Environmental Science & Technology Letters

pubs.acs.org/journal/estlcu

Letter

Machine Learning Based Prediction of Enzymatic Degradation of Plastics Using Encoded Protein Sequence and Effective Feature Representation

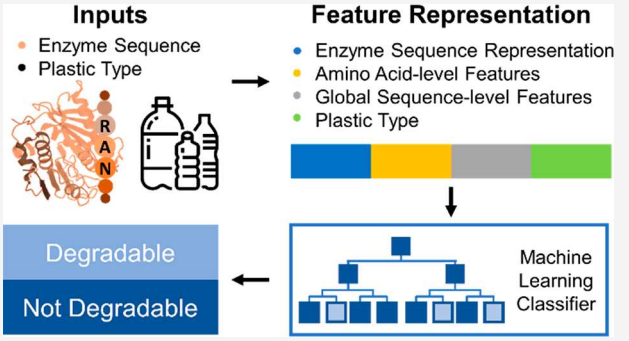
Renjing Jiang,[§] Lanyu Shang,[§] Ruohan Wang, Dong Wang,* and Na Wei*

Cite This: *Environ. Sci. Technol. Lett.* 2023, 10, 557−564

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Enzyme biocatalysis for plastic treatment and recycling is an emerging field of growing interest. However, it is challenging and time-consuming to identify plastic-degrading enzymes with desirable functionality, given the large number of putative enzyme sequences. There is a critical need to develop an effective approach to accurately predict the enzyme activity in degrading different types of plastics. In this study, we developed a machine-learning-based plastic enzymatic degradation (PED) framework to predict the ability of an enzyme to degrade plastics of interest by exploring and recognizing hidden patterns in protein sequences. A data set integrating information from a wide range of experimentally verified enzymes and various common plastic substrates was created. A new context-aware enzyme sequence representation (CESR) mechanism was developed to learn the abundant contextual information in enzyme sequences, and feature extraction was performed for enzymes at both the amino acid level and global sequence level. Thirteen machine learning classification algorithms were compared, and XGBoost was identified as the best-performing algorithm. PED achieved an overall accuracy of 90.2% and outperformed sequence-based protein classification models from the existing literature. Furthermore, important enzyme features in plastic degradation were identified and comprehensively interpreted. This study demonstrated a new tool for the prediction and discovery of plastic-degrading enzymes.



KEYWORDS: *Machine learning, plastic waste, enzymatic degradation, enzyme function, sequence representation*

INTRODUCTION

Plastics are extensively used globally,[1] but improper handling of plastic waste has caused severe environmental problems.[2−4] Treating and recycling postconsumer plastics is critically important for environmental protection and waste valorization.[5] Conventional mechanical and chemical recycling either lead to loss of plastic properties or require high energy input and expensive reagents.[6,7] In contrast, biological enzymes, with high efficiency under mild reaction conditions, degrade plastics into monomers which can then be recovered to synthesize new plastic products for achieving a circular economy.[8−10]

Significant research progress has been made recently in discovering plastic-degrading enzymes,[11] and there is a growing interest in identifying new plastic-degrading enzymes with desirable functionalities by exploring the ever-increasing number of putative enzyme sequences.[12,13] However, searching for plastic-degrading enzymes is a challenging task as evidenced by prior research efforts.[9] First, the enzyme plastic degradation capabilities do not correlate well with the enzyme commissioning (EC) families. For example, a lipase (EC 3.1.1.3) PbsA was active in degrading poly(butylene succinate) (PBS) plastic in one study,[14] while another lipase PETase degraded poly(ethylene furanoate) (PEF) but not PBS.[15] Meanwhile, enzymes degrading the same type of plastic may belong to different EC families (Table S1), making it hard to simply infer degradability based on taxonomy. Second, prior efforts in exploring plastic-degrading enzymes mostly used homology searching based on sequence similarity,[16] but sequence similarity does not always correlate with enzyme plastic-degrading functionality, and thus, homology searching could overlook new plastic-degrading enzymes or lead to incorrect predictions.[16,17] For example, the newly discovered PET degrading enzyme PETase from *Ideonella sakaiensis* 201-F6 shared only 51% sequence similarity with a previously known hydrolase TfH from *Thermobifida fusca* during homology searching.[18] Therefore, there is a critical need to develop an effective method that is less dependent on sequence

Received: May 7, 2023
Revised: June 9, 2023
Accepted: June 12, 2023
Published: June 19, 2023

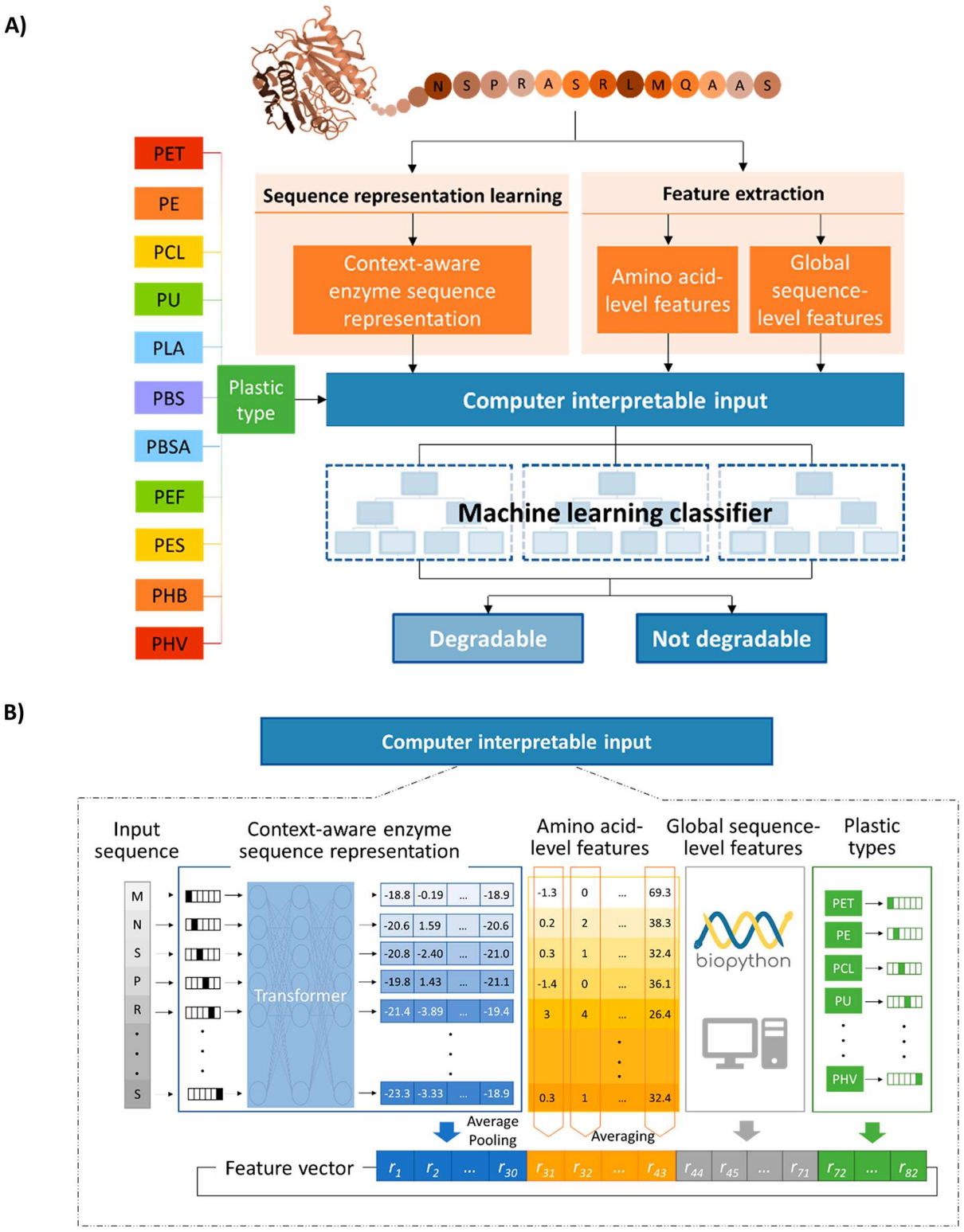


Figure 1. Overview of the plastic enzymatic degradation (PED) framework. (A) Inputs, outputs, and key components of PED. Amino acid sequence of enzyme noted as a string of alphabet letters was converted into computer interpretable input by sequence representation learning, in particular, the context-aware enzyme sequence representation (CESR) method. Enzyme features at amino acid- and global sequence-levels were extracted from raw sequence and used as additional inputs to a machine learning (ML) classifier. The classifier, namely, XGBoost, performed the binary classification by predicting whether the enzyme can degrade the plastic of interest. (B) Construction of the feature vector used as the computer interpretable input in the PED framework. A transformer-based attention mechanism[46] was adopted in CESR learning and outputted a 30-length vector to represent a given sequence. By concatenating CESR, enzyme features at amino acid- and global sequence-levels with the one-hot encoded plastic types, a feature vector was obtained and further used as the input vector of the ML classifier in the PED framework for degradation classification. r_i: feature used as input of the ML classifier (i = 1, 2, ...82).

https://doi.org/10.1021/acs.estlett.3c00293
Environ. Sci. Technol. Lett. 2023, 10, 557−564

similarity and taxonomy to accurately predict enzyme activities in plastic degradation.

Computational methods, particularly machine learning (ML), start to receive increasing attention in recent research efforts.[9] ML is a systematic computational analysis that is capable of capturing hidden patterns from a massive amount of data to make predictions or decisions.[19−21] ML has been applied to protein function prediction using amino acid sequence data, which stores all the information for protein folding and functioning.[22−24] Recent studies have reported prediction of plastic enzymatic degradation using protein sequence information,[25−27] but critical limitations exist. First, the data lack the reliability for model validation. Most studies collected enzyme sequences from UniProt,[25,26] but over 95% of the enzyme sequences in UniProt are from sources without experimental reports of enzymatic activities.[28] Another limitation in previous studies is that the range of plastic substrates considered was relatively narrow. For example, Hemalatha et al.'s study only predicted whether a protein could degraded alkane[26] and Buchholz et al.'s study only identified homologues of PET hydrolases.[27] Gan and Zhang's study included polyhydroxyalkanoate, polyhydroxybutyrate (PHB), polyurethane (PU), poly(vinyl alcohol) and phthalate, but the model did not consider enzymes that were active on multiple plastic substrates.[25] Therefore, a model for reliably predicting enzymatic degradation of different plastic substrates is needed to address the knowledge gap in the current literature.

In this study, we aimed to develop an innovative and effective computational approach to predict the ability of an enzyme with a known sequence to degrade a target plastic of interest with consideration of a variety of common plastics. A ML-based plastic enzymatic degradation (PED) framework was designed, integrating information from a wide range of experimentally verified enzymes and different types of plastic substrates collected from peer-reviewed publications (Figure 1). A new context-aware enzyme sequence representation (CESR) learning mechanism was developed to learn the abundant contextual information in enzyme sequences, feature extraction was performed for enzymes at both amino acid and global sequence-levels and were compared, and XGBoost was selected from 13 ML classification algorithms. Model evaluation results demonstrated that PED significantly outperformed state-of-the-art sequence-based protein classification models. Furthermore, we comprehensively analyzed PED prediction results to understand important features for enzymatic plastic degradation.

2. METHODS

2.1. Data Set Preparation. Information about the enzymatic degradation of plastics was manually collected from relevant peer-reviewed publications between 1995 and 2022. Specific information included enzyme sequences, plastic types, and ground-truth labels corresponding to an enzyme-plastic pair (i.e., degradable or nondegradable) based on experimental studies reported. In all, a data set including 213 records of enzyme-plastic pairs was created. Details of data set preparation were provided in Section S1.

2.2. Context-Aware Enzyme Sequence Representation (CESR) and Feature Extraction. Enzyme sequences are alphabetical letters representing amino acids. In order to convert text of letters into numerical matrices, one-hot encoding[29] was used and each enzyme sequence was represented by a matrix consisting of zeros and ones (detailed in Section S2). Each of the 11 plastic types was one-hot encoded into an 11-bit vector consisting of ten zeros and a one. A CESR learning mechanism was designed to capture latent contextual information on amino acid sequences (Section S3).

By using feature correlation analysis (Figure S1) and domain knowledge (detailed in Section S4), 13 amino acid-level features (Table S2) and 28 global sequence-level features (Table S3) were extracted from sequential information, which characterized the physicochemical properties of amino acid residues and the whole enzyme sequence, respectively.

Feature importance was evaluated using Shapley Additive exPlanations (SHAP).[30] Mean SHAP values were obtained to describe the average impact of the feature on model output, and a scatter plot of SHAP values was obtained to show the influence of a feature value on model output. Details are provided in Section S5.

2.3. Classification Model Development and Optimization. A base model was first developed using one-hot encoded enzyme sequences and plastic types as inputs. Then, the model incorporated enzyme sequence information learned by the CESR mechanism and the extracted features at amino acid- and global sequence-levels features. CESR was concatenated with extracted features and the encoded plastic type as the final feature vector, which was used as the input to a supervised ML classification algorithm to perform the binary classification (Figure 1B).

In the process of model optimization, the performance of 13 classification algorithms from seven algorithm categories were evaluated (Section S6) and the best algorithm was chosen in the PED framework.

The data set was randomly split into training and testing sets with a ratio of 8:2. The random split was repeated ten times, and model performance results were averaged. Model performance was evaluated by a set of widely adopted metrics, including accuracy, precision, recall, and F1 score (Section S6).[31−33] The receiver operating characteristic (ROC) curve and the Area Under the ROC Curve (AUC_{ROC}) were used to evaluate the model performance against different classification thresholds.[34] 5-fold cross-validation was performed on the training set to tune the hyperparameters. In the analysis of model performance on dissimilar train-test sets, the data set was split by using *k*-medoids clustering method,[35] a popular clustering method that groups similar enzyme sequences together. Details are provided in Section S6.

3. RESULTS AND DISCUSSION

3.1. Data Set of Enzymatic Plastic Degradation. Based on a thorough review of literature reports with experimental studies, we organized information regarding enzymatic degradation of plastics including PET, PU, PHB, PBS, PEF, polyethylene (PE), polycaprolactone (PCL), poly(ethylene succinate) (PES), polyhydroxyvalerate (PHV), poly(lactic acid) (PLA), and poly(butylene succinate-*co*-adipate) (PBSA). The data set included 230 enzyme-plastic pairs in total, consisting of 129 unique enzyme sequences and 11 types of plastics. 141 enzyme-plastic pairs were degradable, and 89 pairs were nondegradable. Among all enzymes, PET-degrading enzymes and PHB-degrading enzymes have been frequently reported, with 63 and 40 enzymes tested (Figure S2). As for the other plastics, the specific enzymes contributing to plastic degradation remain underexplored[8,36] and thus constituted a small proportion of the data set. Additionally, pairwise

https://doi.org/10.1021/acs.estlett.3c00293
Environ. Sci. Technol. Lett. 2023, 10, 557−564

sequence similarity (with calculation described in Section S1) showed 78% pairs of all data points had sequence similarities less than 30% and that 7% pairs had similarities larger than 90%, indicating enzyme sequences in the data set were dissimilar (Figure S3). The data set is publicly available in Supporting Information.

3.2. Role of the Key Model Components. The contribution of each key model component, including CESR and feature extraction at amino acid and global sequence-levels, was analyzed in the process of rational design of the overall PED framework (Section S7). We first built a base model where one-hot encoding of amino acids was used to represent enzyme sequences and random forest (RF)[22,37] was used as the default classification algorithm. The base model performed reasonably well (with an overall accuracy of 75.5%) but could be further improved (Table S4 and Section S7). Therefore, we developed and introduced the CESR learning strategy in lieu of one-hot encoding and generated the CESR model. While one-hot encoding has been frequently used to represent protein sequences, this method is inherently sparse, memory-inefficient, and high-dimensional, as it only differentiates the 20 amino acids but ignores the interactions of adjacent amino acids in the surrounding microenvironments.[29,38] In contrast to the uniform amino acid representation by one-hot encoding, the CESR mechanism not only learned the latent vector representation of each amino acid but also used the state-of-the-art bidirectional transformer-based attention mechanism[39] to jointly capture the amino acid information from the forward and backward directions of an enzyme sequence and explore the interactive information from the adjacent amino acids of each specific amino acid. The CESR model outperformed the base model in accuracy by 2.72% and recall by 3.31%, suggesting the benefits of CESR by learning contextual information in enzyme sequences (Table S4 and Section S7).

Next, we extracted enzyme features at amino acid and global sequence-levels as additional informative inputs. The amino acid-level features capture the physicochemical properties of individual amino acid residues within a protein and are of importance in the development of predictive models for protein classification.[24,40−42] The global sequence-levels features capture the characteristics of a whole sequence and have been shown to help solve protein classification problems.[43] When amino acid- and global sequence-level features were incorporated into the CESR model, the new C/AA/GS model outperformed the CESR model in all evaluation metrics (Table S4 and Section S7). In summary, we successfully developed and implemented CESR and feature extraction at amino acid- and global sequence-levels for model optimization and each optimization step was effective.

3.3. Evaluation of ML Algorithms and Generalizability of PED Framework. To select a proper ML algorithm,[44] we investigated the performance of different ML classification algorithms on the basis of the C/AA/GS model and identified XGBoost as the best-performing algorithm for the final PED framework (Table S5). Detailed analysis of the strengths of XGBoost over other classifiers is provided in Section S8. As the performance of ML algorithms is often sensitive to the amount of training data,[45] we also evaluated the impact of the size of training data set on PED performance (Figure S4). PED worked well with data sets of small sizes, and the number of enzyme-plastic pairs in our data set was sufficient to learn the classification problem in this study, though further improvement could be achieved with the expanded data sets as the research area advances (detailed in Section S8). To assess the generalizability of PED, we performed a train-test split by using the *k*-medoids clustering method (Section S6) and found that PED was generalizable to enzyme sequences that were dissimilar to those in the training data set (Table S6, Section S8). Additionally, to facilitate the utilization of PED by potential users on a plastic of interest, we presented the accuracy breakdown of PED by plastic type (Table S7). The PED achieved accuracies higher than 86.6% for all plastic types, except for PHV which had an accuracy of 65.0% due to the small number of training data. More discussion can be found in Section S8.

3.4. Performance of PED Compared to Existing Sequence-Based ML Models. The performance of PED was compared to sequence-based protein classification models from existing literature (referred to as "baseline models", with a description of these models summarized in Table S8). The same data set was used as the input to all the models for evaluation. PED outperformed all of the baseline models (Figure 2A, Table S9). Particularly, PED outperformed the best-performing baseline model Seq2Vec[29] by 21% in terms of F1 score, which collectively represents precision and recall and is commonly used to compare the performance of different classifiers.[32] In the ROC curve, which is a two-dimensional depiction of classification performance in terms of the true positive rate on the *y*-axis and the false positive rate on the *x*-axis,[34] PED had the curve closest to the top-left region, meaning the best performance (Figure 2B). The highest AUC_{ROC} of 0.89 showed a reliable resolution of PED to distinguish degradable versus nondegradable enzyme-plastic pairs.

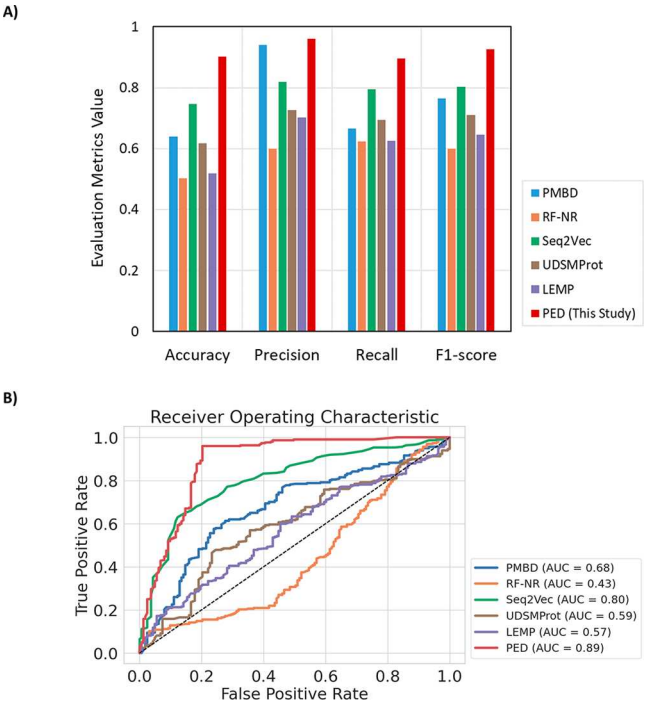


Figure 2. Comparison of the performance of PED with sequence-based protein classification models from existing literature in terms of (A) evaluation metrics (accuracy, precision, recall, and F1-score), and (B) Receiver Operating Characteristic (ROC) curve. AUC: area under the ROC curve.

https://doi.org/10.1021/acs.estlett.3c00293
Environ. Sci. Technol. Lett. 2023, 10, 557−564

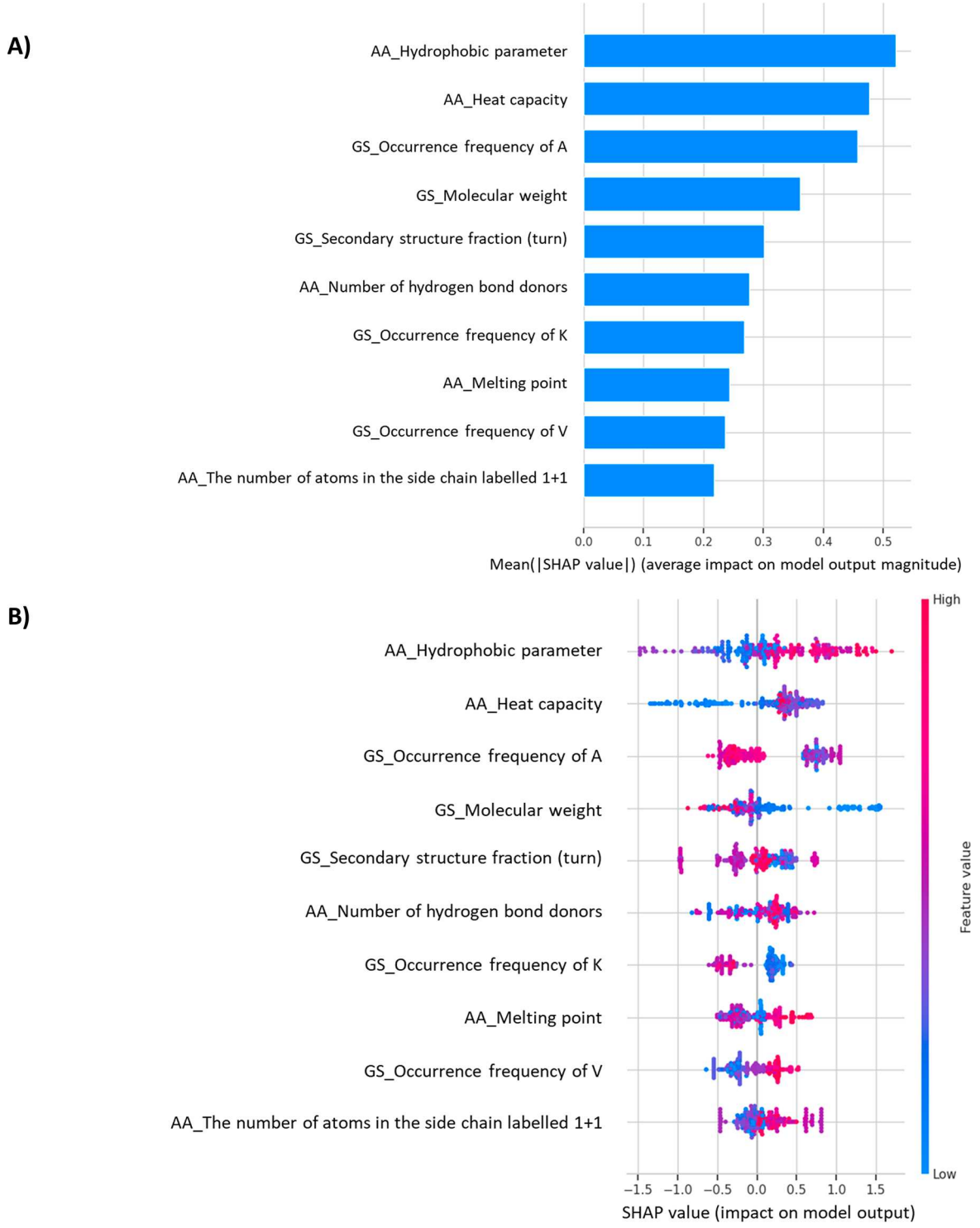


Figure 3. Ranking of top features based on SHapley Additive exPlanations (SHAP). (A) Top ten most important features in enzymatic degradation of plastics. The *x*-axis is the mean absolute SHAP value, which indicates the relative importance of features on enzymatic plastic degradation. (B) The density scatter plot of SHAP values indicating the impact of feature values on degradation. A positive SHAP value means that the feature can increase the predicted possibility of degradation (i.e., more likely to be degradable), whereas a negative SHAP value means that the feature can decrease the predicted possibility of degradation (i.e., less likely to be degradable). AA: amino acid-level features; GS: global sequence-level features.

The superior performance of PED compared to the baseline models could be attributed to three aspects of the model design. First, PED had the unique CESR learning mechanism encoding not only the individual amino acids but also the contextual relationships between neighborhood amino acids, which was distinct from the protein representation strategies (e.g., long short-term memory,[46] Doc2Vec mechanism[29]) in baseline models. Second, PED extracted enzyme features at both amino acid- and global sequence-levels, which brought out key protein information to distinguish the degradable and

https://doi.org/10.1021/acs.estlett.3c00293
Environ. Sci. Technol. Lett. 2023, 10, 557−564

nondegradable enzyme-plastic pairs. In comparison, the baseline models either lacked such comprehensive feature extraction or extracted nonrepresentative features. Third, PED employed the XGBoost classifier, of which the advantage was discussed in Section S8. In comparison, the deep learning-based baseline models underperformed because deep learning classifiers could not achieve desirable performance where the size of the training data set for enzymatic degradation of plastics was relatively small. Other baseline models employed RF classifier, which was outcompeted by XGBoost used in PED since XGBoost classifier was optimized by paying more attention to misclassified data samples to boost the overall classification performance.[47,48] A detailed comparison of PED with the baseline models regarding the three aspects was provided in Section S9.

3.5. Feature Interpretation. The effect of amino acid- and global sequence-level features on degradation prediction was analyzed by calculating SHAP values (Section S5), and the top ten ranked features were identified (Figure 3). First, high amino acid hydrophobicity levels (i.e., low feature value of *AA_Hydrophobic parameter*, defined as required energy (kcal/mol) of transferring an amino acid from water to ethanol at 25 °C[49]) could in general negatively affect enzyme activity in plastic degradation (Figure 3B). The influence of enzyme hydrophobicity could be complex, as shown in previous studies. On one hand, high hydrophobicity on enzyme surface can facilitate the attachment of enzyme onto plastic substrate and thus promote degradation, with an example in Section S10.[50] On the other hand, high hydrophobicity can negatively affect enzymatic plastic degradation due to enzyme aggregation and impairment of catalytic activity caused by intermolecular hydrophobic interactions.[51] Our SHAP analysis showed that the negative effects of hydrophobicity were more profound in the enzymes analyzed. Second, a positive effect of high *AA_Heat capacity* (referring to the amount of heat to be supplied to one molar amino acid to produce a unit change in their temperature (cal/mol-°C)) was observed (Figure 3B). A protein with relatively high heat capacity indicates the protein would be resistant to temperature change and denaturation,[52−54] but it remains an open question how heat capacity correlates with the protein functioning of plastic-degrading enzymes and further study is needed. Third, there was a clear distribution of SHAP values for the feature *GS_molecular weight*, where enzymes with a low molecular weight were favorable for plastic degradation compared to larger enzymes (Figure 3B). The observation was consistent with the literature reports that small enzyme molecules could diffuse into amorphous regions or pores in crystalline regions in polymers relatively easily and thus catalyzed degradation process more efficiently than large enzymes.[55,56] Last but not least, it was observed that a lower frequency of alanine (A) in protein sequences was favorable for the enzymatic degradation of plastics (Figure 3B). Such effects may be attributed to the aromatic interactions between amino acids and plastic substrates[57] and was observed in a study detailed in Section S10.[15] Additional discussion for some other features (e.g., number of hydrogen bond donors and turn structures) is detailed in Section S10.

3.6. Implications. In summary, this study demonstrated the successful application of ML for sequence-based prediction of enzyme activity in degrading various plastics for the first time. PED can simply use an enzyme sequence and plastic type as the input query to generate the output of *degradable/nondegradable*, providing a new data-driven model for enzymatic plastic degradation prediction. Furthermore, critical enzyme features at amino acid and global sequence-levels were identified, which provided insight into critical factors affecting enzyme activities in plastic degradation at the molecular level and suggested directions for future experimental investigation. It is noted that the size of the data set for model training is relatively small, and we envision the data available from experimental studies will expand rapidly as research in this emerging area advances. Our future work will jointly explore the data from both reported experimental studies and databases via weak supervision[58] to enable the use of unreliable and noisy data for creating a strong predictive model.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.3c00293.

Additional details on data set preparation, enzyme representation methods, feature selection and importance analysis, model development and evaluation methods, interpretation of model evaluation results, comparison of ML algorithms, analysis of model generalizability, interpretation of critical enzyme features, and supplementary figures and tables (PDF)

Python codes for the machine learning models used in this study (ZIP)

Complete data set used in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Na Wei − *Department of Civil and Environmental Engineering, University of Illinois Urbana−Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0003-2093-3441; Phone: 217-333-6967; Email: nawei2@illinois.edu

Dong Wang − *School of Information Sciences, University of Illinois Urbana−Champaign, Champaign, Illinois 61820, United States*; Phone: 217-244-6412; Email: dwang24@illinois.edu

Authors

Renjing Jiang − *Department of Civil and Environmental Engineering, University of Illinois Urbana−Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-7267-5520

Lanyu Shang − *School of Information Sciences, University of Illinois Urbana−Champaign, Champaign, Illinois 61820, United States*

Ruohan Wang − *Department of Civil and Environmental Engineering, University of Illinois Urbana−Champaign, Urbana, Illinois 61801, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.estlett.3c00293

Author Contributions

[§](R.J., L.S.) These authors contributed equally

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acs.estlett.3c00293
Environ. Sci. Technol. Lett. 2023, 10, 557−564

■ ACKNOWLEDGMENTS

This work was supported by capital funds from the University of Illinois Urbana−Champaign (UIUC), the National Science Foundation (Grant CNS-1831669), and CEE fellowship to R.J.

■ REFERENCES

(1) Matjasic, T.; Simcic, T.; Medvescek, N.; Bajt, O.; Dreo, T.; Mori, N. Critical evaluation of biodegradation studies on synthetic plastics through a systematic literature review. *Sci. Total Environ.* **2021**, *752*, 141959.
(2) Chang, X.; Xue, Y.; Li, J.; Zou, L.; Tang, M. Potential health impact of environmental micro- and nanoplastics pollution. *J. Appl. Toxicol.* **2020**, *40* (1), 4−15.
(3) Rochman, C. M.; Browne, M. A.; Halpern, B. S.; Hentschel, B. T.; Hoh, E.; Karapanagioti, H. K.; Rios-Mendoza, L. M.; Takada, H.; Teh, S.; Thompson, R. C. Classify plastic waste as hazardous. *Nature* **2013**, *494* (7436), 169−171.
(4) Smith, M.; Love, D. C.; Rochman, C. M.; Neff, R. A. Microplastics in seafood and the implications for human health. *Curr. Environ. Health Rep.* **2018**, *5* (3), 375−386.
(5) Wei, R.; Zimmermann, W. Biocatalysis as a green route for recycling the recalcitrant plastic polyethylene terephthalate. *Microb. Biotechnol.* **2017**, *10* (6), 1302−1307.
(6) Ragaert, K.; Delva, L.; Van Geem, K. Mechanical and chemical recycling of solid plastic waste. *Waste Manage.* **2017**, *69*, 24−58.
(7) Maraveas, C.; Kotzabasaki, M. I.; Bartzanas, T. Intelligent Technologies, Enzyme-Embedded and Microbial Degradation of Agricultural Plastics. *AgriEngineering* **2023**, *5* (1), 85−111.
(8) Chen, C.-C.; Dai, L.; Ma, L.; Guo, R.-T. Enzymatic degradation of plant biomass and synthetic polymers. *Nat. Rev. Chem.* **2020**, *4* (3), 114−126.
(9) Zhu, B.; Wang, D.; Wei, N. Enzyme discovery and engineering for sustainable plastic recycling. *Trends in Biotechnol.* **2022**, *40* (1), 22−37.
(10) Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **2022**, *604* (7907), 662−667.
(11) Roohi; Bano, K.; Kuddus, M.; Zaheer, R. M.; Zia, Q.; Khan, F. M.; Ashraf, M. G.; Gupta, A.; Aliev, G. Microbial enzymatic degradation of biodegradable plastics. *Curr. Pharm. Biotechnol.* **2017**, *18* (5), 429−440.
(12) Mardis, E. R. DNA sequencing technologies: 2006−2016. *Nat. Protoc.* **2017**, *12* (2), 213−218.
(13) Sankara Subramanian, S. H.; Balachandran, K. R. S.; Rangamaran, V. R.; Gopal, D. RemeDB: Tool for rapid prediction of enzymes involved in bioremediation from high-throughput metagenome data sets. *Comput. Biol.* **2020**, *27* (7), 1020−1029.
(14) Uchida, H.; Shigeno-Akutsu, Y.; Nomura, N.; Nakahara, T.; Nakajima-Kambe, T. Cloning and sequence analysis of poly-(tetramethylene succinate) depolymerase from *Acidovorax delafieldii* strain BS-3. *J. Biosci. Bioeng.* **2002**, *93* (2), 245−247.
(15) Austin, H. P.; Allen, M. D.; Donohoe, B. S.; Rorrer, N. A.; Kearns, F. L.; Silveira, R. L.; Pollard, B. C.; Dominick, G.; Duman, R.; El Omari, K.; Mykhaylyk, V.; Wagner, A.; Michener, W. E.; Amore, A.; Skaf, M. S.; Crowley, M. F.; Thorne, A. W.; Johnson, C. W.; Woodcock, H. L.; McGeehan, J. E.; Beckham, G. T. Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (19), E4350−E4357.
(16) Pearson, W. R. An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinform.* **2013**, *42* (1), 3.1.1−3.1.8.
(17) Viljakainen, V. R.; Hug, L. A. New approaches for the characterization of plastic-associated microbial communities and the discovery of plastic-degrading microorganisms and enzymes. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 6191−6200.
(18) Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.; Toyohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K. A bacterium that degrades and assimilates poly(ethylene terephthalate). *Science* **2016**, *351* (6278), 1196−1199.
(19) Mitchell, T. M. *Machine Learning*; McGraw-Hill Education: 1997; Vol. *1*.
(20) Liu, X.; Lu, D.; Zhang, A.; Liu, Q.; Jiang, G. Data-driven machine learning in environmental pollution: Gains and problems. *Environ. Sci. Technol.* **2022**, *56* (4), 2124−2133.
(21) Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55* (10), 78−87.
(22) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **2020**, *10* (2), 1210−1223.
(23) Kai, M. *Essential Cell Biology*, 4th ed.; Yale Journal of Biology and Medicine, 2015.
(24) Xu, Y.; Verma, D.; Sheridan, R. P.; Liaw, A.; Ma, J.; Marshall, N. M.; McIntosh, J.; Sherer, E. C.; Svetnik, V.; Johnston, J. M. Deep dive into machine learning models for protein engineering. *J. Chem. Inf. Model.* **2020**, *60* (6), 2773−2790.
(25) Gan, Z.; Zhang, H. PMBD: A comprehensive plastics microbial biodegradation database. *Database* **2019**, *2019*, baz119.
(26) Hemalatha, N.; Wilson, A.; Akhil, T. Prediction of plastic degrading microbes. *bioRxiv*, Aug. 2, 2021. DOI: 10.1101/2021.08.01.454681
(27) Buchholz, P. C. F.; Feuerriegel, G.; Zhang, H.; Perez-Garcia, P.; Nover, L.-L.; Chow, J.; Streit, W. R.; Pleiss, J. Plastics degradation by hydrolytic enzymes: The plastics-active enzymes database—PAZy. *Proteins: Struct. Funct. Genet.* **2022**, *90* (7), 1443−1456.
(28) *Where do the UniProtKB protein sequences come from?* https://www.uniprot.org/help/sequence_origin (accessed Dec. 22, 2022).
(29) Yang, K. K.; Wu, Z.; Bedbrook, C. N.; Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **2018**, *34* (23), 4138−4138.
(30) Lundberg, S. M.; Lee, S.-I., A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Long Beach, California, USA, 2017; pp 4768−4777.
(31) Olson, D. L.; Delen, D. *Advanced data mining techniques*. Springer Science & Business Media: 2008.
(32) Sasaki, Y. The truth of the F-measure. *Teach Tutor Mater.* **2007**, *1* (5), 1−5 Available from https://www.cs.odu.edu/~mukka/cs795sum11dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf (accessed Jan. 18, 2023)..
(33) Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20* (1), 37−46.
(34) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143* (1), 29−36.
(35) Park, H.-S.; Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert systems with applications* **2009**, *36* (2), 3336−3341.
(36) Wei, R.; Zimmermann, W. Microbial enzymes for the recycling of recalcitrant petroleum-based plastics: How far are we? *Microb. Biotechnol.* **2017**, *10* (6), 1308−1322.
(37) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(38) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16* (8), 687−694.
(39) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*; Association for Computational Linguistics: Minneapolis, Minnesota, USA, 2019; pp 4171−4186.
(40) Herrera-Bravo, J.; Farías, J. G.; Contreras, F. P.; Herrera-Belén, L.; Norambuena, J.-A.; Beltrán, J. F. VirVACPRED: A web server for prediction of protective viral antigens. *Int. J. Pept. Res. Ther.* **2022**, *28* (1), 35.
(41) Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A sequence-based predictor using effective feature representation to

improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34* (23), 4007−4016.
(42) Yerukala Sathipati, S.; Ho, S.-Y. Identification and characterization of species-specific severe acute respiratory syndrome coronavirus 2 physicochemical properties. *J. Proteome Res.* **2021**, *20* (5), 2942−2952.
(43) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16* (12), 1315−1322.
(44) Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2021**, *2* (3), 160.
(45) Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLoS One* **2019**, *14* (11), No. e0224365.
(46) Strodthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal deep sequence models for protein classification. *Bioinformatics* **2020**, *36* (8), 2401−2409.
(47) Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785−794.
(48) Nielsen, D. Tree boosting with xgboost: Why does xgboost win "every" machine learning competition? Ph.D. Thesis, Norweigian University of Science and Technology, Trondheim, Norway, 2016.
(49) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **1976**, *104* (1), 59−107.
(50) Hiraishi, T.; Komiya, N.; Maeda, M. Y443F mutation in the substrate-binding domain of extracellular PHB depolymerase enhances its PHB adsorption and disruption abilities. *Polym. Degrad. Stab.* **2010**, *95* (8), 1370−1374.
(51) Biundo, A.; Steinkellner, G.; Gruber, K.; Spreitzhofer, T.; Ribitsch, D.; Guebitz, G. M. Engineering of the zinc-binding domain of an esterase from *Clostridium botulinum* towards increased activity on polyesters. *Catal. Sci. Technol.* **2017**, *7* (6), 1440−1447.
(52) Prabhu, N. V.; Sharp, K. A. Heat capacity in proteins. *Annu. Rev. Phys. Chem.* **2005**, *56* (1), 521−548.
(53) Makhatadze, G. I.; Privalov, P. L. Heat capacity of proteins: I. Partial molar heat capacity of individual amino acid residues in aqueous solution: Hydration effect. *J. Mol. Biol.* **1990**, *213* (2), 375−384.
(54) Cooper, A. Protein heat capacity: An anomaly that maybe never was. *J. Phys. Chem. Lett.* **2010**, *1* (22), 3298−3304.
(55) Luterbacher, J. S.; Parlange, J.-Y.; Walker, L. P. A pore-hindered diffusion and reaction model can help explain the importance of pore size distribution in enzymatic hydrolysis of biomass. *Biotechnol. Bioeng.* **2013**, *110* (1), 127−136.
(56) Tanaka, M.; Ikesaka, M.; Matsuno, R.; Converse, A. O. Effect of pore size in substrate and diffusion of enzyme on hydrolysis of cellulosic materials with cellulases. *Biotechnol. Bioeng.* **1988**, *32* (5), 698−706.
(57) Hunter, C. A. Meldola Lecture: The role of aromatic interactions in molecular recognition. *Chem. Soc. Rev.* **1994**, *23* (2), 101−109.
(58) Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5* (1), 44−53.

https://doi.org/10.1021/acs.estlett.3c00293
Environ. Sci. Technol. Lett. 2023, 10, 557−564