EC8 for non-enzyme). The

[22] Bateman et al., "UniProt: The Universal Protein Knowledgebase in 2023."

[23] Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."

[24] Prabakaran and Bromberg, "Deciphering Enzymatic Potential in Metagenomic Reads through DNA Language Model."

resulting datasets are 303094 raw reads in total for REBEAN, 50% are synthetic nitrogenase fragments and 50% are sampled from metagenomic dataset. The datasets are split for training:validation:testing by 60%:20%:20%.

Model Training

REBEAN-Halo and REBEAN-Nitro are versions of REBEAN that was finetuned to predict whether an input sequence represent halogenase/nitrogenase or not. REBEAN-Halo and REBEAN-Nitro are both consist of the six encoder layers coupled with one classifier module comprising three dense layers. The classifier annotates a given read as being class one (is halogenase/nitrogenase) or class zero (not is halogenase/nitrogenase) We trained both model using an ADAMW optimizer and cosine restarts scheduler. The model was trained until no significant loss decrease was observed.

Sequence Alignment, Structural Prediction, and Structural Clustering

All sequence alignment described in this work are conducted via MMseqs[25]. The exact mapping was generated via MMseqs map –min-seq-id 1 –coverage 1 2; all other alignment, without further notice, was generated via MMseqs map –min-seq-id 0.8.

Structural Prediction was conducted via ESMFold[26]. Structural clustering was conducted via Foldseek[27].

[25] Steinegger and Söding, "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets."
[26] Lin et al., "Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model."
[27] van Kempen et al., "Fast and Accurate Protein Structure Search with Foldseek."

Bibliography

Babalola, Olubukola Oluranti, and Matthew Chekwube Enebe. “Metagenomes of Maize Rhizosphere Samples after Different Fertilization Treatments at Molelwane Farm, Located in North-West Province, South Africa.” *Microbiology Resource Announcements* 9, no. 43 (October 22, 2020). https://doi.org/10.1128/mra.00937-20.

Dong, Changjiang, Silvana Flecks, Susanne Unversucht, Caroline Haupt, Karl Heinz Van Pée, and James H. Naismith. “Structural Biology: Tryptophan 7-Halogenase (PrnA) Structure Suggests a Mechanism for Regioselective Chlorination.” *Science* 309, no. 5744 (September 30, 2005): 2216–19. https://doi.org/10.1126/science.1116510.

Duarte, Carlos M. “Seafaring in the 21st Century: The Malaspina 2010 Circumnavigation Expedition.” *Limnology and Oceanography Bulletin*. American Society of Limnology and Oceanography Inc., February 1, 2015. https://doi.org/10.1002/lob.10008.

Gu, Baojing, Jie Chang, Yong Min, Ying Ge, Qiuan Zhu, James N. Galloway, and Changhui Peng. “The Role of Industrial Nitrogen in the Global Nitrogen Biogeochemical Cycle.” *Scientific Reports* 3 (2013). https://doi.org/10.1038/srep02579.

ic Sa ers, Er W, Jeff Bec, Evan E Bolt on, J Rodne Brist er, Jessica Chan, Donald C Comeau, Ry an Connor, et al. “Database Resources of the National Center for Biot Ec Hnolog y Inf Ormation.” *Nucleic Acids Research* 52 (2024): 33–43. https://doi.org/10.1093/nar/gkad1044.

Ji, Yanrong, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. “DNABERT: Pre-Trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome.” *Bioinformatics* 37, no. 15 (August 1, 2021): 2112–20. https://doi.org/10.1093/bioinformatics/btab083.

Libbrecht, Maxwell W., and William Stafford Noble. “Machine Learning Applications in Genetics and Genomics.” *Nature Reviews Genetics*. Nature Publishing Group, May 18, 2015. https://doi.org/10.1038/nrg3920.

Navgire, Gauri S., Neha Goel, Gifty Sawhney, Mohit Sharma, Prashant Kaushik, Yugal Kishore Mohanta, Tapan Kumar Mohanta, and Ahmed Al-Harrasi. “Analysis and Interpretation of Metagenomics Data: An Approach.” *Biological Procedures Online*. BioMed Central Ltd, December 1, 2022. https://doi.org/10.1186/s12575-022-00179-7.

Pavlopoulos, Georgios A., Fotis A. Baltoumas, Sirui Liu, Oguz Selvitopi, Antonio Pedro Camargo, Stephen Nayfach, Ariful Azad, et al. “Unraveling the Functional Dark Matter through Global Metagenomics." *Nature* 622, no. 7983 (October 19, 2023): 594–602. https://doi.org/10.1038/s41586-023-06583-7.

Prabakaran, R, and Y Bromberg. "Deciphering Enzymatic Potential in Metagenomic Reads through DNA Language Model," December 11, 2024. https://doi.org/10.1101/2024.12.10.627786.

Ramakrishnan, Prabakaran, and Yana Bromberg. "Functional Profiling of the Sequence Stockpile: A Review and Assessment of in Silico Prediction Tools," July 14, 2023. https://doi.org/10.1101/2023.07.12.548726.

Richardson, Lorna, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, et al. "MGnify: The Microbiome Sequence Data Analysis Resource in 2023." *Nucleic Acids Research* 51, no. D1 (January 6, 2023): D753–59. https://doi.org/10.1093/nar/gkac1080.

Shi, Wei, Hui Yu Zhao, Yin Chen, Jin Song Wang, Bing Han, Cong Ping Li, Jun Yuan Lu, and Li Mei Zhang. "Organic Manure Rather than Phosphorus Fertilization Primarily Determined Asymbiotic Nitrogen Fixation Rate and the Stability of Diazotrophic Community in an Upland Red Soil." *Agriculture, Ecosystems & Environment* 319 (October 1, 2021): 107535. https://doi.org/10.1016/J.AGEE.2021.107535.

Smith, Duncan R.M., Sabine Grüschow, and Rebecca J.M. Goss. "Scope and Potential of Halogenases in Biosynthetic Applications." *Current Opinion in Chemical Biology*, April 2013. https://doi.org/10.1016/j.cbpa.2013.01.018.

Threatt, Stephanie D., and Douglas C. Rees. "Biological Nitrogen Fixation in Theory, Practice, and Reality: A Perspective on the Molybdenum Nitrogenase System." *FEBS Letters*. John Wiley and Sons Inc, January 1, 2023. https://doi.org/10.1002/1873-3468.14534.

Wooley, John C., Adam Godzik, and Iddo Friedberg. "A Primer on Metagenomics." *PLoS Computational Biology* 6, no. 2 (February 26, 2010): e1000667. https://doi.org/10.1371/journal.pcbi.1000667.

Xu, Gangming, and Bin Gui Wang. "Independent Evolution of Six Families of Halogenating Enzymes." *PLOS ONE* 11, no. 5 (May 1, 2016): e0154619. https://doi.org/10.1371/JOURNAL.PONE.0154619.

Yan, Binghao, Yunbi Nam, Lingyao Li, Rebecca A. Deek, Hongzhe Li, and Siyuan Ma. "Recent Advances in Deep Learning and Language Models for Studying the Microbiome," September 15, 2024. https://doi.org/10.3389/fgene.2024.1494474.