NAR Genomics and Bioinformatics, 2025, **7**, lqaf037
https://doi.org/10.1093/nargab/lqaf037
Standard Article

XenoBug: machine learning-based tool to predict pollutant-degrading enzymes from environmental metagenomes

Aditya S. Malwe, Usha Longwani, Vineet K. Sharma*

MetaBioSys Group, Department of Biological Sciences, Indian Institute of Science Education and Research, Bhopal, 462066, India

*To whom correspondence should be addressed. Email: vineetks@iiserb.ac.in

Abstract

Application of machine learning-based methods to identify novel bacterial enzymes capable of degrading a wide range of xenobiotics offers enormous potential for bioremediation of toxic and carcinogenic recalcitrant xenobiotics such as pesticides, plastics, petroleum, and pharmacological products that adversely impact ecology and health. Using 6814 diverse substrates involved in ∼141 200 biochemical reactions, we have developed 'XenoBug', a machine learning-based tool that predicts bacterial enzymes, enzymatic reaction, the species capable of biodegrading xenobiotics, and the metagenomic source of the predicted enzymes. For training, a hybrid feature set was used that comprises 1603 molecular descriptors and linear and circular fingerprints. It also includes enzyme datasets consisting of ∼3.3 million enzyme sequences derived from an environmental metagenome database and ∼16 million enzymes from ∼38 000 bacterial genomes. For different reaction classes, XenoBug shows very high binary accuracies (>0.75) and F1 scores (>0.62). XenoBug is also validated on a set of diverse classes of xenobiotics such as pesticides, environmental pollutants, pharmacological products, and hydrocarbons known to be degraded by the bacterial enzymes. XenoBug predicted known as well as previously unreported metabolic enzymes for the degradation of molecules in the validation set, thus showing its broad utility to predict the metabolism of any input xenobiotic molecules. XenoBug is available on: https://metabiosys.iiserb.ac.in/xenobug.

Introduction

Industrialization, modernized agriculture, and rapid urbanization due to population growth have led to increased production and utilization of synthetic chemicals such as pesticides, chemical fertilizers, and different types of plastic such as polyethylene (PE), polyethylene terephthalate (PET), polyurethane (PU), and polyvinyl chloride (PVC) due to their affordability, inertness, and convenience in day-to-day life [1, 2]. However, inadequate treatment practices for industrial waste and agriculture runoff along with excessive use of chemical fertilizers and pesticides leads to accumulation of synthetic chemicals and heavy metals, resulting in soil and water pollution [3, 4]. Some of the earlier studies had predicted loss of almost 25% of global species over time due to environmental pollution [5]. For example, loss of habitat and species diversity in aquatic ecosystems due to eutrophication is directly related to heavy metal pollution [6, 7]. Similarly, biomagnification of polycyclic hydrocarbons, pesticides, and heavy metals in terrestrial and aquatic ecosystems poses another risk for species loss [8, 9]. Toxic effects of synthetic compounds and environmental pollutants contribute to almost 22% of global disease burden [10]. Environmental pollutants cause long-term kidney diseases [11, 12] and endocrine disorders [13, 14], and also have carcinogenic effects [15, 16]. Recent reports have also highlighted that microplastics arising out of weathering and photodegradation of plastic waste contaminate terrestrial and aquatic ecosystems as well as the human body [17–19].

Bioremediation and phytoremediation techniques involving microbes and plants, respectively, show greater promise in mitigating environmental pollutants as these methods are inexpensive and reduce secondary waste [20]. Phytoremediation can improve soil quality by sequestering soil contaminants in plant biomass and using this biomass in biofuel production, construction, composting, etc. [20, 21]. However, excess land use, use of specific plant varieties, and the requirement for multiple growing seasons for effective clean-up of contaminants limit the use of phytoremediation [20]. On the other hand, bioremediation using microbes and microbial enzymes provides better utility in mitigating contaminants. Bioremediation utilizes biotransformation or absorption of contaminants using bacterial biochemical activities [3]. Many studies have shown the ability of various *Arthrobacter* sp. to adapt to oxidative and heavy metal stress and to degrade triazine herbicides [22–24]. In addition to this, many marine bacteria show important adaptive features that make them suitable for bioremediation [25]. Novel marine bacteria belonging to the *Cycloclasticus* genus show hydrocarbon biodegradation capabilities [26]. Naphthalene-degrading bacterial species such as *Pseudomonas fluorescens*, *Brevundimonas vesicularis*, and *Comamonas testosteroni* were isolated from marine sources [27]. Similarly, enzymes obtained from microbes can be immobilized on an inert matrix and used to accelerate bioremediation of environmental contaminants. Lignin peroxidases and various other types of oxidoreductases are immobilized and used as green catalysts for degradation of contaminants such

Received: September 28, 2024. **Revised:** March 9, 2025. **Editorial Decision:** March 18, 2025. **Accepted:** April 28, 2025

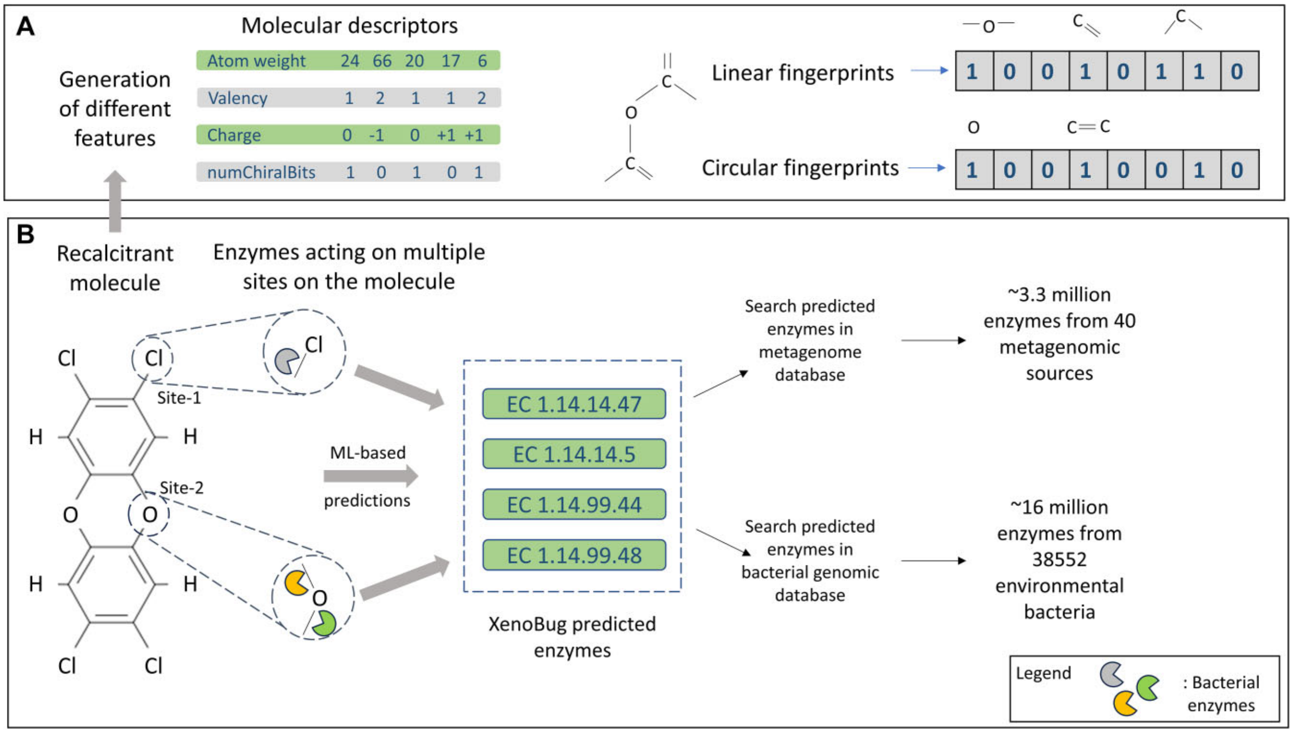
A
Molecular descriptors
Generation of different features
Atom weight 24 66 20 17 6
Valency 1 2 1 1 2
Charge 0 -1 0 +1 +1
numChiralBits 1 0 1 0 1
Linear fingerprints
1 0 0 1 0 1 1 0
Circular fingerprints
1 0 0 1 0 0 1 0
B
Recalcitrant molecule
Enzymes acting on multiple sites on the molecule
Cl
Cl
H
H
O
O
H
H
Cl
Cl
Site-1
Site-2
ML-based predictions
EC 1.14.14.47
EC 1.14.14.5
EC 1.14.99.44
EC 1.14.99.48
XenoBug predicted enzymes
Search predicted enzymes in metagenome database
~3.3 million enzymes from 40 metagenomic sources
Search predicted enzymes in bacterial genomic database
~16 million enzymes from 38552 environmental bacteria
Legend : Bacterial enzymes

Figure 1. Graphical abstract of the XenoBug pipeline showing (**A**) generation of different types of features and (**B**) using machine learning to predict pollutant-degrading enzymes and identify the source of the biodegrading enzymes.

as phenols, bisphenols, and synthetic dyes [28, 29]; however, different enzymes require different types of immobilization as the 3D structure, cofactors, and mode of action are different for each enzyme [30].

Conventional methods for developing bioremediation techniques require isolation, characterization, and identification of environmental microbes and their enzymes capable of biodegradation of contaminants using culture-based methods, molecular biology, and mass spectrometry analysis [31]. Besides these methods, functional metagenomics is also used for discovery of environmental microbes and their enzymes involved in biodegradation of contaminants. For example, using this approach, a novel carboxyesterase CE_Ubrb was identified from a carbamate-degrading uncultivated soil bacterium [32]. A metagenomics approach is also used for identifying plastic-degrading bacterial enzymes [19, 33]. However, these methods rely on known contaminant-degrading enzymes to identify their presence in metagenome sequences. As a result, identifying novel enzymes capable of degrading specific pollutants becomes difficult. In addition, a single pollutant can also be potentially degraded by multiple enzymes, which adds to the complexity of the problem.

In such a scenario, using predictive approaches such as machine learning-based classifiers to predict pollutant-degrading enzymes from bacterial genomes and metagenomes can provide valuable leads towards developing novel bioremediation strategies. All enzymes are identified by their unique Enzyme Commission number (EC number) that is assigned by International Union of Biochemistry and Molecular Biology (IUBMB) [34] which provides information regarding the reaction mechanism as well as primary substrates of the enzyme. Tools and databases such as PlasticDB [35], University of Minnesota-Biocatalysis/Biodegradation database (UM-BDD) [36], RemeDB [37], and BioSurfDB [38] are available that contain information regarding a limited number of known environmental contaminants and specific bacterial enzymes capable of degrading them. RemeDB also provides the option to identify homologous contaminant-degrading enzymes. Here, we have developed 'XenoBug', a machine learning-based tool that predicts EC numbers of bacterial enzymes from various environmental bacteria and metagenomes capable of biodegrading any contaminant of interest (Fig. 1). Prediction of multiple enzymes degrading any pollutant of interest is a multilabel classification problem. As a result, XenoBug uses multiple performance metrics such as F1 score, hamming loss, accuracy, and binary accuracy for efficient evaluation of trained models. To attest to the generalizability and substantiate its performance, XenoBug is also validated using a vast array of pollutants and xenobiotics, which were not a part of the training or test datasets. Using XenoBug, researchers can get leads on specific microbial enzymes from different environments that can potentially degrade pollutants of interest (Fig. 2).

Materials and methods

Construction of the substrate database for training

To predict EC numbers of bacterial enzymes metabolizing environmental contaminants and pollutants, a database consisting of primary substrates for all enzymatic reactions was constructed. For this purpose, multiple publicly available databases containing standard and curated information regarding enzyme reactions, their EC numbers, and substrates, such as the Braunschweig Enzyme Database (BRENDA) [39] and Kyoto Encyclopaedia of Genes and Genomes (KEGG) [40] database were used. To ensure synthetic chemicals and environmental pollutants are also part of the training, various specific databases such as PlasticDB, UM-BDD, and the Plastic Microbial Biodegradation Database (PMBD) [41] were

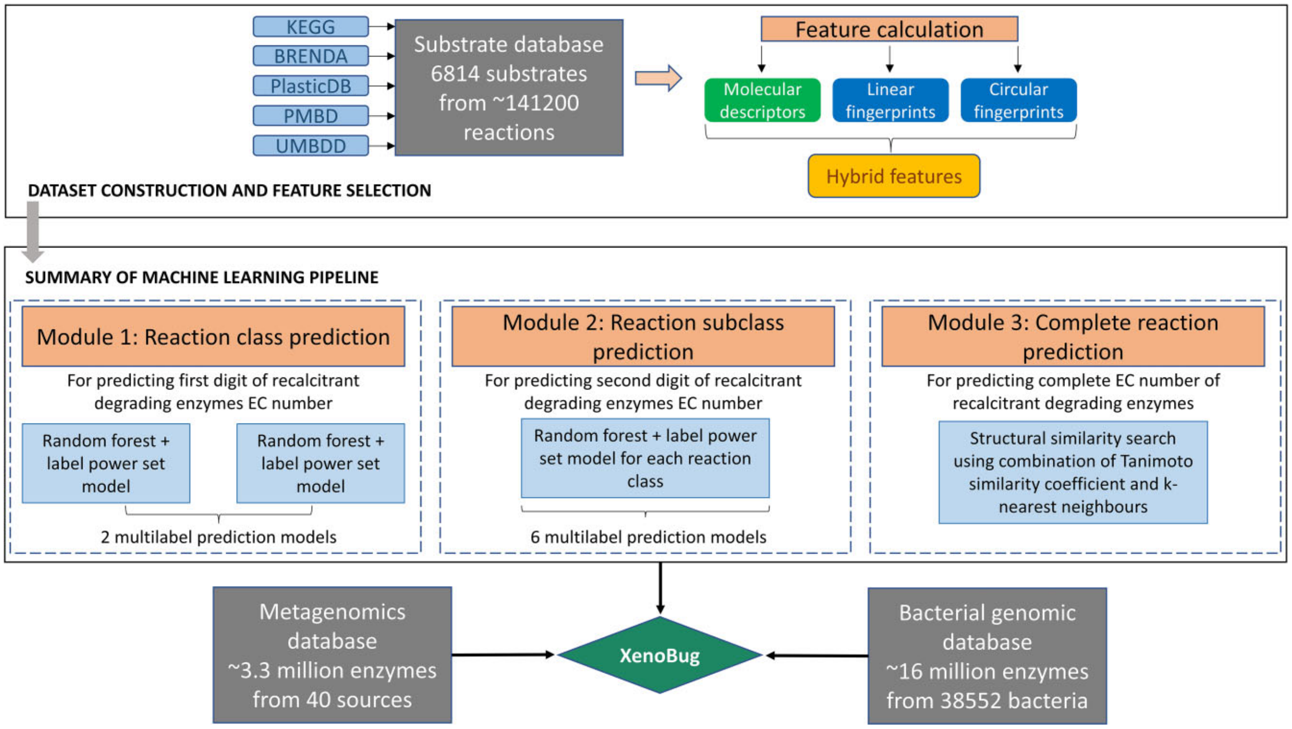
KEGG
BRENDA
PlasticDB
PMBD
UMBDD
Substrate database 6814 substrates from ~141200 reactions
Feature calculation
Molecular descriptors
Linear fingerprints
Circular fingerprints
Hybrid features
DATASET CONSTRUCTION AND FEATURE SELECTION
SUMMARY OF MACHINE LEARNING PIPELINE
Module 1: Reaction class prediction
For predicting first digit of recalcitrant degrading enzymes EC number
Random forest + label power set model
Random forest + label power set model
2 multilabel prediction models
Module 2: Reaction subclass prediction
For predicting second digit of recalcitrant degrading enzymes EC number
Random forest + label power set model for each reaction class
6 multilabel prediction models
Module 3: Complete reaction prediction
For predicting complete EC number of recalcitrant degrading enzymes
Structural similarity search using combination of Tanimoto similarity coefficient and k-nearest neighbours
Metagenomics database ~3.3 million enzymes from 40 sources
XenoBug
Bacterial genomic database ~16 million enzymes from 38552 bacteria

Figure 2. Complete pipeline of XenoBug.

also used. PlasticDB and PMBD contain detailed information regarding the microbial enzymes biodegrading the molecules used in different types of plastics. Similarly, UM-BDD contains microbial degradation pathways for diverse groups of synthetic chemicals/xenobiotics. In total, 21 036 substrates were obtained using all the databases, and .sdf files containing the structural information for each molecule were downloaded from PubChem [42] and ChEBI [43]. Due to a higher degree of overlap between all these databases, an all-versus-all Tanimoto structural similarity search [44] was used to remove redundancies. Molecules with a > 0.95 Tanimoto coefficient were removed, resulting in 6814 molecules that were then tagged with their corresponding complete EC numbers to be used for training.

Construction of the enzyme database from metagenomic data

To identify pollutant-biodegrading enzymes from available metagenomic datasets, we used the NCBI metagenome database (version v5) to retrieve all available protein fasta files. Using the ‘blastcmd’ command, the protein fasta sequences were retrieved from the metagenome database which contained the annotations for each sequence along with the origin of the sample. In the next step, eggNOG-mapper [45] (e-value threshold 0.001, identity threshold 40, query coverage 20, emapper version 2.1.12, eggnog DB version 5.0.2) was used to functionally annotate each sequence and determine the EC number for enzymatic proteins. Using this approach, EC numbers for 3 365 611 enzymatic proteins belonging to 40 metagenome origins were assigned.

Construction of the enzyme database from bacterial genomic data

To predict which environmental bacteria harbour the predicted pollutant-biodegrading enzymes, an enzyme database obtained from bacterial genomes was also constructed. A list of all available bacterial refseq genomes from the NCBI Genomes database was obtained. From this, protein sequences from bacterial genomes with complete and chromosome-level assemblies were downloaded. In total, protein sequences from 38 552 bacterial genomes were obtained. In the next step, eggNOG-mapper (e-value threshold 0.001, identity threshold 40, query coverage 20, emapper version 2.1.12, eggnog DB version 5.0.2) was used to functionally annotate each sequence and determine the EC number for enzymatic proteins. In total, 11 834 667 enzymes from 38 552 bacterial strains were obtained.

Generation of different types of features for molecules in the substrate database

To use molecules/compounds to train and develop a machine learning-based classifier, features that capture structural information as well as physicochemical properties of the molecules are required. Linear and circular fingerprints encode structural information for any molecule and are represented in the form of bits. Linear fingerprints are a set of 10 208 bits and are generated using PaDeL software [46], whereas circular fingerprints are a set of 1024 bits generated using RDKit (RDKit: Open-source cheminformatics; http://www.rdkit.org) (version 2022.09.4). In addition to these, molecular descriptors are a set of 208 features that contain physicochemical information about the molecules such as valency, charge, molecular weight, etc., and are also generated using RDKit. These different types of features are routinely used in machine learning tools that utilize molecules/compounds for training [47–50]. Thus, for all of the 6814 molecules, these three sets of features were generated independently. Molecular descriptors, linear fingerprints, and circular fingerprints are routinely used as features for compound/molecule data as they can capture the structural and physical properties of the molecules. The relatively low number of features generated by each set makes machine learning modelling more efficient.

In addition to the three sets of features, we also combined the most important features from each set to generate a hybrid set of features (referred to as 'hybrid features' throughout the manuscript) using 'feature selection'. Feature selection helps in identifying and selecting the most important features out of all the available features without any loss of performance of the prediction model. Feature selection was executed using the Boruta library [51] available in R (version 3.4). Boruta is a wrapper-based method that uses a random forest (RF) classifier as the base model to calculate feature importance. For each feature in the dataset, Boruta generates its corresponding shadow feature and determines the importance of each feature for the real and shadow feature. All the features that have higher importance than their corresponding shadow feature are selected and the rest are rejected. Using Boruta, feature selection was performed on the training set to determine important linear fingerprints and molecular descriptors using 100 iterations, that yielded 440 linear fingerprints and 139 molecular descriptors. Thus, after combining the complete set of circular fingerprints with the selected linear fingerprints and molecular descriptors, 1603 hybrid features were used as the fourth set of features.

Performance evaluation

Prediction of bacterial enzymes that degrade pollutants and other recalcitrant compounds is a multilabel classification problem. In multilabel classification, each data point belongs to multiple target labels (e.g. a single molecule can be used as a substrate by multiple enzymes) [52, 53]. As a result, it is essential to address the classification performance through multiple evaluation metrics. Accuracy refers to the ability of the model to predict all target labels correctly.

$$\text{Accuracy} = \frac{\text{Number of completely correct predicted instances}}{\text{Total number of instances}}$$

Similarly, binary accuracy is defined as the ability of the model to predict each label separately. Binary accuracy gives 'partial credit' to the model if it is able to predict a label from combination of all labels.

$$\text{Binary accuracy yi} = \frac{\text{Number of correct predictions for yi}}{\text{Total number of instances of yi}}$$

where, yi = ith label from the total number of labels y.

Hamming loss is another key metric routinely used in multilabel classification and is defined as the fraction of incorrectly predicted labels from the combination of labels for each point. The lower the hamming loss, the better is the performance of the model [47, 54].

$$\text{Hamming loss} = \frac{\text{Number of incorrectly predicted labels}}{\text{Total number of labels for that instance}}$$

Finally, the F1 score is another key metric that incorporates precision and recall during multilabel classification, thus giving us information about how a model is able to balance precision and recall despite having imbalance.

$$\text{F1 score} = 2 \text{ x } \frac{2 \text{ x (Precision x recall)}}{\text{Precision} + \text{Recall}}$$

The multilabel classification performance of XenoBug to predict pollutant-degrading bacterial enzymes was evaluated using all these metrics.

Development of machine learning-based classification models to predict pollutant-degrading enzymes

Prediction of the complete EC number of pollutant-degrading enzymes was performed in three steps: in the first step, training is performed using the substrate database to predict the first digit of the EC number that indicates the reaction class of the enzymes; in the second step, reaction class-specific training is performed to predict the reaction subclass or second digit of the EC number; and, in the third step, based on the results of reaction subclass prediction, a k-nearest neighbour (kNN) and Tanimoto similarity search is performed between the query molecule and the molecules in the substrate database to assign the complete EC number to the query molecule. Thus, using this modular architecture, multiple models were compared and optimized to develop the final pollutant-degrading enzyme prediction tool.

Prediction of reaction class (Module-1) was addressed as a multilabel classification problem as a single molecule can be involved in multiple types of reactions. In multilabel classification, a single data point may belong to multiple labels [55]. As a result, for evaluation of performance of different multilabel classification models, multiple metrics such as hamming loss (proportion of incorrectly predicted labels out of all labels), F1 score, overall accuracy, and binary accuracy for each label was calculated. The complete substrate database was used for training Module-1 and data were split iteratively into training and test data in an 80:20 ratio [56]. During training, 10-fold iterative cross-validation was also performed on training data. RF [57] and artificial neural networks (ANNs) [58] were used as base classifier models as both these algorithms are suitable for multilabel classification [47, 59–60]. Various problem transformation methods that convert a multilabel classification problem into simpler binary or multiclass problems [53], such as One-vs-Rest classifier (OVR), binary relevance (BR), classifier chain (CC), and label powerset (LPS), were compared and the best performing method was selected. Training performance of all the methods was compared using all four types of features, and the models generated using the best set of features and methods were further optimized using the GridSearchCV method [61]. Further, multiple RF parameters such as 'n_estimators' and 'max_features' and ANN hyperparameters such as number of hidden layers in the network, number of nodes per hidden layer, classifier alpha value, learning rate, and max_iter were optimized using GridSearchCV.

Development of the reaction subclass prediction module (Module-2) was also a multilabel classification problem. Since reaction subclasses are specific to a reaction class, and a variable number of subclasses are present in different reaction classes (Supplementary Table S2), six independent multilabel models for six reaction classes were developed. RF was used as the base classifier along with different problem transformation methods as RF shows good performance for multilabel classification even with limited data [62, 63]. The hybrid feature set and 10-fold cross-validation were used for training and comparing different models. In the case of EC 5 and EC 6, 5-fold cross-validation was performed as both reaction classes have a smaller number of substrates to be used for training (Supplementary Table S2).

The final module (Module-3) includes a kNN and Tanimoto coefficient-based structural similarity search. Reaction subclass-specific primary substrates are binned together and tagged with their complete EC number. Based on the prediction obtained from Module-1 and Module-2, predicted reaction subclass bins are selected and a structural similarity search of the primary substrates in the select bin is performed with the query molecule. Using various Tanimoto coefficient thresholds, substrates similar to the query molecule and their corresponding complete EC numbers are thus obtained. Thus, based on machine learning-based predictions and a structural similarity search, complete EC numbers of enzymes degrading pollutants are obtained. The predicted EC numbers are then searched in the metagenomic enzyme database constructed using eggNOG-mapper to obtain details regarding the enzyme pathway, environmental source of the predicted enzyme, and its sequence.

Development of the XenoBug webtool

XenoBug is deployed as a webtool using HTML and PHP. Users can enter PubChem ID or upload an sdf file of the query molecule as input. XenoBug provides detailed information regarding predicted enzymes, their metagenome source, pathway information, as well as the Tanimoto threshold at which EC numbers were predicted. Users can also download the predictions obtained using RF and ANN. A metagenome-assembled genomes (MAGs) pipeline is also provided in the Downloads section on the Results page, using which users can obtain eggNOG-mapper-derived EC numbers for the protein sequences in the MAGs and integrate those with predictions obtained from XenoBug to identify potential environmental pollutant-degrading enzymes in their specific MAGs. XenoBug is available at https://metabiosys.iiserb.ac.in/xenobug.

Results

Complexity of the substrate database used for training

The substrate database used for training machine learning models predicting enzymes degrading environmental pollutants was constructed using multiple databases such as BRENDA, KEGG, PlasticDB, PMBD, and UM-BDD. Among these, BRENDA contributed the highest number (141 201) of biochemical reactions, followed by the KEGG database that contributed 11 468 reactions across the six reaction classes. UM-BDD contained 992 specific reactions involving pollutant degradation, whereas databases specific to plastic degradation such as PlasticDB and PMBD contained 1670 and 132 reactions, respectively (Supplementary Table S1). From all these reactions, sdf files for 21 036 substrates were obtained, from which many secondary substrates such as cofactors, metal ions, electrons, and water molecules were removed. To remove redundant molecules from the substrate database, an all-vs-all Tanimoto structural similarity search was performed, and molecules with > 95% similarity were removed. As a result, 6814 substrates were finalized and their distribution in different EC reaction classes is provided in Supplementary Table S2. EC 1 and EC showed the highest representation of substrates, whereas EC 5 and EC 6 showed the lowest representation of substrates, which mimics the natural distribution of participating substrates across the six reaction classes.

Metagenomic enzyme database constructed using the NCBI metagenome database

To identify pollutant-degrading enzymes in metagenomic sequences, we constructed an enzyme database from the NCBI metagenome database that contains sequences functionally annotated using eggNOG-mapper, which uses the eggNOG database [64] of pre-computed orthologue groups, and DIAMOND search [65] for orthology and annotation assignments to input sequences. Protein fasta sequences were retrieved from the NCBI metagenome database using the 'blastcmd' command that retrieved 10 951 228 protein sequences. eggNOG-mapper was used to functionally annotate and identify enzymes from these protein sequences that resulted in 3 365 611 enzymes from 40 metagenome sources as available in the NCBI metagenome database. These metagenome sources include animal sources (e.g. the bird metagenome, pig metagenome, wasp and termite metagenome, human gut metagenome, etc.) and various environmental niches such as the aquatic metagenome, mine drainage metagenome, soil and sediment metagenome, etc.

Bacterial enzyme database constructed using the NCBI bacterial genome database

To identify the bacterial strains harbouring the pollutant-degrading enzymes, protein.fasta files for all complete bacterial genomes along with bacterial genomes with chromosome-level assembly were downloaded. In total, 4976 chromosome-level bacterial assemblies and 33 576 complete bacterial genomes were downloaded from the NCBI Genomes database. These genomes provided 132 866 270 proteins, out of which 16 826 385 EC number-annotated enzymes were obtained using eggNOG-mapper. Thus, ∼16 million bacterial enzymes were obtained from 38 552 bacteria found in various environments.

Module-1: reaction class prediction

Prediction of the reaction class of pollutant-degrading enzymes was addressed as a multilabel problem since a single substrate can potentially be metabolized by enzymes from more than one reaction classes. Iterative stratification was used to split the substrate dataset into a training and test set in an 80:20 ratio, which helps in retaining the distribution of labels in the training and test set in an unbiased manner. As a result, the training set included 5451 substrates, and the remaining 1363 substrates were included in the test set. Performance of different problem transformation methods such as LPS, CC, BR, and OVR that act as a wrapper on base RF and base ANN models was compared on all four feature types to select the best performing multilabel classification method. The single hidden layer ANN network was implemented using MLPClassifier available on scikit-learn, and all the above-mentioned problem transformation methods were implemented using scikit-learn and the scikit-multilearn library available on Python 3.7.9. In terms of accuracy and F1 score, RF with the LPS wrapper performed better as compared with other methods (Fig. 3A; Supplementary Figs S1 and S2), whereas in the case of hamming loss, all the methods showed a similar performance, with LPS having slightly higher ham-

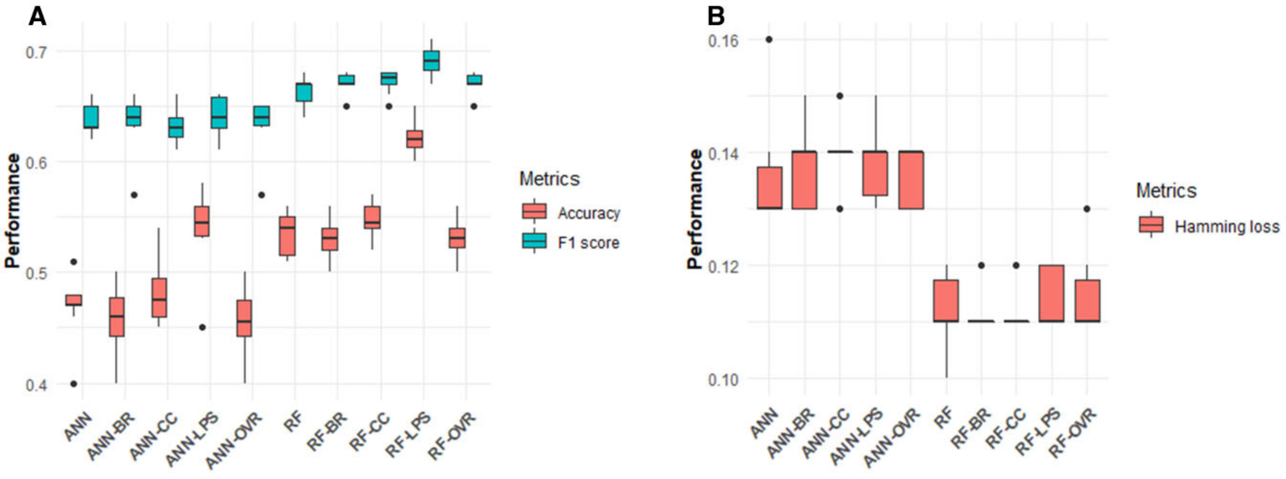

Figure 3. (**A**) Ten-fold cross-validation performance of various problem transformation methods with ANN and RF models evaluated using accuracy and F1 score. (**B**) Ten-fold cross-validation performance of various problem transformation methods with ANN and RF models evaluated using hamming loss.

ming loss (Fig. 3B; Supplementary Table S3). This trend was observed for all four types of features; however, the performance of RF with the LPS wrapper was found to be the best with the hybrid feature set (Supplementary Fig. S3). Moreover, hybrid features perform better than other individual types of features (Supplementary Table S3).

A similar trend was observed for ANN-based models where LPS together with a single layer neural network showed higher accuracy and F1 score as compared with other methods, whereas hamming loss was slightly higher in the case of LPS as compared with other methods. The best performance was obtained using hybrid features; however, the ANN + LPS method showed slightly higher accuracy and F1 score with linear fingerprints. Nevertheless, performance using a single hidden layer ANN + LPS was still marginally lower as compared with the performance obtained using the RF + LPS model (Fig. 1; Supplementary Table S3). As a result, we decided to further optimize both the RF + LPS model and ANN + LPS model using hybrid features to determine the best performing model after optimization. For RF, n_estimators specify the number of trees in the RF, and the max_features parameter specifies the number of features to be considered while selecting the best split in a tree. The vbest performance was obtained with n_estimators = 1000 and max_features = 100 (Supplementary Table S4). For ANN, best mean accuracy was obtained using two hidden layers (400 nodes and 250 nodes, respectively) with max_iters = 1000 (Supplementary Table S5). The performance of the optimized RF + LPS model on the test set yielded an accuracy of 0.6, a hamming loss at 0.12, and an F1 score of 0.67 which is similar to that obtained on the training set, indicating that the model does not suffer from any issues related to overfitting (Supplementary Table S6). Binary accuracy for each label, defined as the ability of the model to correctly predict individual labels, was also calculated for the test set (Supplementary Table S7). For each label, very high binary accuracy (>0.77) was observed. In the case of the ANN + LPS model, test set performance showed a mean accuracy of 0.53, hamming loss of 0.14, and F1 score of 0.61, which was lower than the RF + LPS model. However, binary accuracies using both the optimized models was high and similar (Supplementary Tables S6 and S7). As a result, the optimized ANN + LPS model was also selected along with the RF + LPS model for development of Module-1.

Module-2: reaction subclass prediction

Prediction of the reaction subclass of enzymes, that is represented by the second digit of the EC number, is also a multilabel problem. However, the number of substrates is drastically reduced within reaction classes (Supplementary Table S2), resulting in a reduced number of molecules for training. Moreover, the number of subclasses within different reaction classes is also variable; as a result, six different reaction subclass prediction multilabel models are required for six reaction classes. Due to the reduced number of molecules available for training, only RF-based methods were selected for constructing the subclass prediction module as RF is known to perform well in a multilabel classification problem even with fewer data [47, 57, 63]. The performance of different types of problem transformation methods applied to RF was compared for each EC class during training using hybrid features. For all EC classes, RF + LPS showed the highest accuracy and F1 score as compared with other methods (Fig. 4; Supplementary Figs S4 and S5), whereas hamming loss was similar for each method for all EC classes (Supplementary Table S8; Supplementary Fig. S6). As a result, the RF + LPS method was selected for optimization for all EC classes. RF parameter optimization was performed using GridSearchCV, and the results are provided in Supplementary Table S9. The optimized RF + LPS models showed a test set performance that was similar to their training performance for each EC class (Supplementary Table S10), thus ruling out the possibilities of overfitting and indicating sufficient training. As a result, the multilabel RF + LPS models optimized using hybrid features were finalized for each EC class that collectively form the reaction subclass prediction module.

Module-3: identification of complete EC number of enzymes potentially degrading environmental pollutants

The final module that predicts and identifies the complete EC number of enzymes degrading environmental pollutants includes performing a structural similarity search between the query molecule and molecules in the substrate database. Molecules in the substrate database are tagged with their complete EC number and binned together according to their respective reaction subclass. Based on the reaction class and re-

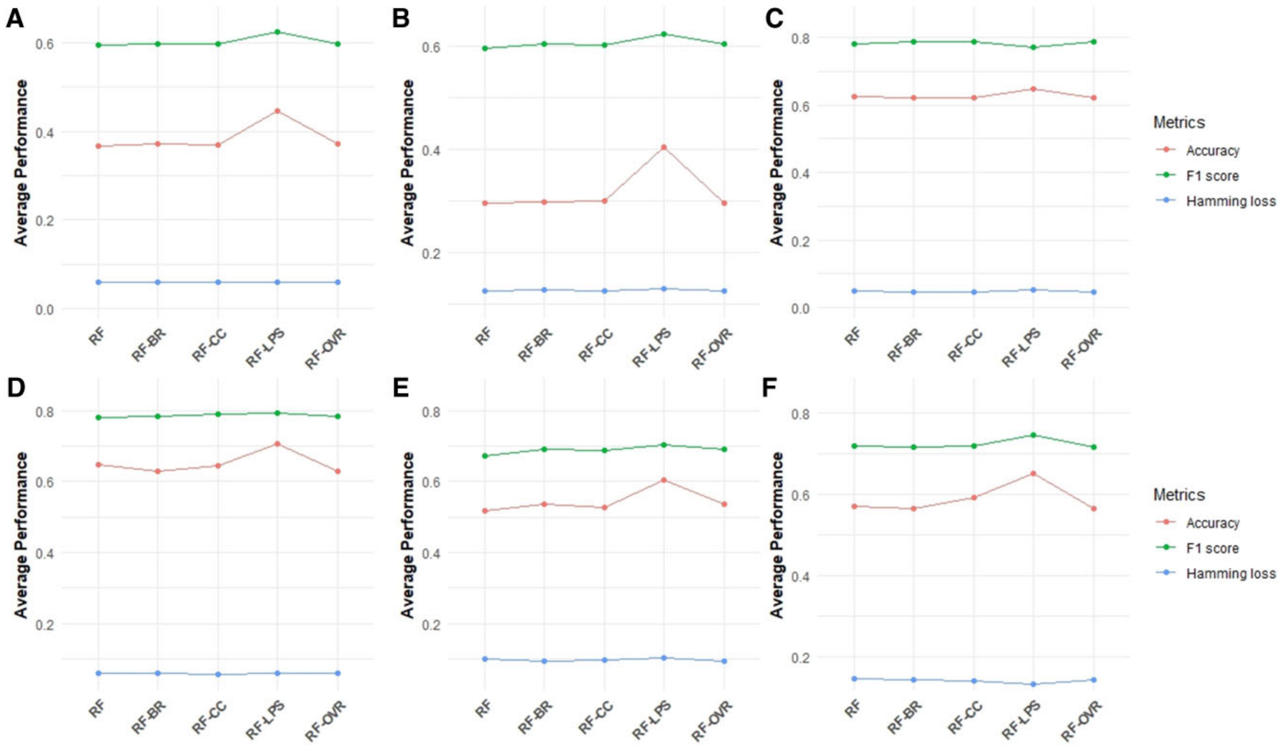

Figure 4. Average 10-fold cross-validation accuracy, F1 score, and hamming loss for multiple problem transformation methods and the RF-based model for (**A**) EC 1 subclass classification, (**B**) EC 2 subclass classification, (**C**) EC 3 subclass classification, (**D**) EC 4 subclass classification, (**E**) EC 5 subclass classification. and (**F**) EC 6 subclass classification.

action subclass predictions for the query molecule obtained from Module-1 and Module-2, respectively, a structural similarity search is performed with the substrates in the predicted reaction subclass. In the next step, using multiple Tanimoto similarity coefficients, the number of structurally similar molecules is obtained at each Tanimoto similarity coefficient threshold. This number is used as input-k for executing a kNN-based search to identify substrates (nearest neighbours) similar to the query molecule. Finally, the complete EC number of such similar molecules/nearest neighbours is assigned to the query molecule. Thus, using a combination of machine learning models, Tanimoto structural similarity search, and kNN-based identification of similar molecules, the complete EC number of enzymes potentially degrading environmental pollutants can be predicted using XenoBug.

Performance of XenoBug on the validation set

To assess the ability of XenoBug to predict novel as well as known bacterial enzymes involved in bioremediation of persistent recalcitrant compounds and pollutants, a validation set comprising 25 environmental pollutants with known bacterial biodegradation was constructed. These compounds were not part of the training or test set. The validation set comprised diverse classes of pollutants such as chlorinated and phosphorus-containing insecticides, pesticides, halogenated aliphatic hydrocarbons, aromatic compounds, and pharmacological pollutants. XenoBug predicted the complete EC number of the bacterial enzymes that can potentially degrade these molecules, and their environmental source, and provided a rich output with biochemical information, potential pathways involved in the biotransformation, and the sequence of the predicted enzymes. The predictions obtained are provided in Table 1.

p-Xylene is a major aromatic petroleum hydrocarbon used as a solvent in paints and aeroplane fuel production, and also used as a primary compound for synthesizing polymers used in PET plastics and polyester that is known to contaminate soil and groundwater [66]. Cold-active monooxygenases isolated from *Pseudomonas* sp. such as xylene monooxygenase are able to oxidize *p*-xylene [67]. XenoBug predicted multiple monooxygenases (EC 1.14) from various bacterial strains from *Pseudomonas*, *Halopseudomonas*, and *Xenorhabdus* genera that can potentially biodegrade *p*-xylene, and were found in marine and hydrocarbon metagenomes. Triazophos is an organophosphorus pesticide that is extremely toxic to human beings and other animals [68]. Triazophos-degrading hydrolase (EC 3.1) was identified from *Ochrobacterium* sp. isolated from contaminated soil and was correctly predicted by XenoBug.

Similarly, XenoBug predicted reducing as well as oxidizing enzymes from multiple bacterial genera such as *Mycobacterium*, *Pseudomonas*, and *Gordonia*, for herbicides, pesticides, and hydrocarbons such as acetochlor, endosulfan, *n*-hexadecane, and pentachlorophenol. Experimental evidence also highlights the role of reducing and oxidative enzymes found in *Sphingobium* sp., *Mycobacterium* sp., and *Gordonia* sp. isolated from various sources such as activated sludge, contaminated soil, and marine sediments that can degrade acetochlor [69, 70], endosulfan [71, 72], *n*-hexadecane [73], and pentachlorophenol [74].

XenoBug was also able to predict known hydrolysing enzymes involved in biotran4sformation of atrazine and alpha-hexachlorocyclohexane (popularly known as Lindane).

Table 1. Performance of XenoBug on the validation set consisting of diverse classes of recalcitrant molecules

Compound	Known bacterial enzymes involved in degradation	Known information regarding source of enzymes	XenoBug-predicted enzymes involved in degradation (subclass level) and bacteria (genus level)	XenoBug-predicted metagenomic sources	References
p-Xylene	Xylene monooxygenase (EC 1.14)	*Pseudomonas* sp.	EC 1.14, *Pseudomonas*, *Halopseudomonas*, *Xenorhabdus*	Marine metagenome, hydrocarbon metagenome	(Miri *et al*., 2022)
Triazophos	Triazophos hydrolase (EC 3.1)	*Ochrobacterium* sp. isolated from soil	EC 3.1, *Ochrobactrum*, *Novosphingobium*, *Blastomonas*	Marine metagenome, compost metagenome	(Gu *et al*., 2006)
Decabromodiphenyl ether	NA (debromination, hydroxylation)	*Brevibacillus* sp., *Achromobacter* sp.	EC 1.14, *Achromobacter*, *Brevibacillus*, *Arthrobacter*	Marine metagenome, compost metagenome	(Hu *et al*., 2022)
Acetochlor	Oxygenase (EC 1.14)	*Sphingobium* sp. isolated from activated sludge	EC 1.14, *Sphingobium*, *Sphingobacterium*, *Pseudoarthrobacter*	Marine metagenome, anaerobic digester metagenome, compost metagenome	(Li *et al*., 2013; Liu *et al*., 2019)
Toxaphene	NA (reduction)	*Bjerkandera* sp. BOL13	EC 1.14, *Streptomyces*, *Bacillus*, *Brachybacterium*	Marine metagenome, bioreactor metagenome	(Romero *et al*., 2006; Prieto *et al*., 2021)
Dibenzofuran	Catechol-2,3-dioxygenase (EC 1.13), dibenzofuran dioxygenase (EC 1.14)	*Serratia marcescens*	EC 1.14, *Pseudomonas*, *Escherichia*, *Serratia*	Compost metagenome, bioreactor metagenome, marine metagenome	(Kasuga *et al*., 1997; Jaiswal and Thakur, 2007)
Atrazine	Atrazine chlorohydrolase (EC 3.8)	*Pseudomonas* sp. strain ADP and *Arthrobacter* sp. from contaminated soil	EC 3.5, *Arthrobacter*, *Arsenophonus*, *Pseudomonas*	Marine sediment metagenome, hydrocarbon metagenome, soil metagenome	(Neumann *et al*., 2004; Sajjaphan *et al*., 2004)
3,3′,4,4′-Tetrachlorobiphenyl	Biphenyl dioxygenase (EC 1.14), dihydrodiol dehydrogenase (EC 1.3)	*Burkholderia xenovorans* LB400, *Pseudomonas pseudoalcaligens* KF707	EC 1.1, *Pseudomonas*, *Bacillus*, *Burkholderia*	Marine metagenome, compost metagenome, termite gut metagenome	(Furukawa and Fujihara, 2008)
Aldrin	NA	*Pseudonocardia* sp. KSF27, *Pseudomonas fluorescens*	EC 1.14, *Bacillus*, *Bradyrhizobium*, *Pseudonocardia*	Marine metagenome, compost metagenome	(Sakakibara *et al*., 2011; Pang *et al*., 2022)
Glyphosate	Glycerol-3-phosphate dehydrogenase (EC 1.1)	*Pseudomonas pseudomallei* isolated from soil	EC 1.14, *Pseudomonas*, *Streptomyces*, *Escherichia*	Marine metagenome, soil metagenome	(Peñaloza-Vazquez *et al*., 1995; Liu *et al*., 2019)
Alpha-hexachlorocyclohexane (lindane)	Haloalkane dehalogenase (EC 3.8)	*Sphingomonas paucimobilis*	EC 3.8, *Bacillus*, *Sphingomonas*, *Bifidobacterium*	Subsurface metagenome, marine metagenome,	(Camacho-Pérez *et al*., 2012)
2,3,7,8-Tetrachlorodibenzo-*p*-dioxin	Lignin peroxidase (EC 1.11), angular dioxygenase (EC 1.13)	*Sphingomonas* sp., *Burkholderia* sp.	EC 1.14, *Escherichia*, *Novosphingonium*, *Paraburkholderia*	Marine metagenome, freshwater metagenome	(Field and Sierra-Alvarez, 2008; Sakaki and Munetsuna, 2010)
N-Hexadecane	*n*-Alkane hydroxylase (EC 1.14)	*Gordonia* sp. isolated from Mediterranean shoreline	EC 1.14, *Gordonia*,*Klebsiella*, *Erwinia*	Marine metagenome, marine sediment metagenome	(Quatrini *et al*., 2007)
Chlordecone	NA (reductive sulfidation)	*Citrobacter* sp.86, *Desulfovibrio* sp.86	EC 1.14, *Streptomyces*, *Desulfosediminicola*, *Burkholderia*	Marine metagenome, compost metagenome	(Chaussonnerie *et al*., 2016; Della-Negra *et al*., 2020)
Clofenotane (DDT)	NA (dechlorination)	*Chryseobacterium* sp. PYR2, *Alcaligenes eutropha*	EC 1.14, *Alcaligenes*, *Chryseobacterium*, *Comamonas*	Marine metagenome, compost metagenome, freshwater metagenome	(Mansouri *et al*., 2017)

Table 1. Continued

Compound	Known bacterial enzymes involved in degradation	Known information regarding source of enzymes	XenoBug-predicted enzymes involved in degradation (subclass level) and bacteria (genus level)	XenoBug-predicted metagenomic sources	References
Pentachlorobenzene	Dienelactone hydrolase (EC 3.1), maleylacetate reductase (EC 1.3)	*Xanthobacter flavus* 14p1	EC 2.7, EC 1.20, *Xanthobacter*, *Acinetobacter*, *Scytonema*	Bioreactor metagenome, subsurface metagenome, marine sediment metagenome	(Sommer and Görisch, 1997)
Endosulfan	Monooxgenase (EC 1.14), phenol hydroxylase (EC 1.14)	*Mycobacterium* sp.	EC 1.14, *Pseudomonas*, *Shigella*, *Mycobacterium*	Compost metagenome, freshwater metagenome, marine metagenome	(Sutherland *et al.*, 2002; Singh et al., 2019)
Hexabromocyclododecane	NA (debromination, hydroxylation)	*Pseudomonas* sp. HB01, *Achromobacter* sp. HBCD-1	EC 1.14, *Streptomyces*, *Saccharomonospora*, *Synechococcus*	Mine drainage metagenome, marine metagenome	(Huang *et al.*, 2020)
3-Fluorobenzoic acid	Benzoate dioxygenase (EC 1.14)	*Sphingomonas* sp. HB1	EC 1.14, *Pseudomonas*, *Sphingomonas*, *Vogesella*	Freshwater metagenome, marine metagenome	(Hidde Boersma *et al.*, 2004; Murphy, 2016)
2,5,6,7,8,11,15-Heptachloroheptadecane (short chained chlorinated paraffin)	Haloalkane dehalogenase (EC 3.8)	*Sphingomonas paucimobilis*	EC 1.14, *Streptomyces*, *Kutzneria*, *Sphingomonas*	Compost metagenome, marine metagenome	(Knobloch *et al.*, 2021)
Pentachlorophenol	Pentachlorophenol monooxygenase (EC 1.14), TCBQ reductase (EC 1.1)	*Sphingobium chlorophenolicum*	EC 1.14, *Pseudomonas*, *Sphingobium*	Marine metagenome, compost metagenome	(Dai and Copley, 2004)
Dieldrin	NA	*Pseudonocardia* sp. KSF27, *Pseudomonas fluorescens*	EC 3.3, EC 1.14, *Pseudomonas*, *Pseudonocardia*, *Xanthomonas*	Marine metagenome, hydrothermal vent metagenome, plant metagenome	(Sakakibara *et al.*, 2011; Pang *et al.*, 2022)
Carbamazepine	Amidohydrolase (EC 3.5)	*Gordonia polyophrenivorans*	EC 1.14, *Escherichia*, *Shigella*, *Gordonia*	Human gut metagenome, marine metagenome	(Wang *et al.*, 2023)
Fluorene	Angular dioxygenase (EC 1.13)	*Sphingomonas* sp. LB126	EC 1.14, *Streptomyces*, *Sphingomonas*, *Pseudomonas*	Insect metagenome, marine metagenome	(Schuler *et al.*, 2008)
Phenanthrene	Biphenyl-2,3-diol dioxygenase (EC 1.13), carboxylase (EC 4.1)	*Sphingobium* sp.	EC 1.14, *Novosphingobium*, *Sphingobium*, *Paenibacillus*	Marine metagenome, termite metagenome	(Waigi *et al.*, 2015; Sun *et al.*, 2022)

Atrazine and Lindane are highly toxic to mammals including human beings and aquatic animals. Atrazine chlorohydrolase (EC 3.8) shows specific activity in degradation of atrazine and was isolated from multiple bacteria found in contaminated soil such as *Pseudomonas* sp. ADP and *Arthrobacter* sp. [22, 75]. XenoBug was able to predict various hydrolysing enzymes belonging to EC 3.5 from strains belonging to *Arthrobacter, Arsenophonus*, and *Pseudomonas* genera that can potentially biodegrade atrazine, and these enzymes were mapped to the soil metagenome and the marine sediment metagenome. Similarly, experimental evidence shows that enzymes belonging to the haloalkane dehalogenase family (EC 3.8) isolated from *Sphingomonas paucimobilis* are capable of hexachlorocyclohexane biodegradation [76]. XenoBug correctly predicted multiple enzymes belong to EC 3.8 that can potentially metabolize hexachlorocyclohexane.

Severe toxicity of halogenated and polycyclic aromatic pollutants highlights the importance of mitigation of such compounds. XenoBug predicted redox enzymes belonging to EC 1.14 that can potentially degrade 3-fluorobenzoic acid, fluorene, and phenanthrene. Benzoate dioxygenase (EC 1.14) isolated from *Sphingomonas* sp. HB1 is known to metabolize 3-fluorobenzoic acid, whereas biphenyl-2,3-diol dioxygenase (EC 1.13) and an angular dioxygenase (EC 1.13) have been experimentally shown to degrade phenanthrene and fluorene, respectively [77, 78].

For many pollutants in the validation set, the bioremediation product and mechanism of biodegradation are known, but bacterial enzymes involved in their degradation are not known. For such pollutants, XenoBug was able to predict bacterial enzymes, along with their sequence and metagenomic source, that have a similar mechanism of biotransformation to that experimentally reported. Decabromodiphenyl ether is subjected to debromination and hydroxylation by *Brevibacterium* sp. and *Achromobacter* sp., but the enzymes involved are not yet identified [79]. XenoBug predicted multiple hydroxylating enzymes such as polyprenyl-methoxyphenol hydroxylase, geranyl reductase, and phenylpropionate dioxygenase (all belonging to EC subclass 1.14) from known decabromodiphenyl ether-degrading bacterial genera such as *Brevibacterium* and *Achromobacter*, along with novel *Arthrobacter* strains that can potentially hydroxylate and biodegrade decabromodiphenyl ether.

Toxaphene, a very potent carcinogen [80], is subjected to reduction by the metabolic activity of *Bjerkandera* sp. BOL13, but the enzymes involved in this metabolism are yet to be identified [81, 82]. XenoBug predicted various reductases such as methylene tetrahydromethanopterin reductase, ketopantoate reductase, and other flavin-dependent reductases that can perform reduction of toxaphene. Aldrin and dieldrin are closely related chlorinated pesticides that have neurotoxic effects on animals. Multiple studies reported degradation of aldrin and dieldrin by *Pseudonocardia* sp. KSF27 and *Pseudomonas fluorescens,* but the biodegrading enzymes are not yet known [83, 84]. XenoBug predicted reductase, dioxygenase, and hydroxylase (EC 1.14) from multiple known bacterial genera known to degrade aldrin and dieldrin such as *Pseudomonas* and *Pseudonocardia*, as well as previously unknown *Xanthomas* and *Bradyrhizobium* genus and multiple metagenomic sources such as the plant metagenome, compost metagenome, and marine metagenome that can potentially metabolize and degrade aldrin and dieldrin. Clofenotane (popularly known as DDT) is a persistent and highly toxic pesticide for which bioremediating enzymes have not yet been identified, but experimental evidence suggests that clofenotane is subjected to dechlorination by the activity of *Chryseobacterium* sp. PYR2 and *Alcaligenes eutropha* [85]. XenoBug was able to correctly predict halogenase from *Chryseobacterium* and *Alcaligenes* that can potentially perform dechlorination of clofenotane.

Thus, the performance of XenoBug on the validation set highlights its ability to predict novel as well as existing bacterial enzymes involved in bioremediation of important environmental pollutants and further attests to its utility and applicability.

Discussion

Diverse classes of chemical compounds such as pesticides, petroleum and plastic products, and polycyclic aromatic compounds have considerable importance in the developed and developing world because of their utility, and ease of production and usage. However, most of these compounds are recalcitrant and toxic to human beings and wildlife, and also persist in the environment, causing pollution. The detrimental effect of such recalcitrant compounds and contaminants on the ecology and human health necessitates efficient pollutant mitigation strategies. Bioremediation is the most efficient and inexpensive method of pollution mitigation as compared with other methods and has higher utility in addressing this massive environmental crisis [2, 86]. However, current culturomics and computational methods are inadequate for discovering novel bioremediating enzymes for such diverse classes of contaminants. A machine learning-based tool such as XenoBug can, however, address this challenging task and can predict all potential bacterial enzymes involved in biodegradation of pollutants, contaminants, and recalcitrant molecules, identify the metagenomic source of such enzymes, and also provide protein sequences of the predicted enzymes.

The incorporation of a comprehensive dataset of 6814 primary substrates of metabolic enzymes from multiple biochemical databases, and a vast database of ∼2.8 million metagenomic enzymes from 40 metagenomic sources in XenoBug helps to identify and discover novel bioremediating enzymes directly from available metagenomic data. In addition, it can identify the bacteria that harbour the predicted pollutant-degrading enzymes from the 16 million bacterial enzymes from 38 552 environmental bacteria. Thus, using XenoBug's inbuilt metagenomic enzyme database, prediction of enzymes involved in bioremediation of specific pollutants and their protein sequence provides valuable leads for further experimental studies.

Another important characteristic of XenoBug is that all the artificial intelligence machine learning models are trained using hybrid features that capture physical and structural properties of substrates by combining important molecular descriptors, linear fingerprints, and circular fingerprints. Moreover, using a structural similarity search helps in identifying bacterial enzymes with promiscuous ability to potentially metabolize and biodegrade pollutants. Utilizing RF and ANNs along with a problem transformation method such as label powerset helps in executing multilabel classification for predicting diverse classes of enzymes. XenoBug is organized in a modular architecture that will make future updates of the tool easier based on the availability of more data. Furthermore, the complete biochemical information regarding the reactions catalysed by predicted enzymes, their biochemical pathways, and protein sequence as provided by XenoBug furthers our understanding of the interaction of environmental bacterial enzymes and pollutants and their potential bioremediation. The training scheme and development of XenoBug tool is explained in Fig. 2.

The performance of XenoBug on diverse recalcitrant molecules from the validation set such as *p*-xylene, triazophos, endosulfan, and *n*-hexadecane is in agreement with the experimental evidence, whereas in the case of pollutants such as toxaphene, aldrin, dieldrin, DDT, etc., for which the biodegrading enzymes are yet to be discovered, XenoBug is able to predict potential bacterial enzymes that can perform their biotransformation. These validations attest to its applicability and utility, and the predictions provide valuable leads for experimental confirmation which can help in developing bacterial enzymes for bioremediation of existing harmful and toxic pollutants, recalcitrant molecules, and contaminants on a large scale. This approach currently uses conventional cheminformatic features for training and prediction of biochemical enzymes using substrate data. Recent advances in natural language processing (NLP)-based feature extraction using SMILEs data [87] can also be used for such applications provided the computational power and memory are scaled efficiently. Training using 3D cheminformatic features for substrates can be compared with currently used 2D descriptors and fingerprints. One of the current limitations is the unequal number of enzymes and their corresponding substrates naturally known in the different EC classes (Supplementary Table S2). This results in data imbalance during training. To address this issue, the approach to artificially oversample minority classes such as EC 5 and EC 6 to the levels of EC 1 or EC 3, or to undersample the majority classes to the levels of minority classes may not be appropriate and may also lead to overfitting. Thus, in this study, the evaluation of models constructed in XenoBug was performed using multiple evaluation metrics such as F1 score, hamming loss, accuracy, and binary accuracy that take into account various aspects of data and accommodate data imbalance. An improved understanding of enzyme–substrate interactions, advanced algorithms, and the availability of a higher number of substrates for minority EC classes in the future will further help in improving data balance and performance of the prediction models. XenoBug is free to use and publicly accessible online at https://metabiosys.iiserb.ac.in/xenobug.

Acknowledgements

We thank IISER Bhopal for providing the intramural funds and the Department of Biotechnology for providing the funds to carry out this work.

Author contributions: A.S.M. and U.L. were involved in dataset preparation, data curation, and data organization. A.S.M. performed and constructed machine learning training and the webtool. A.S.M. and V.K.S. wrote the manuscript, and all authors were involved in editing and reviewing the manuscript.

Supplementary data

Supplementary data are available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

Department of Biotechnology, Ministry of Science and Technology, India [BT/PR51934/BTIS/137/86/2024].

Data availability

XenoBug is free to use and publicly accessible online at https://metabiosys.iiserb.ac.in/xenobug.

References

1. Johnson LM, Talbot HW. Detoxification of pesticides by microbial enzymes. *Experientia* 1983;**39**:1236–46. https://doi.org/10.1007/BF01990361
2. Singh Jadaun J, Bansal S, Sonthalia A *et al.* Biodegradation of plastics for sustainable environment. *Bioresour Technol* 2022;**347**:126697. https://doi.org/10.1016/j.biortech.2022.126697
3. Sharma P, Kumar S, Pandey A. Bioremediated techniques for remediation of metal pollutants using metagenomics approaches: a review. *J Environ Chem Eng* 2021;**9**:105684. https://doi.org/10.1016/j.jece.2021.105684
4. Randika JLPC, Bandara PKGSS, Soysa HSM *et al.* Bioremediation of pesticide-contaminated soil: a review on indispensable role of soil bacteria. *J Agric Sci – Sri Lanka* 2022;**17**:19–43.
5. McNeely JA. The sinking ark: pollution and the worldwide loss of biodiversity. *Biodivers Conserv* 1992;**1**:2–18. https://doi.org/10.1007/BF00700247
6. Gagneten AM, Paggi JC. Effects of heavy metal contamination (Cr, Cu, Pb, Cd) and eutrophication on zooplankton in the lower basin of the Salado River (Argentina). *Water Air Soil Pollut* 2009;**198**:317–34. https://doi.org/10.1007/s11270-008-9848-z
7. Ansari AA, Gill SS, Khan FA. Eutrophication: threat to aquatic ecosystems. In: Chislock MF, Doster E, Zitomer RA *et al.* (eds.), *Eutrophication: causes, consequences and control*. Dordrecht: Springer, 2010, 143–70.
8. Wang D-Q, Yu Y-X, Zhang X-Y *et al.* Polycyclic aromatic hydrocarbons and organochlorine pesticides in fish from Taihu Lake: their levels, sources, and biomagnification. *Ecotoxicol Environ Saf* 2012;**82**:63–70. https://doi.org/10.1016/j.ecoenv.2012.05.010
9. Gupta S, Gupta K. Bioaccumulation of pesticides and its impact on biological systems. In:Srivastava PK, Singh VP, Singh A, Tripathi DK, Singh S, Prasad SM, Chauhan DK (eds.), *Pesticides in crop production*. Chichester, UK: Wiley, 55–67. 2020.
10. Xu X, Nie S, Ding H *et al.* Environmental pollution and kidney diseases. *Nat Rev Nephrol* 2018;**14**:313–24. https://doi.org/10.1038/nrneph.2018.11
11. Coresh J, Selvin E, Stevens LA *et al.* Prevalence of chronic kidney disease in the United States. *JAMA* 2007;**298**:2038. https://doi.org/10.1001/jama.298.17.2038
12. Xie Y, Bowe B, Li T *et al.* Long-term kidney outcomes among users of proton pump inhibitors without intervening acute kidney injury. *Kidney Int* 2017;**91**:1482–94. https://doi.org/10.1016/j.kint.2016.12.021
13. Barrios-Estrada C, de J, Rostro-Alanis M *et al.* Emergent contaminants: endocrine disruptors and their laccase-assisted degradation—a review. *Sci Total Environ* 2018;**612**:1516–31. https://doi.org/10.1016/j.scitotenv.2017.09.013
14. Kumar V, Majumdar C, Roy P. Effects of endocrine disrupting chemicals from leather industry effluents on male reproductive system. *J Steroid Biochem Mol Biol* 2008;**111**:208–16. https://doi.org/10.1016/j.jsbmb.2008.06.005
15. Nunes G, Barceló D. Analysis of carbamate insecticides in foodstuffs using chromatography and immunoassay techniques. *Trends Anal Chem* 1999;**18**:99–107. https://doi.org/10.1016/S0165-9936(98)00076-4
16. Bonfiglio R, Scimeca M, Mauriello A. The impact of environmental pollution on cancer: risk mitigation strategies to consider. *Sci Total Environ* 2023;**902**:166219. https://doi.org/10.1016/j.scitotenv.2023.166219
17. Ragusa A, Svelato A, Santacroce C *et al.* Plasticenta: first evidence of microplastics in human placenta. *Environ Int* 2021;**146**:106274. https://doi.org/10.1016/j.envint.2020.106274
18. Vethaak AD, Legler J. Microplastics and human health. *Science* 2021;**371**:672–4. https://doi.org/10.1126/science.abe5041
19. Kim D-W, Ahn J-H, Cha C-J. Biodegradation of plastics: mining of plastic-degrading microorganisms and enzymes using metagenomics approaches. *J Microbiol* 2022;**60**:969–76. https://doi.org/10.1007/s12275-022-2313-7
20. Muthusaravanan S, Sivarajasekar N, Vivek JS *et al.* Phytoremediation of heavy metals: mechanisms, methods and enhancements. *Environ Chem Lett* 2018;**16**:1339–59. https://doi.org/10.1007/s10311-018-0762-3
21. Rai PK. Heavy metal pollution in aquatic ecosystems and its phytoremediation using wetland plants: an ecosustainable approach. *Int J Phytoremediation* 2008;**10**:133–60.
22. Sajjaphan K, Shapir N, Wackett LP *et al.* *Arthrobacter aurescens* TC1 atrazine catabolism genes trzN, atzB, and atzC are linked on a 160-kilobase region and are functional in *Escherichia coli*. *Appl Environ Microbiol* 2004;**70**:4402–7.
23. Zhang Y, Jiang Z, Cao B *et al.* Chemotaxis to atrazine and detection of a xenobiotic catabolic plasmid in *Arthrobacter* sp. DNS10. *Environ Sci Pollut Res* 2012;**19**:2951–8. https://doi.org/10.1007/s11356-012-0805-4
24. Guo X, Xie C, Wang L *et al.* Biodegradation of persistent environmental pollutants by *Arthrobacter* sp. *Environ Sci Pollut Res* 2019;**26**:8429–43. https://doi.org/10.1007/s11356-019-04358-0
25. Dash HR, Mangwani N, Chakraborty J *et al.* Marine bacteria: potential candidates for enhanced bioremediation. *Appl Microbiol*

Biotechnol 2013;**97**:561–71.
https://doi.org/10.1007/s00253-012-4584-0

26. Teramoto M, Suzuki M, Okazaki F *et al.* Oceanobacter-related bacteria are important for the degradation of petroleum aliphatic hydrocarbons in the tropical marine environment. *Microbiology* 2009;**155**:3362–70. https://doi.org/10.1099/mic.0.030411-0
27. Samanta SK, Singh OV, Jain RK. Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation. *Trends Biotechnol* 2002;**20**:243–8. https://doi.org/10.1016/S0167-7799(02)01943-1
28. Chang Q, Jiang G, Tang H *et al.* Enzymatic removal of chlorophenols using horseradish peroxidase immobilized on superparamagnetic Fe_3O_4/graphene oxide nanocomposite. *Chin J Catal* 2015;**36**:961–8. https://doi.org/10.1016/S1872-2067(15)60856-7
29. Bilal M, Adeel M, Rasheed T *et al.* Emerging contaminants of high concern and their enzyme-assisted biodegradation—a review. *Environ Int* 2019;**124**:336–53. https://doi.org/10.1016/j.envint.2019.01.011
30. Zdarta J, Meyer A, Jesionowski T *et al.* A general overview of support materials for enzyme immobilization: characteristics, properties, practical utility. *Catalysts* 2018;**8**:92. https://doi.org/10.3390/catal8020092
31. Chouhan S, Tuteja U, Flora SJS. Isolation, identification and characterization of fluoride resistant bacteria: possible role in bioremediation. *Appl Biochem Microbiol* 2012;**48**:43–50. https://doi.org/10.1134/S0003683812010036
32. Ufarté L, Laville E, Duquesne S *et al.* Discovery of carbamate degrading enzymes by functional metagenomics. *PLoS One* 2017;**12**:e0189201. https://doi.org/10.1371/journal.pone.0189201
33. Danso D, Schmeisser C, Chow J *et al.* New insights into the function and global distribution of polyethylene terephthalate (PET)-degrading bacteria and enzymes in marine and terrestrial metagenomes. *Appl Environ Microb* 2018;**84**:e02773-17. https://doi.org/10.1128/AEM.02773-17
34. Cornish-Bowden A. Current IUBMB recommendations on enzyme nomenclature and kinetics. *Perspect Sci* 2014;**1**:74–87. https://doi.org/10.1016/j.pisc.2014.02.006
35. Gambarini V, Pantos O, Kingsbury JM *et al.* PlasticDB: a database of microorganisms and proteins linked to plastic biodegradation. *Database* 2022;**2022**:baac008. https://doi.org/10.1093/database/baac008
36. Gao J, Ellis LBM, Wackett LP. The University of Minnesota Biocatalysis/Biodegradation Database: improving public access. *Nucleic Acids Res* 2010;**38**:D488–91. https://doi.org/10.1093/nar/gkp771
37. Sankara Subramanian SH, Balachandran KRS, Rangamaran VR *et al.* RemeDB: tool for rapid prediction of enzymes involved in bioremediation from high-throughput metagenome data sets. *J Comput Biol* 2020;**27**:1020–9. https://doi.org/10.1089/cmb.2019.0345
38. Oliveira JS, Araújo W, Sales L *et al.* BioSurfDB: knowledge and algorithms to support biosurfactants and biodegradation studies. *Database* 2015;**2015**:bav033. https://doi.org/10.1093/database/bav033
39. Scheer M, Grote A, Chang A *et al.* BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* 2011;**39**:D670–6. https://doi.org/10.1093/nar/gkq1089
40. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30. https://doi.org/10.1093/nar/28.1.27
41. Gan Z, Zhang H. PMBD: a comprehensive plastics microbial biodegradation database. *Database* 2019;**2019**:baz119. https://doi.org/10.1093/database/baz119
42. Kim S, Chen J, Cheng T *et al.* PubChem 2023 update. *Nucleic Acids Res* 2023;**51**:D1373–80. https://doi.org/10.1093/nar/gkac956
43. Degtyarenko K, de Matos P, Ennis M *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2007;**36**:D344–50. https://doi.org/10.1093/nar/gkm791
44. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 2015;**7**:20. https://doi.org/10.1186/s13321-015-0069-3
45. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 2021;**38**:5825–9. https://doi.org/10.1093/molbev/msab293
46. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 2011;**32**:1466–74. https://doi.org/10.1002/jcc.21707
47. Malwe AS, Srivastava GN, Sharma VK. GutBug: a tool for prediction of human gut bacteria mediated biotransformation of biotic and xenobiotic molecules using machine learning. *J Mol Biol* 2023;**435**: 168056. https://doi.org/10.1016/j.jmb.2023.168056
48. Sharma AK, Jaiswal SK, Chaudhary N *et al.* A novel approach for the prediction of species-specific biotransformation of xenobiotic/drug molecules by the human gut microbiota. *Sci Rep* 2017;**7**:9751. https://doi.org/10.1038/s41598-017-10203-6
49. Srivastava GN, Malwe AS, Sharma AK *et al.* Molib: a machine learning based classification tool for the prediction of biofilm inhibitory molecules. *Genomics* 2020;**112**:2823–32. https://doi.org/10.1016/j.ygeno.2020.03.020
50. Cheng F, Li W, Zhou Y *et al.* admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model* 2012;**52**:3099–105. https://doi.org/10.1021/ci300367a
51. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Soft* 2010;**36**:1–13. https://doi.org/10.18637/jss.v036.i11
52. Malwe AS, Sharma VK. Application of artificial intelligence approaches to predict the metabolism of xenobiotic molecules by human gut microbiome. *Front Microbiol* 2023;**14**:1254073. https://doi.org/10.3389/fmicb.2023.1254073
53. Pushpa M, Karpagavalli S. Multi-label classification: problem transformation methods in Tamil Phoneme classification. *Procedia Comput Sci* 2017;**115**:572–9. https://doi.org/10.1016/j.procs.2017.09.116
54. Wu G, Zhu J. Multi-label classification: do hamming loss and subset accuracy really conflict with each other? arXiv, https://arxiv.org/abs/2011.07805, 16 November 2020, preprint: not peer reviewed.
55. Tsoumakas G, Katakis I, Vlahavas I. A review of multi-label classification methods. *Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006)*.
56. Sechidis K, Tsoumakas G, Vlahavas I. On the stratification of multi-label data. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M (eds.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science*. Vol. **6913**. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23808-6_10
57. Cutler A, Cutler DR, Stevens JR. Random Forests. In: *Ensemble Machine Learning*. New York, NY: Springer New York, 2012, 157–75.
58. Krogh A. What are artificial neural networks? *Nat Biotechnol* 2008;**26**:195–7. https://doi.org/10.1038/nbt1386
59. Jaiswal SK, Agarwal SM, Thodum P *et al.* SkinBug: an artificial intelligence approach to predict human skin microbiome-mediated metabolism of biotics and xenobiotics. *iScience* 2021;**24**:101925. https://doi.org/10.1016/j.isci.2020.101925
60. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of Enzyme Commission numbers. *Proc Natl Acad Sci USA* 2019;**116**:13996–4001. https://doi.org/10.1073/pnas.1821905116

61. Ahmad GN, Fatima H, Ullah S *et al.* Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. *IEEE Access* 2022;**10**:80151–73. https://doi.org/10.1109/ACCESS.2022.3165792
62. Ali J, Khan R, Ahmad N *et al.* Random forests and decision trees. *IJCSI* 2012;**9**:272.
63. Wu X, Gao Y, Jiao D. Multi-label classification based on random forest algorithm for non-intrusive load monitoring system. *Processes* 2019;**7**:337. https://doi.org/10.3390/pr7060337
64. Huerta-Cepas J, Szklarczyk D, Heller D *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;**47**:D309–14. https://doi.org/10.1093/nar/gky1085
65. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.
66. Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological profiles for xylene. Atlanta, GA: US Department of Health and Human Services. 2007.
67. Miri S, Davoodi SM, Robert T *et al.* Enzymatic biodegradation of highly p-xylene contaminated soil using cold-active enzymes: a soil column study. *J Hazard Mater* 2022;**423**:127099.
68. Yang F-W, Li Y-X, Ren F-Z *et al.* Toxicity, residue, degradation and detection methods of the insecticide triazophos. *Environ Chem Lett* 2019;**17**:1769–85.
69. Li Y, Chen Q, Wang C-H *et al.* Degradation of acetochlor by consortium of two bacterial strains and cloning of a novel amidase gene involved in acetochlor-degrading pathway. *Bioresour Technol* 2013;**148**:628–31.
70. Liu L, Bilal M, Duan X *et al.* Mitigation of environmental pollution by genetically engineered bacteria—current challenges and future perspectives. *Sci Total Environ* 2019;**667**:444–54.
71. Singh NS, Sharma R, Singh DK. Identification of enzyme(s) capable of degrading endosulfan and endosulfan sulfate using in silico techniques. *Enzyme Microb Technol* 2019;**124**:32–40.
72. Sutherland TD, Horne I, Harcourt RL *et al.* Isolation and characterization of a *Mycobacterium* strain that metabolizes the insecticide endosulfan. *J Appl Microbiol* 2002;**93**:380–9. https://doi.org/10.1046/j.1365-2672.2002.01728.x
73. Quatrini P, Scaglione G, De Pasquale C *et al.* Isolation of Gram-positive n-alkane degraders from a hydrocarbon-contaminated Mediterranean shoreline. *J Appl Microbiol* 2007;**104**;251–9. https://doi.org/10.1111/j.1365-2672.2007.03544.x
74. Dai M, Copley SD. Genome shuffling improves degradation of the anthropogenic pesticide pentachlorophenol by *Sphingobium chlorophenolicum* ATCC 39723. *Appl Environ Microb* 2004;**70**:2391–7. https://doi.org/10.1128/AEM.70.4.2391-2397.2004
75. Neumann G, Teras R, Monson L *et al.* Simultaneous degradation of atrazine and phenol by *Pseudomonas* sp. strain ADP: effects of toxicity and adaptation. *Appl Environ Microb* 2004;**70**:1907–12.
76. Camacho-Pérez B, Ríos-Leal E, Rinderknecht-Seijas N *et al.* Enzymes involved in the biodegradation of hexachlorocyclohexane: a4 mini review. *J Environ Manage* 2012;**95**:S306–18.
77. Waigi MG, Kang F, Goikavi C *et al.* Phenanthrene biodegradation by sphingomonads and its application in the contaminated soils and sediments: a review. *Int Biodeter Biodegrad* 2015;**104**:333–49.
78. Schuler L, Ní Chadhain SM, Jouanneau Y *et al.* Characterization of a novel angular dioxygenase from fluorene-degrading *Sphingomonas* sp. strain LB126. *Appl Environ Microb* 2008;**74**:1050–7.
79. Hu D, Wu J, Fan L *et al.* Aerobic degradation characteristics and mechanism of decabromodiphenyl ether (BDE-209) using complex bacteria communities. *Int J Environ Res Public Health* 2022;**19**:17012. https://doi.org/10.3390/ijerph192417012
80. Bonefeld Jorgensen E, Autrup H, Hansen JC. Effect of toxaphene on estrogen receptor functions in human breast cancer cells. *Carcinogenesis* 1997;**18**:1651–4. https://doi.org/10.1093/carcin/18.8.1651
81. Prieto I, Klimm A, Roldán F *et al.* Evidence for cometabolic transformation of weathered toxaphene under aerobic conditions using camphor as a co-substrate. *J Appl Microbiol* 2021;**131**:221–35. https://doi.org/10.1111/jam.14963
82. Romero ML, Terrazas E, van Bavel B *et al.* Degradation of toxaphene by *Bjerkandera* sp. strain BOL13 using waste biomass as a cosubstrate. *Appl Microbiol Biotechnol* 2006;**71**:549–54. https://doi.org/10.1007/s00253-005-0174-8
83. Sakakibara F, Takagi K, Kataoka R *et al.* Isolation and identification of dieldrin-degrading *Pseudonocardia* sp. strain KSF27 using a soil–charcoal perfusion method with aldrin trans-diol as a structural analog of dieldrin. *Biochem Biophys Res Commun* 2011;**411**:76–81. https://doi.org/10.1016/j.bbrc.2011.06.096
84. Pang S, Lin Z, Li J *et al.* Microbial degradation of aldrin and dieldrin: mechanisms and biochemical pathways. *Front Microbiol* 2022;**13**:713375. https://doi.org/10.3389/fmicb.2022.713375
85. Mansouri A, Cregut M, Abbes C *et al.* The environmental issues of DDT pollution and bioremediation: a multidisciplinary review. *Appl Biochem Biotechnol* 2017;**181**:309–39. https://doi.org/10.1007/s12010-016-2214-5
86. Azubuike CC, Chikere CB, Okpokwasili GC. Bioremediation techniques—classification based on site of application: principles, advantages, limitations and prospects. *World J Microbiol Biotechnol* 2016;**32**:180. https://doi.org/10.1007/s11274-016-2137-x
87. Sharma R, Saghapour E, Chen JY. An NLP-based technique to extract meaningful features from drug SMILES. *iScience* 2024;**27**:109127. https://doi.org/10.1016/j.isci.2024.109127

Received: September 28, 2024. **Revised:** March 9, 2025. **Editorial Decision:** March 18, 2025. **Accepted:** April 28, 2025