Machine Learning Applied to the Optimization of Plastic-Degrading Enzymes

Romeu Fernandes[1], Pedro Soares[2,3], and João Carneiro[2,3]

[1] Informatics Department, University of Minho, 4710-057 Braga, Portugal
[2] Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, 4710-057 Braga, Portugal
[3] Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, 4710-22057 Braga, Portugal

1 Introduction

1.1 The Global Plastic Crisis

Plastic pollution has become a major problem over the years and is degrading the quality of the surroundings worldwide. Plastic production has surged globally since the 1950s, reaching 380 million tons annually (1). Traditional plastics take centuries to decompose, worsening the problem (2).

A noteworthy fact is that only 9% of plastic waste is recycled while the remaining percent is either burnt or dumped into landfills where it continues to pollute the environment. This has led to nearly 8 million tons of plastic ending up in the oceans each year (3). As a result, this excess of plastic has created marine garbage patches around the world, the largest being near the Pacific Ocean, which is more than three times the size of France (4).

Often, marine animals get caught in plastic waste or swallow plastic pieces, resulting in injury, hunger, and eventually death (5). Microplastics - particles smaller than 5 mm - have spread through all levels of the food web and can be found in remote regions, from deep ocean trenches to Arctic ice (6). Recently, microplastics have been found in human blood, placenta, and lung tissue, raising substantial alarm for future health problems (7).

Moreover, traditional methods of producing plastic are very fossil fuel intensive, as almost 6% of global oil consumption is directed towards the production of plastic, which is expected to grow to 20% by 2050 (8). As such, the search for biological methods for the breakdown of plastic, especially by using enzymes, has become a key focus of research to address this multifaceted problem (9)(10).

1.2 Enzymatic Solutions for Plastic Degradation

The identification of plastic-degrading enzymes in a range of microorganisms has created an exciting new avenue in the biological management of plastic waste. Synthetic polymers, such as plastics, were previously believed to be extremely resistant to biological degradation but now can be cleaved into monomers and oligomers (11)(12).

The microbial sources of these enzymes are rather broad, including bacteria from plastic polluted environments like *Ideonella sakaiensis*, which produces PETase for degrading polyethylene terephthalate (PET), and numerous fungi that produce cutinases that can hydrolyze several synthetic polymers (9)(13). The range of known plastic degrading enzymes includes polyethylene terephthalate hydrolases (PETases), cutinases, lipases, esterases, laccases, manganese peroxidases, and alkane hydroxylases, which all look for different types of plastics or certain chemical bonds within the plastic (14)(15).

These enzymes act through a variety of mechanisms: hydrolases such as PETase hydrolyze ester bonds in polyesters via hydrolysis, while oxidative enzymes (e.g., laccases and peroxidases) degrade polymer chains via free radical mechanisms (10). Catalytic processes usually occur by the initial generation of enzyme–substrate complexes on the plastic surface, which is then followed by nucleophilic attacks on susceptible bonds within the polymer backbone (16). Although, with reaction rates that are far too slow for cost-effective plastic waste disposal, their inherent catalytic efficiencies are sometimes insufficient for industrial-scale applications (17). Additionally, many natural plastic-degrading enzymes exhibit limited stability under the harsh conditions often encountered in waste processing facilities, such as elevated temperatures, extreme pH values, or the presence of inhibitory compounds (12).

The heterogeneity of plastic waste streams, which include, among others, various polymer types and polymer additives, as well as contaminants, adds more challenges to enzymatic degradation systems (18). These limitations emphasize the necessity of using enzyme engineering techniques to improve the functional properties of the natural plastic degrading enzymes for real applications in waste management and recycling systems.

1.3 Machine Learning Approaches in Enzyme Engineering

Not surprisingly, machine learning (ML) is a game changer in enzyme engineering, marking a shift towards data-driven optimization strategies away from traditional rational design and directed evolution approaches. Such a computational strategy enables tuning of enzyme traits while skipping a complete mechanistic interpretation of sophisticated structure-function correlations (19)(20). Conventional approaches to enzyme engineering have largely relied on heuristic expertise or random mutagenesis and selection, which can be costly and highly limited.

Different ML architectures have been shown to be effective for enzyme engineering. Trained in existing experimental data, supervised learning approaches (e.g., random forest, support vector machines, neural networks) can predict the performance of an enzyme based upon sequence or structural features (21)(22). Deep learning approaches, particularly convolutional neural networks and graph neural networks, have shown to be particularly effective at modeling the intricate relationships between protein sequence, structure, and function (23)(24).

When it comes to plastic degrading enzymes, ML models have been used to identify optimal mutations to improve key properties such as catalytic efficiency, thermal stability and substrate specificity (25). Often these include sequence-derived features, such as properties of amino acids and evolutionary data, structural datasets including distance maps, solvent accessibility and secondary structure elements and also molecular dynamics simulations (19)(26).

ML-based methodologies have emerged as transformative tools in the field of enzyme engineering, offering the capability to uncover latent patterns and relationships within experimental datasets that are typically imperceptible through traditional analysis. This advantage significantly accelerates the conventional process of biocatalyst development by predicting and systematically excluding non-viable enzyme variants, thereby enabling the efficient design of optimized biocatalysts with reduced time and resource expenditure (26)(27). Moreover, their models can be further trained on more data, leading to an increasingly virtuous cycle of enzyme design, where each round of predictions and experimental validation allows for improved model accuracy for the next round of predictions.

1.4 Computational Solutions to Plastic Pollution

With production of 380 million tons annually and only 9% being recycled, the global plastic crisis calls for quick and radical solutions. Two promising avenues for this project involve leveraging machine learning techniques in enzyme engineering and exploring biological degradation through the utilization of plastic-degrading enzymes. Through hydrolysis or oxidative processes, naturally occurring enzymes including PETase from *Ideonella sakaiensis* and several fungal cutinases have shown the capacity to break resistant synthetic polymers. Nevertheless, natural enzymes exhibit substantial limitations, particularly in terms of their catalytic efficiencies for industrial applications, their stability under adverse processing conditions, and their efficacy in addressing the complexity of heterogeneous waste streams composed of diverse polymers and chemical additives.

From conventional rational design and directed evolution approaches towards data-driven optimization strategies, the integration of machine learning marks a paradigm change. ML approaches hasten the creation of strong biocatalysts by predicting beneficial mutations and effectively navigating the large sequence space of possible enzyme variants. Models of the complex interactions between protein sequence, structure, and function have shown success using many ML architectures including supervised learning techniques and deep learning methods.

Combining biotechnology with computational techniques applies the inherent benefits of both areas to possibly overcome the constraints of conventional plastic waste management and generate a virtuous cycle of enzyme design in which every round of predictions and experimental validation increases the model's accuracy. This project adopts a multidisciplinary approach aimed at addressing the environmental challenges associated with plastic waste management through innovative solutions.

2 Project Objectives and Aims

2.1 Primary Objective

Develop and implement a machine learning framework for the identification, optimization, and characterization of novel plastic-degrading enzymes, with enhanced catalytic efficiency and substrate specificity.

2.2 Specific Aims

Database Development and Knowledge Integration

- Improve the Plastizyme database of known plastic-degrading enzymes by integrating data from UniProt, NCBI Protein, PDB, and AlphaFold.
- Standardize and pre-process enzyme sequence and structural data to ensure quality and consistency for downstream machine learning applications.

Predictive Model Development

- Design and train machine learning models to predict enzyme functionality based on sequence and structural features.
- Compare performance of advanced deep learning approaches with classical machine learning methods to identify optimal modeling strategies for plastic degrading enzyme prediction.

Structure-Function Relationship Analysis

- Clarify the structure-function relationships of plastic-degrading enzymes via computational modeling and docking simulations.
- Use AlphaFold3 software to create precise structural representations of promising enzyme candidates.
- Implement automated HADDOCK docking workflows for evaluating enzyme-substrate interactions with different plastics polymers.

Integrated Computational Pipeline Development

- Develop a seamless computational pipeline that uses sequence analysis, structural prediction, molecular docking, and machine learning to find and verify high-potential enzyme variants.
- Incorporate docking parameters and binding affinity data into machine learning models to improve forecast accuracy of enzyme activity.

Performance Validation and Optimization

- Systematically evaluate the predictive performance of developed models using Orange Data Mining software and suitable validation metrics.
- Maximize accuracy in detecting enzymes with increased catalytic activity by optimizing model parameters.

Biotechnological Application Assessment

- Evaluate the potential of identified enzyme candidates for practical applications in industrial plastic recycling operations and environmental bioremediation.

This project is meant to tackle the international plastic pollution crisis by leveraging computational approaches to accelerate the discovery and optimization of enzymes capable of degrading various plastic polymers. By combining machine learning with structural biology approaches, one will have a methodical system for identifying enzyme candidates for experimental confirmation and later biotechnological application.

3 Methodologies

This study aimed to enhance the previous Plastizyme machine learning (ML) algorithm by integrating high-confidence structural data generated with AlphaFold3 and comprehensive molecular docking parameters for all enzyme structures. The updated workflow was specifically designed to leverage these new data sources, thereby improving the predictive power and interpretability of the Plastizyme platform.

3.1 Structural Modeling and Database Integration

Over 125 enzyme structures, representing the main classes relevant to plastic degradation (laccases, peroxidases, and hydrolases), were modeled using AlphaFold3. Each sequence was submitted to the official AlphaFold3 server, and the resulting models were evaluated for quality using per-residue pLDDT scores and Predicted Aligned Error (PAE) matrices. To ensure compatibility with downstream tools, a custom Python script utilizing Biopython was developed to batch-convert all AlphaFold3 output files from .cif to .pdb format.

3.2 Docking Simulations and Feature Extraction

All processed 3D models underwent molecular docking simulations using HADDOCK. Structures were pre-processed in PyMOL to remove solvent molecules and alternate atom locations, ensuring clean input files. Docking outputs were further analyzed with PRODIGY-Ligand to extract quantitative binding affinity and interaction interface features. These docking parameters, together with structural descriptors from AlphaFold3 models, were incorporated as new features in the machine learning pipeline.

3.3 Machine Learning Workflow

The full computational workflow was executed for the primary enzyme types (laccases, peroxidases, and hydrolases). Using Orange Data Mining, raw assay data and newly generated structural and docking features were pre-processed, balanced (via SMOTE), and split into training and validation sets. Feature selection was performed to reduce multicollinearity, and six classification algorithms were trained and evaluated using fivefold cross-validation. The integration of AlphaFold3-derived features and docking parameters into the Plastizyme ML model was successful, as evidenced by improved preliminary performance metrics for the tested enzyme classes.

3.4 Ongoing Accuracy Assessment

Given the scale of the dataset—encompassing over 700 enzyme models—the full evaluation of model accuracy and generalizability will only be possible after the workflow has been applied to all available 3D structures. Initial results from the main enzyme types demonstrate the feasibility and benefit of incorporating advanced structural and docking data, setting the stage for a comprehensive performance assessment upon completion of the entire workflow.

4 Results

4.1 Structural Predictions with AlphaFold3

A total of 125 biomolecular structures were modeled using AlphaFold3, the latest advancement in DeepMind's protein structure prediction suite. AlphaFold3 represents a significant improvement over previous iterations, offering high-accuracy predictions for complex targets, including protein-ligand assemblies, nucleic acid interactions, and multimeric complexes. All sequences were submitted to the AlphaFold3 official server, and the resulting models exhibited consistently high per-residue pLDDT scores (above 80), as well as reliable Predicted Aligned Error (PAE) matrices, indicating strong confidence in the spatial organization of structured regions (Fig. 1). These quality metrics are in line with recent benchmarks, which have shown that AlphaFold3 achieves high accuracy for protein structure and protein–protein interaction predictions, with most models exceeding established thresholds for structural reliability. Such predictive quality confirms the suitability of these models for downstream applications, including molecular docking and interaction analysis.

Conversion of AlphaFold3 Models from CIF to PDB Format

AlphaFold3 outputs structural models in the .cif (Crystallographic Information File) format, which, while standard for crystallography, is not universally compatible with downstream bioinformatics tools that require the .pdb (Protein Data Bank) format. To address this, a custom Python script utilizing the Biopython Bio.PDB module was developed to automate the batch conversion of AlphaFold3 outputs to .pdb files. The script recursively identified model_0.cif files, parsed them using MMCIFParser, converted the structures with PDBIO,

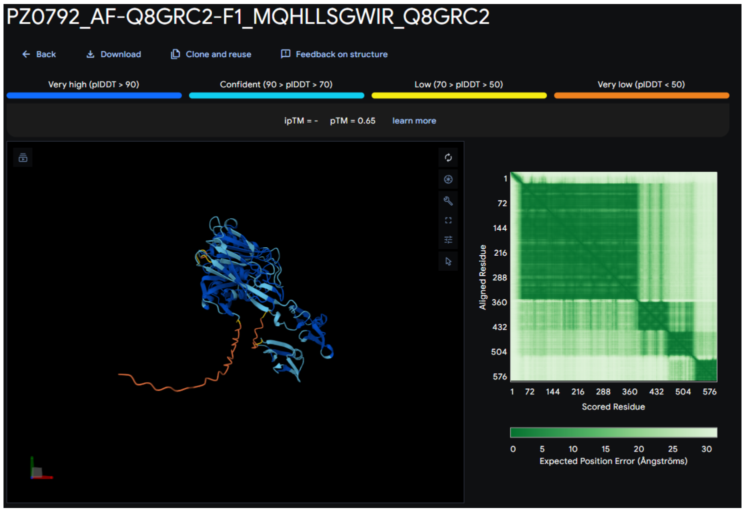


Fig. 1. Example of an AlphaFold3 result page for a single target protein structure prediction. The 3D model is colored by pLDDT score, representing prediction confidence levels. The right panel displays summary statistics, confidence plots, and model download options, providing insights into the reliability and accuracy of the predicted structure.

and saved the resulting files to a dedicated "Converted_Pdbs" directory. This automated approach ensured consistent processing and error checking for all 125 models used in this study.

4.2 Docking Simulations with HADDOCK

Input File Preparation

Prior to docking, all structural models underwent visual inspection and preprocessing in PyMOL. Solvent molecules (HOH) were removed using the command remove resn HOH, and the cleaned structures were saved in PDB format. To resolve alternate atom locations, the pdb_selaltloc tool from the pdbtools suite was employed:

```
pdb_selaltloc –A no_waters_file.pdb > file_no_altloc.pdb
```

Retaining only atoms with alternate location identifier "A" or none. This preprocessing ensured that the final structures were optimized for HADDOCK docking simulations.

Docking and Binding Affinity Analysis

Docking simulations were performed using HADDOCK via Python script utilizing Selenium to automate batch submissions to the HADDOCK web server, with the top-ranked complex from each run selected based on HADDOCK's scoring functions. To efficiently estimate binding affinity and characterize the interaction interface for all 125 complexes, PRODIGY-Ligand web server intended to be used. This approach could enable high-throughput extraction of quantitative metrics on the stability and strength of the predicted complexes, thereby complementing the docking results with robust biophysical parameters.

4.3 Machine Learning Workflow and Integration of New Parameters

The initial machine learning models were developed and validated using data from the existing Plastizyme database, which included raw assay data and molecular descriptors. Two Orange Data Mining workflows were implemented for this purpose. The first workflow (Plastizyme_balanced_ML_workflow_09062025.ows-figure 2) managed data import, preprocessing, SMOTE-based class balancing, feature filtering, model training, cross-validation, and performance visualization. The second workflow (Plastizyme_ML_predictions_09062025.ows – figure 3) applied the trained models to hold-out data for class and probability prediction.

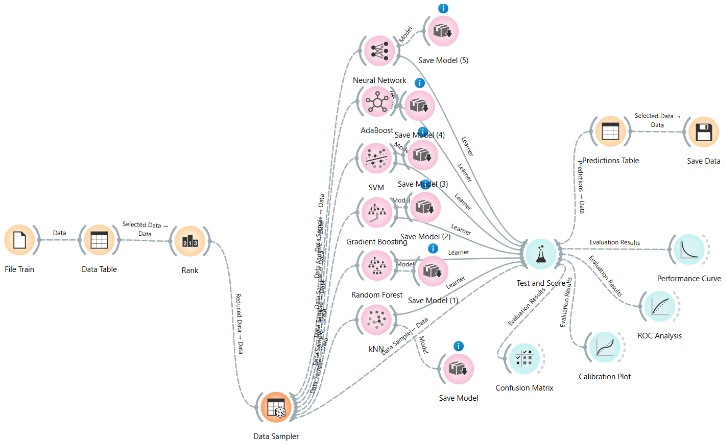


Fig. 2. Orange workflow for training and evaluation of Machine Learning models

The dataset was randomly split (70% for training, 30% for validation), with cases containing missing values excluded. Feature normalization to zero mean and unit variance was applied, and SMOTE oversampling (k=5) was used to balance class distributions.

As new structural and docking parameters generated from AlphaFold3 models and HADDOCK simulations became available, these features were incrementally incorporated into the workflow. However, due to the computational demands of generating AlphaFold3 predictions and docking results for all 3D structures, the full integration of these new parameters into the Plastizyme machine learning models is still in progress. Consequently, comprehensive retraining

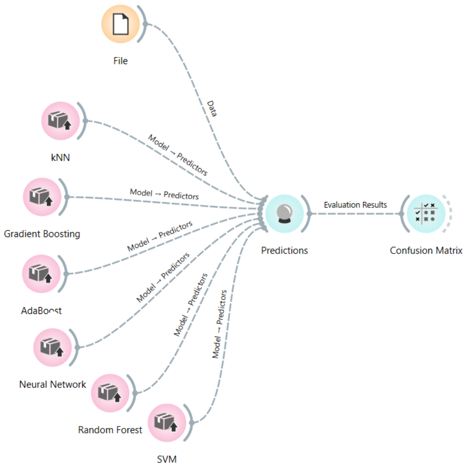


Fig. 3. Orange workflow for generating predictions using trained models

and evaluation of the models using the complete set of updated features will be performed once all AlphaFold3 and docking calculations are finalized.

For the current phase, feature selection was conducted by ranking features according to Pearson correlation with the target variable, removing descriptors with $|r| > 0.90$ to minimize multicollinearity. Constant features were excluded using a variance threshold of zero. Six classification algorithms—k-Nearest Neighbors (kNN), Random Forest, Gradient Boosting, Support Vector Machine (SVM), AdaBoost, and Neural Network—were trained using Orange's Test & Score widget with fivefold cross-validation. Hyperparameter grids were defined for each model to optimize performance, and the best-performing models were subsequently used to generate predictions on the validation dataset.

This staged approach demonstrates the ongoing integration of advanced structural prediction and molecular docking data into the Plastizyme machine learning framework, in line with recent advances in computational structural biology and protein engineering. Full model performance metrics will be reported once the expanded dataset is complete and the final models are retrained and validated.

Comparing Sampling Strategies: Balanced Random vs. Balanced Similar

The training dataset was balanced in this study using two different approaches: Balanced Random Sampling and Balanced Similar Sampling. Both strategies aimed to reduce class disparity while maintaining significant biological patterns in the data.

Balanced Random Sampling: Entails choosing an equal number of positive and negative examples at random from the dataset, disregarding feature similarity. Although this method guarantees class balance, there is a chance that it will include negative samples that are trivially simple to classify because they differ from the positive class.

Limitations:

- May select negatives that are easily distinguished, artificially simplifying the classification process.
- Less representative of borderline or ambiguous cases.
- May lead to model's overestimation performance.

Balanced Similar Sampling: Uses molecular descriptors to choose negative examples that are comparable to the positive class. Because the model has to learn to discern small variations in molecular patterns, this makes the classification task more difficult and realistic.

Advantages:

- Focuses training on the most borderline or ambiguous cases.
- Encourages the model to learn finer-grained class distinctions.
- Improves generalization, particularly in situations where class distinctions are ambiguous.

Model Evaluation Using Test & Score

The `Test & Score` widget from the Orange data mining platform was used to evaluate the performance of different machine learning classifiers trained on the balanced dataset. A stratified k-fold cross-validation was carried out by this algorithm, guaranteeing a reliable and objective assessment of each model's capacity for generalization.

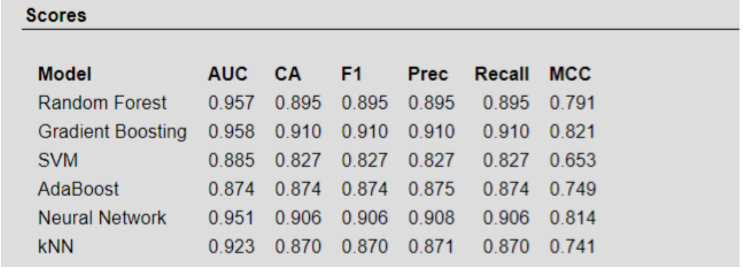
A number of performance metrics were used to assess each model such as: Area Under the ROC Curve (AUC), Classification Accuracy (CA), F1 Score, Precision, Recall, Matthews Correlation Coefficient (MCC).

Best Performing Model: Gradient Boosting

Gradient Boosting, an ensemble learning technique that builds multiple weak learners constructed one after the other, performed the best overall across all evaluation metrics out of all the models that were tested. As a result, a robust predictive model is produced that can manage intricate, non-linear relationships and lower bias and variance. Figure 4 displays the performance metrics for each algorithm, including the Gradient Boosting model.

Confusion Matrix Analysis

Figure 5 illustrates the confusion matrix that was analyzed in order to gain a better understanding of the Gradient Boosting model's classification behavior.

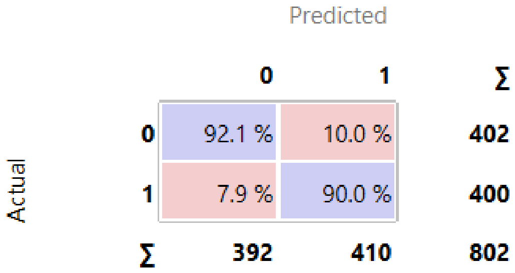


Scores

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.957	0.895	0.895	0.895	0.895	0.791
Gradient Boosting	0.958	0.910	0.910	0.910	0.910	0.821
SVM	0.885	0.827	0.827	0.827	0.827	0.653
AdaBoost	0.874	0.874	0.874	0.875	0.874	0.749
Neural Network	0.951	0.906	0.906	0.908	0.906	0.814
kNN	0.923	0.870	0.870	0.871	0.870	0.741

Fig. 4. Performance comparison of classification models using stratified 10-fold cross-validation. Metrics shown include AUC, CA, F1 Score, Precision, Recall, and MCC.

This matrix offers a thorough analysis of model performance at the class level by displaying the number of true positives, false negatives, false positives, and true negatives. The model's robustness and dependability on the balanced dataset are demonstrated by the high TP and TN values and comparatively low misclassification rates.



Predicted

Actual	0	1	∑
0	92.1 %	10.0 %	402
1	7.9 %	90.0 %	400
∑	392	410	802

Fig. 5. Confusion matrix of the classification model.

ROC Curve – Gradient Boosting

One popular tool for assessing how well classification models perform is the Receiver Operating Characteristic (ROC) curve. Plotting the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at different threshold settings demonstrates the diagnostic power of a binary classifier. The ROC curve for the Gradient Boosting classifier is displayed in Figure 6. The curve shows that even at low false positive rates, the model maintains a high sensitivity while achieving strong discriminative power. The robustness of the model is demonstrated by error bars, which show the variation in performance across various cross-validation folds.

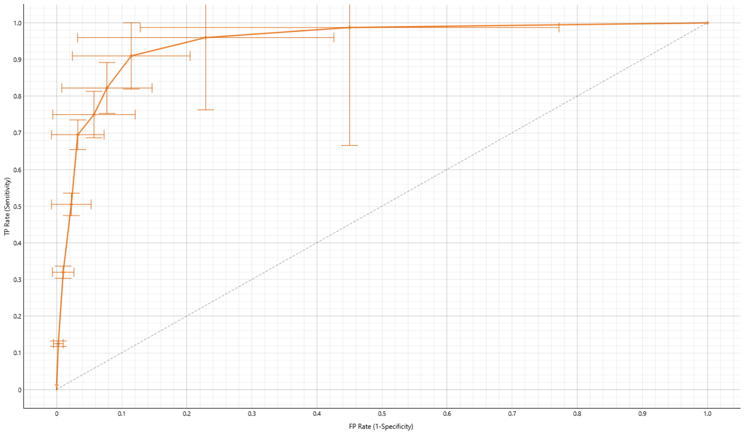


Fig. 6. ROC curve of the Gradient Boosting classifier with error bars representing variability across cross-validation folds.

5 Outlook and Future Directions

This work successfully established an integrated computational pipeline for the discovery and optimization of plastic-degrading enzymes. Key objectives were achieved by enhancing the Plastizyme database with high-quality sequence and structural data from major repositories (UniProt, NCBI Protein, PDB, AlphaFold), and implementing rigorous data standardization for robust machine learning applications. Advanced predictive models, leveraging structure-based features (AlphaFold3), were developed and validated, outperforming classical approaches in enzyme function prediction. Structure-function relationships were clarified through precise AlphaFold3 modeling and automated HADDOCK docking simulations, while the incorporation of docking parameters and binding affinity data into the machine learning workflow further improve predictive accuracy. The pipeline's performance was systematically validated.

Despite these advances, the full integration of new AlphaFold3 and docking-derived features is ongoing, with comprehensive retraining and validation of the models to follow upon completion of all 3D structure calculations. While the computational framework provides a powerful platform for enzyme discovery and engineering, empirical validation remains essential to confirm biological relevance. Future work should focus on expanding training datasets with experimentally validated enzyme activities, and on the experimental expression and testing of prioritized variants under realistic conditions.

Looking ahead, the modular and scalable architecture of this pipeline positions it for broader application in fields such as pharmaceutical protein design, biocatalysis, and the biodegradation of emerging pollutants. Ultimately, the convergence of computational and experimental workflows is expected to deliver accurate, scalable, and sustainable solutions for environmental biotechnology and beyond.

Bibliography

[1] Geyer, R., Jambeck, J. R., & Law, K. L. (2017). Production, use, and fate of all plastics ever made. Science Advances, 3(7), e1700782. https://doi.org/10.1126/sciadv.1700782

[2] Barnes, D. K. A., Galgani, F., Thompson, R. C., & Barlaz, M. (2009). Accumulation and fragmentation of plastic debris in global environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1526), 1985-1998. https://doi.org/10.1098/rstb.2008.0205

[3] Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., & Law, K. L. (2015). Plastic waste inputs from land into the ocean. Science, 347(6223), 768-771. https://doi.org/10.1126/science.1260352

[4] Lebreton, L., Slat, B., Ferrari, F., Sainte-Rose, B., Aitken, J., Marthouse, R., Hajbane, S., Cunsolo, S., Schwarz, A., Levivier, A., Noble, K., Debeljak, P., Maral, H., Schoeneich-Argent, R., Brambini, R., & Reisser, J. (2018). Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic. Scientific Reports, 8(1), 4666. https://doi.org/10.1038/s41598-018-22939-w

[5] Wilcox, C., Van Sebille, E., & Hardesty, B. D. (2015). Threat of plastic pollution to seabirds is global, pervasive, and increasing. Proceedings of the National Academy of Sciences, 112(38), 11899-11904. https://doi.org/10.1073/pnas.1502108112

[6] Bergmann, M., Mützel, S., Primpke, S., Tekman, M. B., Trachsel, J., & Gerdts, G. (2019). White and wonderful? Microplastics prevail in snow from the Alps to the Arctic. Science Advances, 5(8), eaax1157. https://doi.org/10.1126/sciadv.aax1157

[7] Leslie, H. A., van Velzen, M. J. M., Brandsma, S. H., Vethaak, A. D., Garcia Vallejo, J. J., & Lamoree, M. H. (2022). Discovery and quantification of plastic particle pollution in human blood. Environment International, 163, 107199. https://doi.org/10.1016/j.envint.2022.107199

[8] World Economic Forum. (2016). The new plastics economy: Rethinking the future of plastics. World Economic Forum. https://www.weforum.org/reports/the-new-plastics-economy-rethinking-the-future-of-plastics

[9] Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., Toyohara, K., Miyamoto, K., Kimura, Y., & Oda, K. (2016). A bacterium that degrades and assimilates poly(ethylene terephthalate). Science, 351(6278), 1196-1199. https://doi.org/10.1126/science.aad6359

[10] Carr, C. M., Clarke, D. J., & Dobson, A. D. W. (2020). Microbial polyethylene terephthalate hydrolases: current and future perspectives. Frontiers in Microbiology, 11, 571265. https://doi.org/10.3389/fmicb.2020.571265

[11] Danso, D., Schmeisser, C., & Streit, W. R. (2019). Plastic biodegradation: Biotechnology and microbiology. Applied and Environmental Microbiology, 85(16), e01095-19. https://doi.org/10.1128/AEM.01095-19

[12] Wei, R., Song, C., Gräsing, D., Schneider, T., Bielytskyi, P., Böttcher, D., Matysik, J., Bornscheuer, U. T., & Zimmermann, W. (2019). Conformational fitting of a flexible oligomeric substrate does not explain the enzymatic PET degradation. Nature Communications, 10(1), 5581. https://doi.org/10.1038/s41467-019-13492-9

[13] Shirke, A. N., White, C., Englaender, J. A., Zwarycz, A., Butterfoss, G. L., Linhardt, R. J., & Gross, R. A. (2018). Stabilizing leaf and branch compost cutinase (LCC) with glycosylation: mechanism and effect on PET hydrolysis. Biochemistry, 57(7), 1190-1200. https://doi.org/10.1021/acs.biochem.7b01189

[14] Taniguchi, I., Yoshida, S., Hiraga, K., Miyamoto, K., Kimura, Y., & Oda, K. (2019). Biodegradation of PET: Current status and application aspects. ACS Catalysis, 9(5), 4089-4105. https://doi.org/10.1021/acscatal.8b05171

[15] Palm, G. J., Reisky, L., Böttcher, D., Müller, H., Michels, E. A. P., Walczak, M. C., Berndt, L., Weiss, M. S., Bornscheuer, U. T., & Weber, G. (2019). Structure of the plastic-degrading Ideonella sakaiensis MHETase bound to a substrate. Nature Communications, 10(1), 1717. https://doi.org/10.1038/s41467-019-09326-3

[16] Joo, S., Cho, I. J., Seo, H., Son, H. F., Sagong, H. Y., Shin, T. J., Choi, S. Y., Lee, S. Y., & Kim, K. J. (2018). Structural insight into molecular mechanism of poly(ethylene terephthalate) degradation. Nature Communications, 9(1), 382. https://doi.org/10.1038/s41467-018-02881-1

[17] Tournier, V., Topham, C. M., Gilles, A., David, B., Folgoas, C., Moya-Leclair, E., Kamionka, E., Desrousseaux, M. L., Texier, H., Gavalda, S., Cot, M., Guémard, E., Dalibey, M., Nomme, J., Cioci, G., Barbe, S., Chateau, M., André, I., Duquesne, S., & Marty, A. (2020). An engineered PET depolymerase to break down and recycle plastic bottles. Nature, 580(7802), 216-219. https://doi.org/10.1038/s41586-020-2149-4

[18] Salvador, M., Abdulmutalib, U., Gonzalez, J., Kim, J., Smith, A. A., Faulon, J. L., Wei, R., Zimmermann, W., & Jimenez, J. I. (2019). Microbial genes for a circular and sustainable Bio-PET economy. Genes, 10(5), 373. https://doi.org/10.3390/genes10050373

[19] Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. Nature Methods, 16(8), 687-694. https://doi.org/10.1038/s41592-019-0496-6

[20] Xu, Y., Verma, D., Sheridan, R. P., Liaw, A., Ma, J., Marshall, N. M., McIntosh, J., Sherer, E. C., Svetnik, V., & Voigt, J. H. (2021). Deep dive into machine learning models for protein engineering. Journal of Chemical Information and Modeling, 60(6), 2773-2790. https://doi.org/10.1021/acs.jcim.0c00073

[21] Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J., & Arnold, F. H. (2021). Machine learning-assisted directed protein evolution with combina-

torial libraries. Proceedings of the National Academy of Sciences, 118(10), e2012588118. https://doi.org/10.1073/pnas.2012588118

[22] Fox, R. J., Davis, S. C., Mundorff, E. C., Newman, L. M., Gavrilovic, V., Ma, S. K., Chung, L. M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Whitman, J. C., Sheldon, R. A., & Huisman, G. W. (2007). Improving catalytic function by ProSAR-driven enzyme evolution. Nature Biotechnology, 25(3), 338-344. https://doi.org/10.1038/nbt1286

[23] Alley, E. C., Gimpel, J. A., Davies, J. H., & Colwell, L. J. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nature Methods, 16(12), 1315-1322. https://doi.org/10.1038/s41592-019-0598-1

[24] Amidi, A., Amidi, S., Vlachakis, D., Megalooikonomou, V., Paragios, N., & Zacharaki, E. I. (2022). EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ, 10, e13380. https://doi.org/10.7717/peerj.4750

[25] Chen, Z., Wang, Z., Ren, J., Li, G., & Jiang, F. (2021). Machine learning-guided directed evolution for protein engineering. ACS Synthetic Biology, 10(12), 3141-3159. https://doi.org/10.1021/acssynbio.1c00297

[26] Mazurenko, S., Prokop, Z., & Damborsky, J. (2020). Machine learning in enzyme engineering. ACS Catalysis, 10(2), 1210-1223. https://doi.org/10.1021/acscatal.9b04321

[27] Kurinov, A., Gaetani, M., Mondal, A., & Greener, J. G. (2023). Advances and limitations of protein engineering with machine learning. Nature Communications, 14(1), 1-12. https://doi.org/10.1038/s41467-023-39796-5