Github link to project: https://github.com/kbliu22/QBIO-460.git

Exploring Carbohydrate Turnover Through Predicting Enzyme Kinetics From Sequence and Substrate Data

By Kevin Liu

Abstract

Marine microbes are central to the ocean's carbon cycle, with carbohydrate metabolism playing a key role in carbon turnover. This process is mediated by carbohydrate-active enzymes (CAZymes), which catalyze the breakdown and transformation of complex polysaccharides. However, studying these enzymes is challenging, as approximately 99% of carbohydrate metabolizing marine microbes remain unculturable (Bodor et al., 2020). Advances in computational protein structure prediction offer a new opportunity to study enzyme function directly from amino acid sequences. This project aims to leverage these tools to predict enzyme activity and kinetic parameters using a combination of models, including graph convolutional networks and feedforward neural networks. By integrating data from UniKP with ocean-specific CAZymes, we aim to identify structure–function relationships, validate them against previous lab findings (Bodor et al., 2015), and improve predictions of enzyme kinetics for marine carbohydrate-processing microbes.

Introduction

Marine microbes play a crucial role in the ocean's carbon cycle, strongly influencing its climate impact. Phytoplankton absorb sunlight and atmospheric carbon dioxide (CO_2), converting it into organic carbon, while heterotrophic bacteria break down this organic matter, releasing CO_2 back into the atmosphere. Microbial enzymes act as a "bottleneck" in this process, controlling the rate and extent of carbon turnover. Characterizing these enzymes in marine microbes is challenging, as roughly 99% of carbohydrate-metabolizing microbes remain unculturable (Bodor et al., 2020). However, amino acid sequences are more readily available and can be leveraged for computational analyses. A key group of these enzymes are carbohydrate-active enzymes, or CAZymes, which catalyze the synthesis, modification, and degradation of complex carbohydrates. CAZymes are central to understanding microbial contributions to carbon cycling, as they mediate the breakdown and transformation of polysaccharides in marine environments.

Previously, the Levine Lab attempted to calculate binding energies across diverse protein–substrate interactions. However, factors such as substrate directionality made it computationally infeasible to manually curate all the data. Other published models have successfully predicted enzyme activity across a broad range of reactions (Yu et al., 2023), but when applied to a specific set of ocean microbes, the resulting relationships were weak. This project aims to build a more targeted dataset, integrating UniKP data with ocean-specific CAZymes. Using this combined dataset, we will explore potential structure–function

relationships and retrain the predictive model based on the UniKP framework, with the goal of improving activity predictions for marine microbial enzymes.

Methods

The data used in this study originated from two sources. The first consisted of general enzyme–substrate pairs obtained from the UniKP training sets, which aggregate enzymatic reaction data from public databases including UniProt and BRENDA. Because kinetic information is inconsistently reported across databases, the Km and kcat training sets contained different sets of unique enzyme–substrate pairs. The second source was a curated collection of CAZyme enzyme–substrate pairs compiled by the Levine Lab, for which both Km and kcat values were available. These two datasets had already been merged into a single file prior to processing. Enzyme sequences were represented using single-letter amino acid codes, and substrates were represented by their SMILES strings.

The combined dataset required several preprocessing steps. Some CAZyme entries listed multiple substrates for a single enzyme; these were expanded into separate rows while copying the corresponding kinetic parameters. Any entries missing Km or kcat values were removed. Leading or trailing whitespace in substrate SMILES strings was cleaned. Additionally, duplicate enzyme–substrate pairs were identified and removed. This resulted in 421 duplicates for kcat and 252 duplicates for Km being removed. After cleaning, the final dataset contained 7,063 general enzymes and 1,569 CAZymes for kcat, and 7,182 general enzymes and 1,530 CAZymes for Km.

To characterize the dataset prior to modeling, we examined the empirical distributions of Km and kcat using boxplots, and visualized the number of unique sequences and substrates using histograms. Dimensionality reduction was then performed using PCA, t-SNE, and UMAP. For these analyses, enzyme sequences and substrate SMILES had to be converted into numerical representations. We adopted the embedding approach used in the UniKP project: enzyme sequences were embedded using ProtT5-XL-UniRef50, and substrates were embedded using UniKP's pretrained SMILES transformer. This produced 1024-dimensional embedding vectors for both sequences and substrates.

Because the ExtraTrees model was identified by UniKP (Yu et al., 2023) as a top-performing method for kinetic parameter prediction, we first trained separate ExtraTrees regressors to predict Km and kcat using the generated embeddings. Hyperparameter tuning was performed to identify stronger model performances that could be used for comparison.

To evaluate whether neural architectures could improve performance, we developed a custom deep learning model. Enzyme and substrate embeddings were processed in parallel: each 1024-dimensional input vector was passed through two fully connected layers with ReLU activation, reducing dimensionality before concatenation into a fused representation. This fused vector was then processed through several additional hidden layers, followed by a final regression output layer. Separate networks were trained for Km and kcat predictions.

Hyperparameter tuning was performed to identify the best model and assess whether neural architectures could achieve performance comparable to other methods.

Results

As shown in Figure 1, the dataset is skewed toward general enzymes, with CAZyme entries representing a smaller fraction of the total data. We also observe relatively few unique substrate SMILES strings, suggesting constrained chemical diversity that may influence model generalization. Notably, the Km and kcat datasets are comparable in their distributions and exhibit parallel structural patterns.

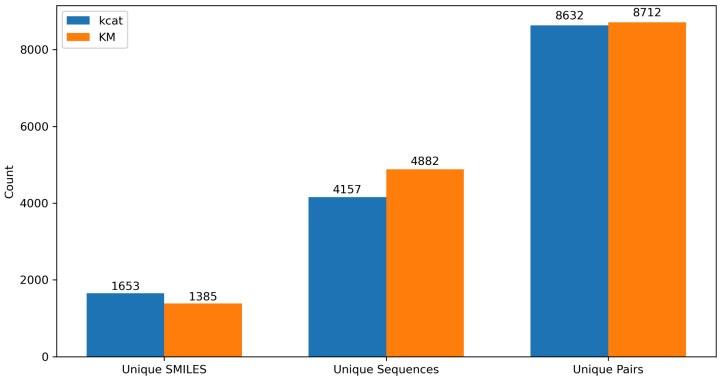

Figure 1. Enzyme-Substrate Pairing Overview

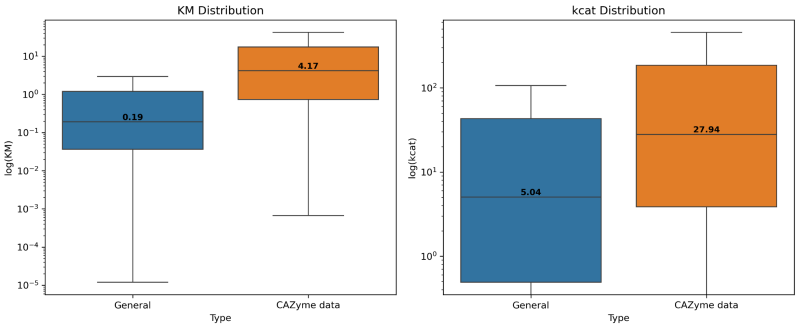

Figure 2. Enzyme-Substrate Kinetic Value Overview

Figure 2 further illustrates the imbalance in the distribution of kinetic values. We applied a logarithmic transformation to normalize the scale across measurements. While CAZyme kcat and Km values generally lie within the distribution of the general enzyme data, the coverage is not exact. In both metrics, CAZymes display marginally higher mean values.

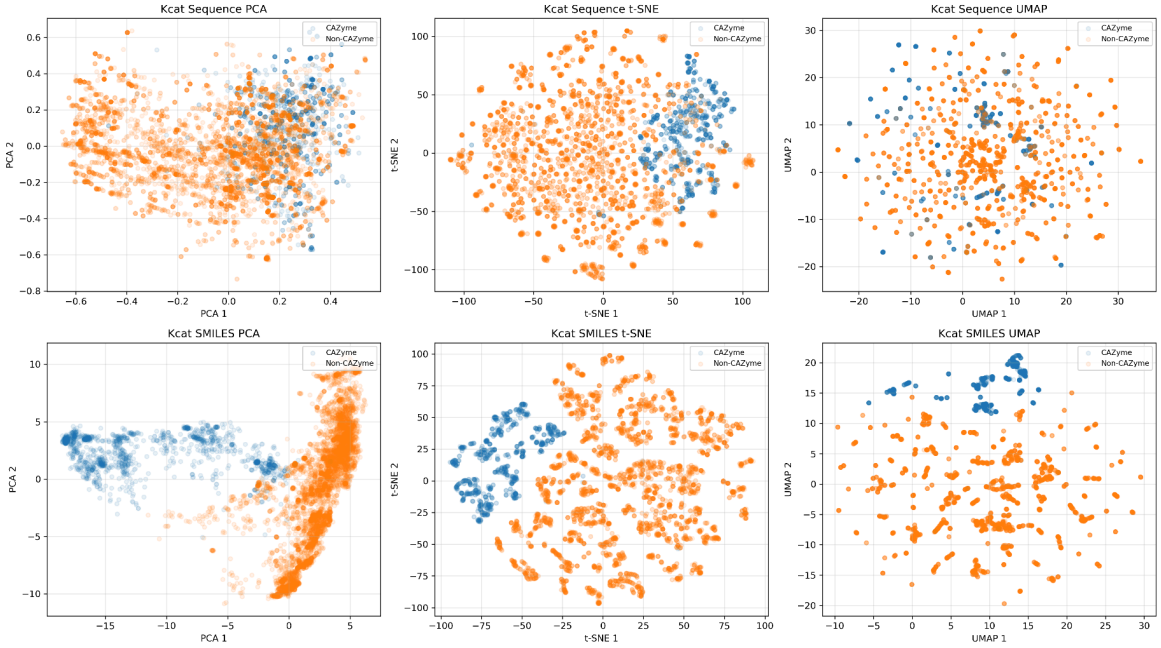

Figure 3a. kcat Dimensionality Reduction

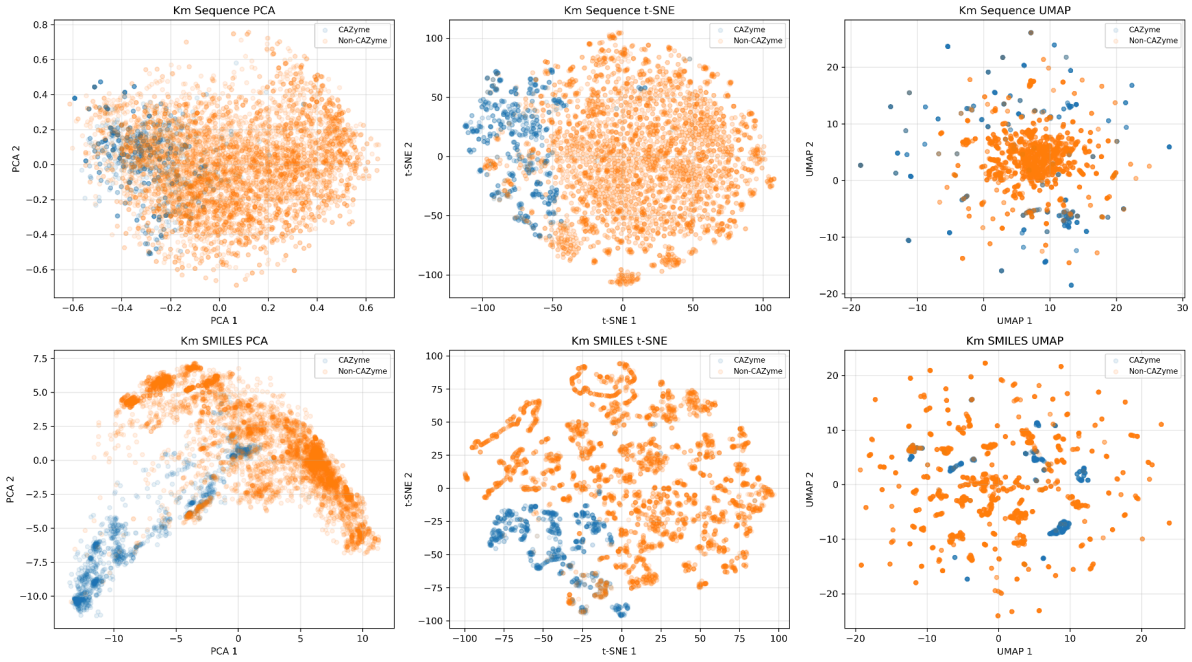

Figure 3b. Km Dimensionality Reduction

Dimensionality-reduction analyses allow us to characterize the structure of the embedding space. In Figure 3a, the PCA projection of enzyme sequence embeddings shows that CAZymes overlap with general enzymes in their global distribution but occupy a more restricted region of the space. Local structure inferred from t-SNE similarly reveals that CAZyme sequences form a compact cluster positioned near the boundary of the general enzyme distribution. UMAP further accentuates this separation, with CAZymes displaced from the dense core of general enzymes. For the kcat substrate SMILES embeddings, the PCA plot indicates that CAZyme substrates deviate more strongly from general-enzyme substrates than the sequences do, a pattern supported by both t-SNE and UMAP. It is clear that the CAZymes occupy a slightly different space than general enzymes.These trends are reproduced in Figure 3b for the Km embeddings, though the CAZyme substrates in the Km set show slightly greater overlap with general substrates than those in the kcat set. Overall, CAZymes form distinct clusters in both sequence and substrate embedding spaces, which helps explain the reduced performance of the original UniKP model when evaluated on CAZyme-specific data.

In the next part, ExtraTrees models were trained on the embeddings to predict the kcat/Km values. The data was separated with an 80/20 training/test split among both general enzymes and CAZymes in order to maintain an equal representation of both types across the datasets. A simple hyperparameter search was then conducted to identify some stronger combinations that could be used as a comparison to the UniKP results and the neural networks.

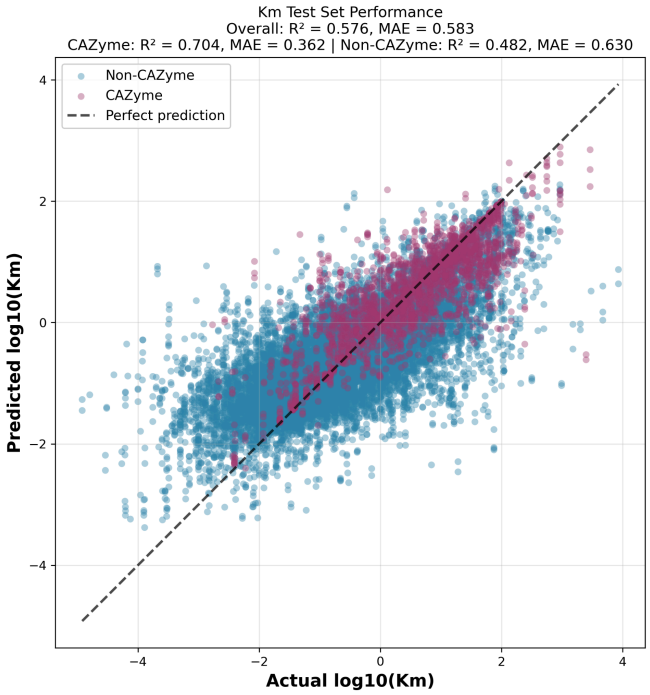

Figure 4a. Km ExtraTrees Performance

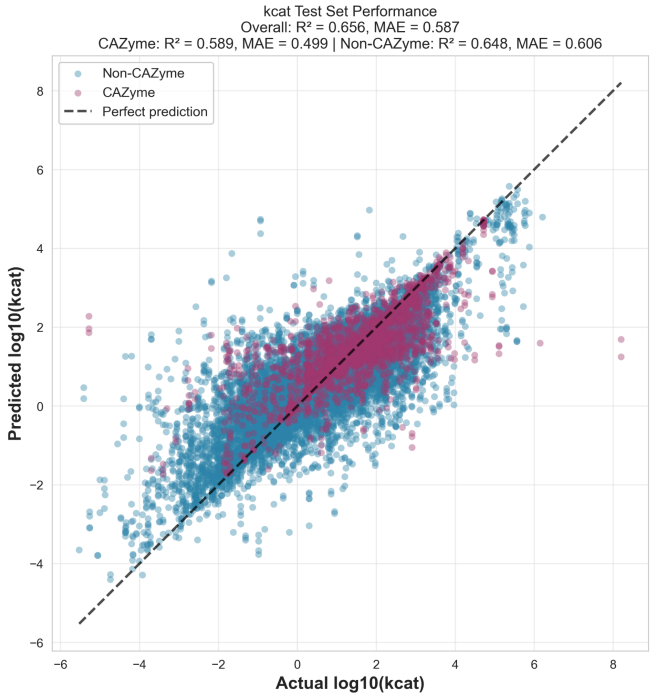

Figure 4b. kcat ExtraTrees Performance

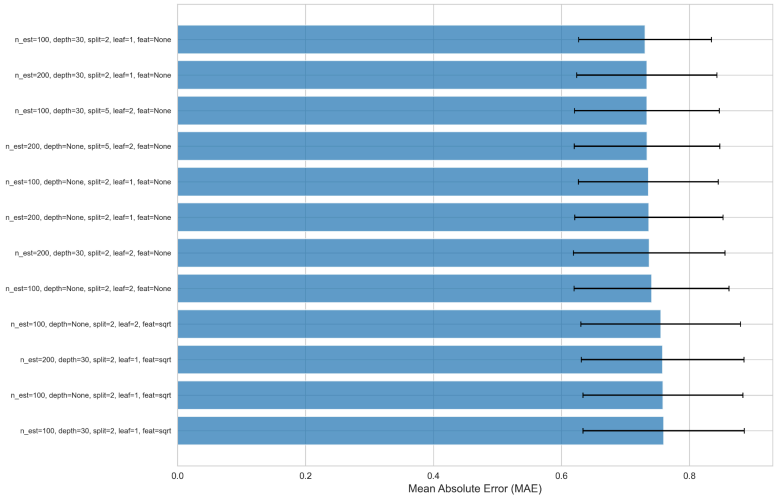

Figure 5a. Km Hyperparameter Tuning

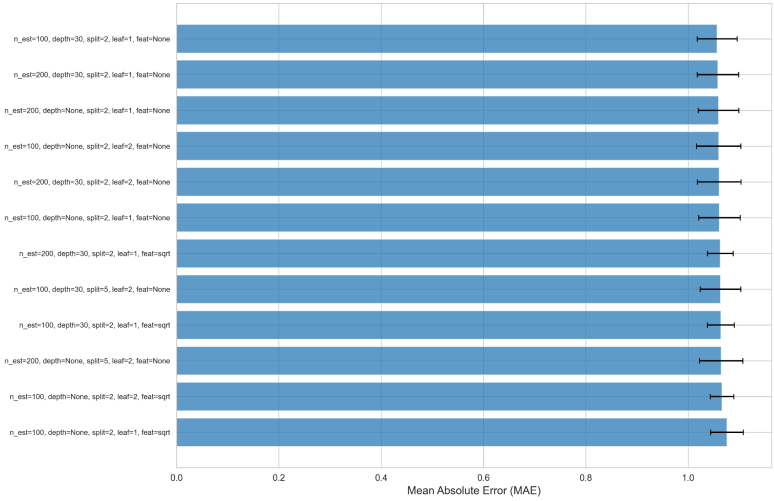

Figure 5b. kcat Hyperparameter Tuning

Figures 5a and 5b indicate that the hyperparameter configurations tested resulted in only minor variation in model performance. Although additional optimization could be explored, these results imply that the ExtraTrees architecture itself may be the limiting factor. Looking at figures 4a and 4b, the close agreement between our R^2 values and those reported by the UniKP framework also supports the notion that performance may be approaching the model's intrinsic upper bound. The plots show relatively strong performance on CAZyme data, with the Km model achieving a notably better fit than the kcat model. Nevertheless, interpretation should be made cautiously, as the CAZyme subset is substantially smaller and the resulting R^2 values are correspondingly more susceptible to the influence of outliers.

Next, we trained the neural network models described in the Methods section. As with the ExtraTrees models, the Km and kcat datasets were split into 80% training and 20% testing, with proportional representation of both general enzymes and CAZymes maintained in each partition. The best model selection was performed using 10-fold cross-validation to identify the best-performing networks.

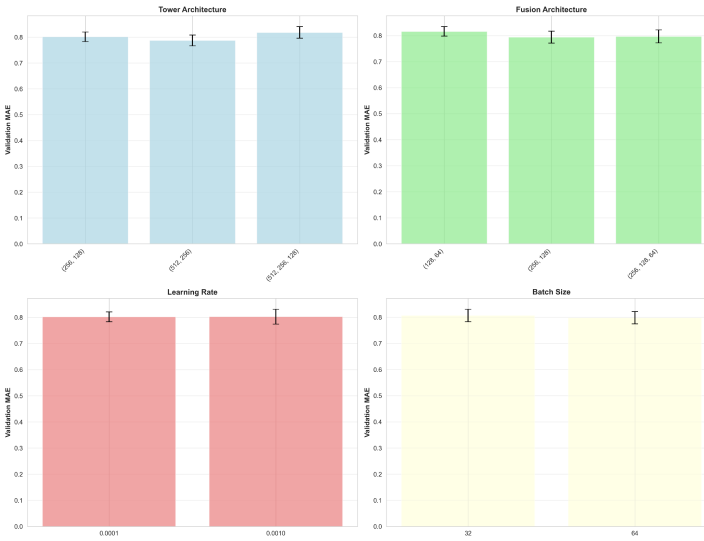

Figure 6a. kcat Neural Network Tuning

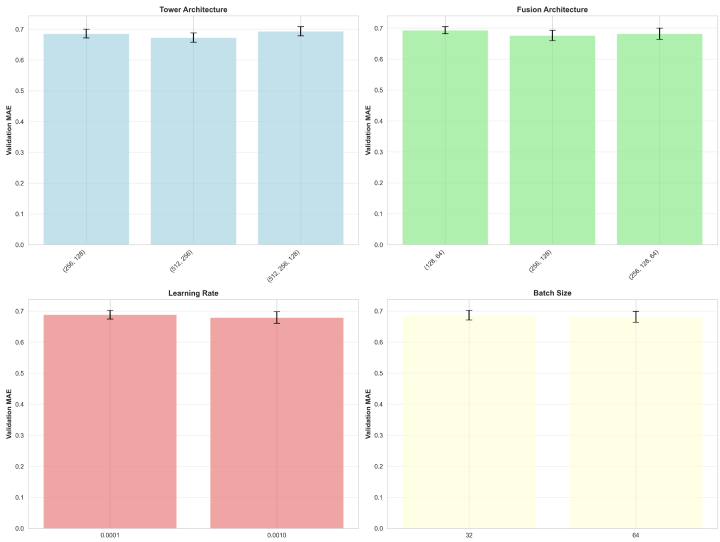

Figure 6b. Km Neural Network Tuning

From the hyperparameter tuning in figures 6a and 6b, the differences in loss are not very significant among both the Km and kcat models. The best parameters from the tuning were then used for the architecture of the following best models. The resulting architectures were nearly identical with only the kcat model utilizing one further hidden layer within its fusion architecture. Potentially, more drastic changes and more parameters could be measured for greater improvements to the model performance.

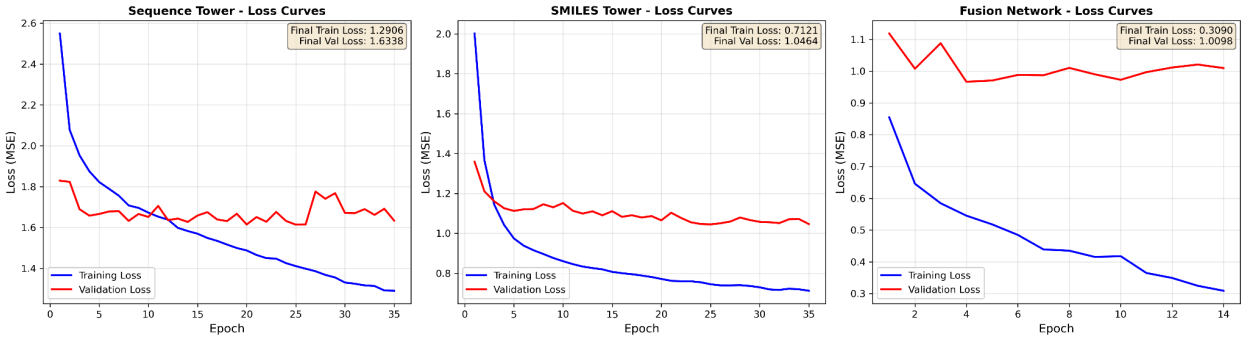

Figure 7a. kcat Neural Network Loss Curves

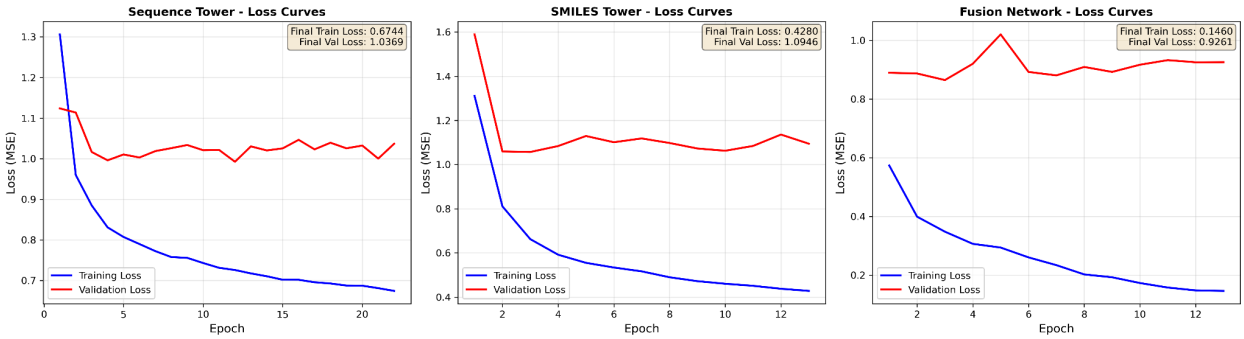

Figure 7b. Km Neural Network Loss Curves

The neural network models yielded performance metrics comparable to but slightly lower than those obtained with the ExtraTrees regressors. The Km model achieved a testing R^2 of 0.468 (training R^2 = 0.736), while the kcat model reached a testing R^2 of 0.534 (training R^2 = 0.741). Figures 7a and 7b show that although the training loss decreased steadily, the test loss improved only marginally, suggesting overfitting. This limited generalization ability, despite 10-fold cross-validation, points to potential shortcomings in the model architecture.This could potentially point to issues in the strategy used to merge the enzyme and substrate embeddings. Alternative design choices, such as earlier fusion of the vectors or ensemble methods (e.g., stacking), may enhance performance. Nevertheless, the predicted versus actual scatter plots still show a reasonable fit, indicating that the neural network framework has potential with further refinement.

Discussion

Taken together, our findings support the feasibility of predicting CAZyme kinetics *in silico*. The dimensionality-reduction analyses revealed that CAZymes occupy a distinct, though overlapping, region of the embedding space relative to general enzymes, providing a mechanistic explanation for the poor transferability of the UniKP model. Retraining the ExtraTrees regressors on the combined dataset enabled us to recover predictive performance for CAZymes that closely matched that of the general enzymes, demonstrating that the underlying kinetic patterns are detectable. The neural network results suggest additional avenues for methodological refinement. In contrast to ExtraTrees, which showed limited sensitivity to hyperparameter tuning, the neural architecture offers substantial room for improvement. Although current performance lags slightly behind the ExtraTrees models, enhancements to vector-fusion strategies, architectural depth, or ensemble methods may yield stronger results. It is also important to acknowledge limitations introduced during data preprocessing. Removing duplicate entries reduced the overall sample size, and expanding enzyme–substrate pairs where a single enzyme had multiple substrates may not accurately capture biological variability, as substrates may not share identical kinetic values. These considerations highlight opportunities for refining data handling in future work.

Importantly, this work has meaningful implications for understanding ocean microbial communities and the carbon exchange processes they drive. Much of the ocean remains

undersampled, especially deep, remote, or rapidly changing regions, making it difficult to measure enzymatic activity directly. Because biochemical characterization requires cultured isolates, which represent only a small fraction of marine microbes, our knowledge of microbial carbon transformations is inherently limited. By enabling in silico prediction of CAZyme kinetics from sequence alone, this framework provides a way to infer microbial carbon-processing capacity even in environments where direct measurements are not feasible. These models could be paired with metagenomic, transcriptomic, or proteomic time-series data from oceanographic expeditions to estimate potential carbon-degradation rates across space and time. Broader impacts include improving global carbon-cycle models, which currently lack mechanistic microbial parameters, and enhancing predictions of how the ocean's carbon-sequestration capacity may shift under climate change. More generally, this approach helps functionally interpret the vast diversity of uncultured marine microbes, offering a scalable tool for linking enzyme sequence data to ecosystem-level processes.

References

Badur, Ahmet H., Matthew J. Plutz, Geethika Yalamanchili, Sujit Sadashiv Jagtap, Thomas Schweder, Frank Unfried, Stephanie Markert, Martin F. Polz, Jan-Hendrik Hehemann, and Christopher V. Rao. "Exploiting Fine-Scale Genetic and Physiological Variation of Closely Related Microbes to Reveal Unknown Enzyme Functions." Journal of Biological Chemistry 292, no. 31 (August 2017): 13056–67. https://doi.org/10.1074/jbc.m117.787192.

Badur, Ahmet H., Sujit Sadashiv Jagtap, Geethika Yalamanchili, Jung-Kul Lee, Huimin Zhao, and Christopher V. Rao. "Alginate Lyases from Alginate-Degrading Vibrio Splendidus 12b01 Are Endolytic." Applied and Environmental Microbiology 81, no. 5 (March 2015): 1865–73. https://doi.org/10.1128/aem.03460-14.

Bodor, Attila, Naila Bounedjoum, György Erik Vincze, Ágnes Erdeiné Kis, Krisztián Laczi, Gábor Bende, Árpád Szilágyi, Tamás Kovács, Katalin Perei, and Gábor Rákhely. "Challenges of Unculturable Bacteria: Environmental Perspectives." Reviews in Environmental Science and Bio/Technology 19, no. 1 (February 1, 2020): 1–22. https://doi.org/10.1007/s11157-020-09522-4.

Yu, Han, Huaxiang Deng, Jiahui He, Jay D. Keasling, and Xiaozhou Luo. "UNIKP: A Unified Framework for the Prediction of Enzyme Kinetic Parameters." Nature Communications 14, no. 1 (December 11, 2023). https://doi.org/10.1038/s41467-023-44113-1.