Deep Learning Enables Rapid Identification of Mycotoxin-Degrading Enzymes

Dachuan Zhang[1, 2, †], Huadong Xing[1, †], Dongliang Liu[1, †], Mengying Han[1], Pengli Cai[1], Huikang Lin[1], Yu Tian[3], Yinghao Guo[4], Bin Sun[4], Ye Tian[1, *], Aibo Wu[1, *] and Qian-Nan Hu[1, *]

[1] CAS Key Laboratory of Computational Biology, CAS Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2] Institute of Environmental Engineering, ETH Zurich, Zurich 8093, Switzerland

[3] School of Biology and Pharmaceutical Engineering, Wuhan Polytechnic University, Wuhan 430023, China

[4] Department of Pharmacology, Harbin Medical University, Harbin 150081, China

[†] Contributed equally

* Qian-Nan Hu

Email: qnhu@sibs.ac.cn

Author Contributions: D.Z. and Q-N.H. designed the research. D.Z., H.X., and D.L. developed the deep learning model. Y.T., H.L., and A.W. performed the experiments. M.H., Y.T., and P.C. collected the enzyme data. Y.G. and B.S. performed molecular dynamics simulations. D.Z., Y.T., H.X., and Q-N.H. wrote the paper. All authors approved the final paper.

Competing Interest Statement: The authors declare no competing interests.

Keywords: synthetic biology, biotransformation, machine learning, food safety, mycotoxins.

Abstract

The identification of functional enzymes for the catalysis of specific biochemical reactions is a major bottleneck in the de novo design of biosynthesis and biodegradation pathways. Conventional methods based on microbial screening and functional metagenomics require long verification periods and incur high experimental costs; recent data-driven methods are only applicable to a few common substrates. To enable rapid and high-throughput identification of enzymes for complex and less-studied substrates, we propose a robust enzyme promiscuity prediction model based on positive unlabeled learning, which shortens the time needed for new enzyme discovery from several years to 29 days. Using this model, we identified 15 new degrading enzymes specific for the mycotoxins ochratoxin A and zearalenone, of which six could degrade > 90% mycotoxin content within 3 h. We anticipate that this model will become indispensable for the identification of new functional enzymes, thereby advancing the fields of synthetic biology, metabolic engineering, and pollutant biodegradation.

Introduction

Mycotoxins are toxic secondary fungal metabolites that frequently contaminate food and feed and adversely affect animal welfare and productivity [1]. Mycotoxins in food and livestock products can also be transferred to human consumers, compromising public health [2]. At a time when the global production of agricultural commodities is barely sustaining the growing population, the Food and Agriculture Organization of the United Nations has estimated that at least 25% of the world's food crops are ruined annually due to mycotoxins, and global investigations revealed that mycotoxins were detected in ~70% of crop samples [3-5]. With climate change and an increased frequency of extreme weather events, the levels of mycotoxin contamination are expected to increase in the future [6-9].

Enzymatic elimination of mycotoxins is considered a promising method owing to its specificity, safety, and environmental friendliness [10]. However, effective mycotoxin-degrading enzymes are rare, owing to economic limitations and the potential toxicity of metabolic products [11,12]. In a previous study, we predicted > 550,000 enzymatic reactions involving biogenic toxins based on reaction rules [13]. Although these reactions could potentially eliminate mycotoxins, ~94.7% of these have not been experimentally verified owing to the long verification period and high experimental costs of traditional enzyme-mining methods based on microbial screening and functional metagenomics [13]. To date, 4×10^9 proteins have been sequenced, and with the rapid development of proteomics, the number of known proteins doubles every 24 months [14,15]. A framework for the rapid and high-throughput prediction and verification of new enzymes with desired activities, e.g., mycotoxin degradation, is urgently required.

In recent years, the traditional view that enzymes have strict substrate specificity has been replaced by the promiscuity theory [16]. Enzyme promiscuity is defined as the capability of enzymes to catalyze other substrates besides their natural substrate [17]. More than 37% of enzymes in *E. coli* are promiscuous, and these enzymes catalyze >65% of the known metabolic reactions in *E. coli* [18]. This finding inspired us to establish a prediction model for enzyme promiscuity and use it to screen enzymes that can specifically degrade mycotoxins. A few methods have recently been proposed to predict enzyme promiscuity based on the similarity theory, assuming the substrates catalyzed by an enzyme generally possess similar structural characteristics [19]. However, being limited by poor generalization abilities, these methods cannot yield reliable results for complex and less-studied molecules, such as mycotoxin ochratoxin A (OTA) and zearalenone (ZEA) (Fig S1). Another strategy used to identify functional enzymes for the catalysis of a specific biochemical reaction is predicting enzyme commission (EC) numbers based on sequence characteristics [20]; however, this strategy only applies to common reactions involved in the EC system. For complex molecules and less-studied reactions, no effective computational method exists to identify their catalytic enzymes. The feasibility of enzymatic reactions is determined by complex interactions between enzymes and substrates. Hence, an enzyme's substrate promiscuity is more suitably predicted based on the characteristics of both the enzyme and substrate [21]. Many active enzyme–substrate pairs have been annotated in bioinformatics databases [22-27]. However, owing to the lack of negative samples (e.g., a certain enzyme that cannot catalyze a certain substrate), the feasibility of predicting enzyme promiscuity based on both enzyme and substrate characteristics remains tenuous. Moreover, although data-driven models enable the rapid prediction of potential enzymes, enzyme screening is a slow process owing to the long experimental period of traditional verification systems.

In this study, we aimed to resolve the obstacles of low efficiency and accuracy in traditional enzyme-mining methods with a framework that combines data-driven prediction models with a rapid cell-free protein expression (CFPE) system (Fig 1). We showcased the feasibility and efficiency of

the proposed framework with two case studies, identification of new enzymes for ZEA and OTA degradation. Using the framework, we successfully found 15 new enzymes with prominent degradation activity within 29 d and identified critical residues on enzymes based only on sequence-level information without any prior knowledge. Based on our findings, we anticipate that this framework will serve as an indispensable tool for the identification of new functional enzymes.

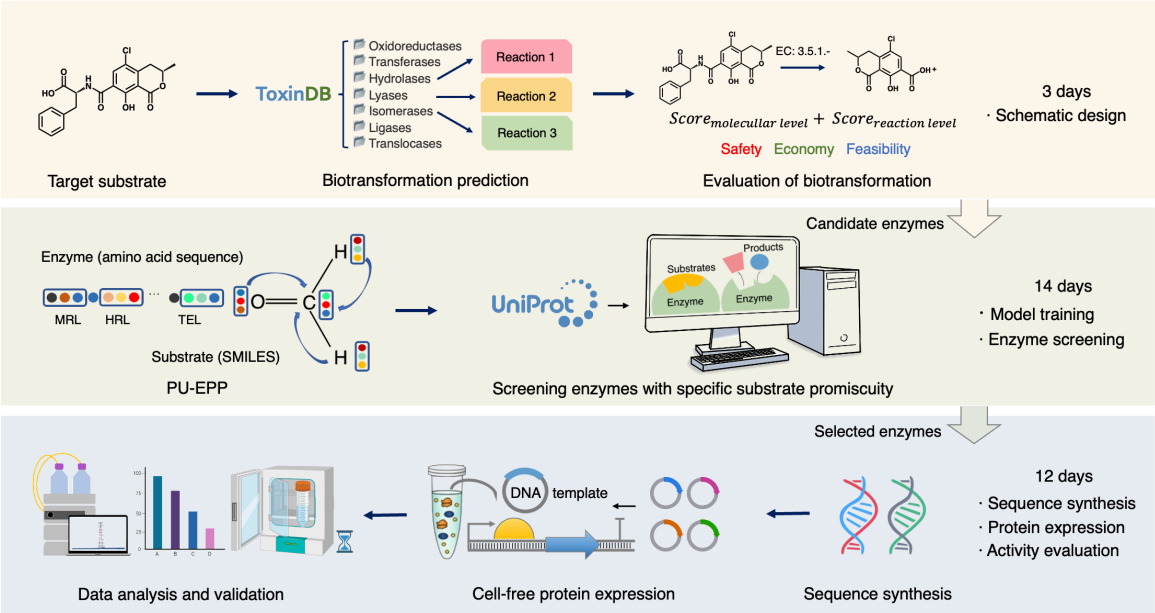

Figure 1. Discovery of new functional enzymes based on the positive unlabeled learning-based enzyme promiscuity prediction (PU-EPP) model and cell-free protein expression. The framework mainly consists of three parts: schematic design (~3 d), enzyme screening (~14 d), and experimental verification (~12 d). First, potential biotransformation reactions and metabolic products of the target substrate (e.g., mycotoxins) were predicted using reaction rules in ToxinDB. The potential biotransformations were evaluated with regards to safety, economy, and feasibility. Next, optimal biotransformations and the corresponding categories of enzymes were selected based on enzyme commission (EC) numbers for downstream screening. The molecular structure of mycotoxins formatted in the simplified molecular input line entry system (SMILES) and the amino acid sequences of candidate enzymes were analyzed using the positive unlabeled (PU) learning-based enzyme promiscuity prediction (PU-EPP) model to screen enzymes that could specifically catalyze the biotransformation. Finally, selected enzymes were expressed using a CFPE system and assayed for their catalytic activity.

Results

The positive unlabeled learning-based enzyme promiscuity prediction (PU-EPP) model. Deep learning relies on large-scale data to make high-quality predictions. To afford a comprehensive dataset on enzyme promiscuity, we integrated data from the Rhea [22], Kyoto Encyclopedia of Genes and Genomes (KEGG) [23], MetaCyc [24], Brenda [25], and RxnFinder [26,27] databases and acquired 170,179 enzymes, 5,837 substrates, and 606,555 corresponding enzyme–substrate pairs. Since negative data are rarely reported, previous studies usually augmented negative samples for training classification models based on activity thresholds or random sampling [28,29]. However, low activity is not equivalent to inactivity, and certain randomly sampled enzyme–substrate pairs may actually be functional (positive) because of enzyme promiscuity, which would result in the deep learning models acquiring inaccurate results. Therefore, we proposed a strategy that combines weighted random sampling and positive unlabeled (PU) learning to minimize the impact of inaccurate negative samples (see the methods section for detail). The weighted random sampling strategy assumes that enzymes with fewer reported substrates have higher catalytic specificity. Therefore, to reflect both enzyme promiscuity and specificity, more corresponding negative enzyme–substrate pairs should be included in the training set. Accordingly, we generated 6,488,914 "negative" samples with a ratio of ~10 times the positive samples (Fig 2A and S2). However, since some enzymes may be promiscuous, some positive samples might remain in the generated negative samples. If these cases are ignored, the model would ignore the enzyme's promiscuity. Therefore, to ensure that the model accurately predicts enzyme behavior, we considered these generated samples as unlabeled data and proposed a PU learning-based enzyme promiscuity prediction (PU-EPP) model. This model employs a PU learning mechanism during training to exclude potential positive samples in the unlabeled data according to the probability ranges of positive samples (Fig 2B) [30]. We chose the 90th percentile of the probability distribution of positive samples as the threshold because it presents both good performance and robust recognition ability of potential positive samples (Fig S3 and S4). The PU-EPP model was built based on a multi-head attention mechanism (Fig 2C) [31]. The substrate features were extracted through graph neural networks, while the protein sequence was encoded using a Continuous Bag-of-Words [32-34].

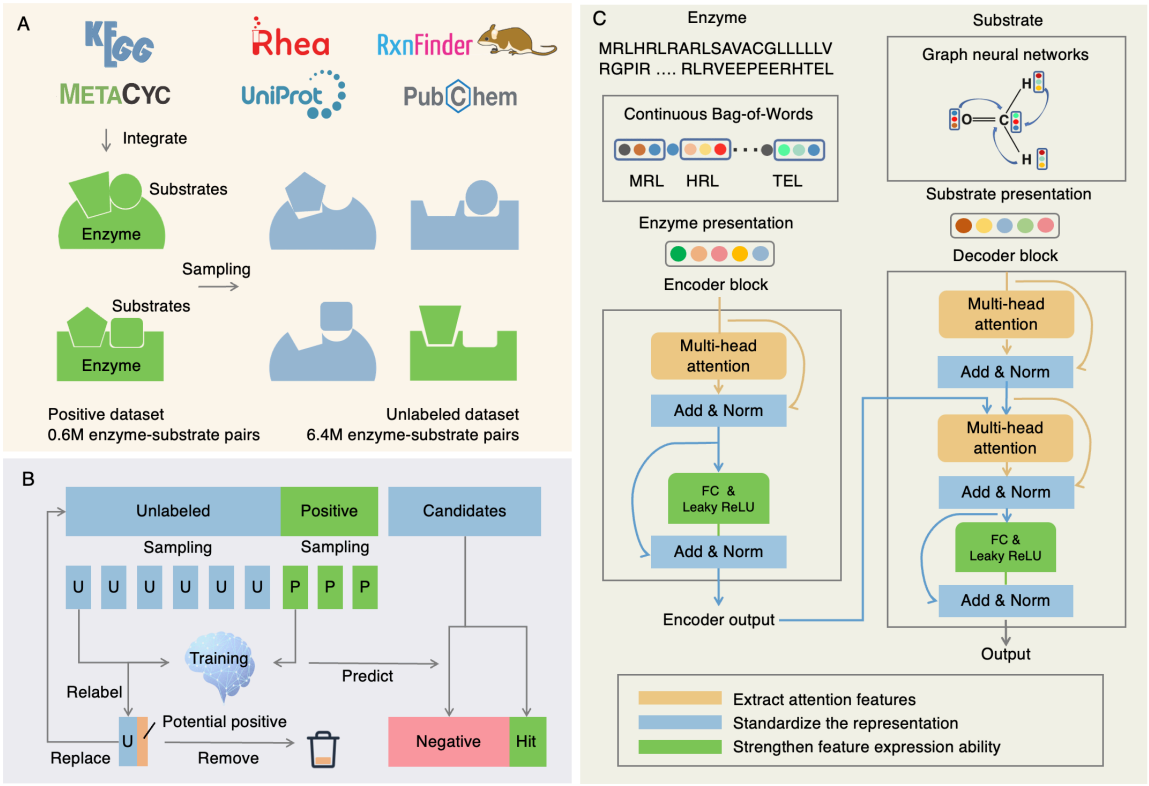

Figure 2. Framework of the PU-EPP model. (A) Known enzyme–substrate pairs were collected from Rhea, KEGG, MetaCyc, and RxnFinder. The molecular structures of substrates were collected from PubChem, while the amino acid sequences of enzymes were collected from UniProt. Afterward, unlabeled enzyme–substrate pairs were obtained through weighted random sampling of known (positive) enzyme–substrate pairs. (B) Framework of PU learning. For each epoch, equal numbers of unlabeled and positive samples were sampled and input into the model. Then, the probability ranges of positive samples in each epoch were calculated, which were used for removing potential positive samples in the unlabeled dataset to avoid the model learning inaccurate knowledge. The potential positive samples were removed from the dataset while others were put back. These processes were repeated until the end of the training. The final model was used to screen potential enzymes (hit) for mycotoxin elimination. (C) Detailed implementation of PU-EPP. The enzyme sequence was encoded with the Continuous Bag-of-Words model, while substrate features were extracted from the molecular structure of substrates with graph neural networks. Extracted enzyme features were input into the encoder. Extracted substrate features were processed through a self-attention layer and then combined with enzyme features. Then, the combined features were input into another self-attention layer (interaction layer) to extract interaction information between the substrates and the enzymes. Finally, the features were input into a fully connected layer and a layer normalization step to achieve the final output by softmax. FC: fully connected layer; Add: residual connection; Norm: layer normalization.

Recovery experiments demonstrate the strong robustness of PU-EPP. An independent test set consisting of 20,000 enzyme-substrate pairs was used for model evaluation, among which >60% of pairs had not been included in the EC system. The final PU-EPP model showed remarkable performance with a receiver operating characteristic area under the curve (ROC-AUC)

of 0.985 and precision recall curve (PRC)-AUC of 0.988, indicating it can achieve reliable predictions on both well-studied and less-studied enzyme-substrate pairs. On the contrary, previous models for enzyme function prediction, e.g., EPP-HMCNF [17] and CLEAN [20], can only apply to hundreds of common substrates involved in the EC system. PU-EPP's wide applicability domain enables it to serve as a more useful tool for the identification of new enzymes with specific catalytic activities. To evaluate the improvement of the new proposed ML framework, we also compared the PU-EPP with ML models in the drug discovery field that utilize both protein and compound features [28,33,34]. We re-trained and tested PU-EPP and other models on a dataset comprising 20,000 positive samples and ~10 times the number of unlabeled samples. PU-EPP showed the best performance, followed by TransformerCPI [28], DeepConv-DTI [33], and GNN-PT [34] (Fig 3A). We also tested the performance of PU-EPP on five independent small-scale datasets collected by Goldman et al., including data on phosphatase, esterase, glycosyltransferase, β-keto acid cleavage enzymes, and halogenase, that comprised real positive and negative enzyme–substrate pairs (Table S1) [35]. We then compared the performance of PU-EPP with the best models reported in their study [35]. PU-EPP also showed better performance on each dataset, illustrating its superiority (Fig 3B).

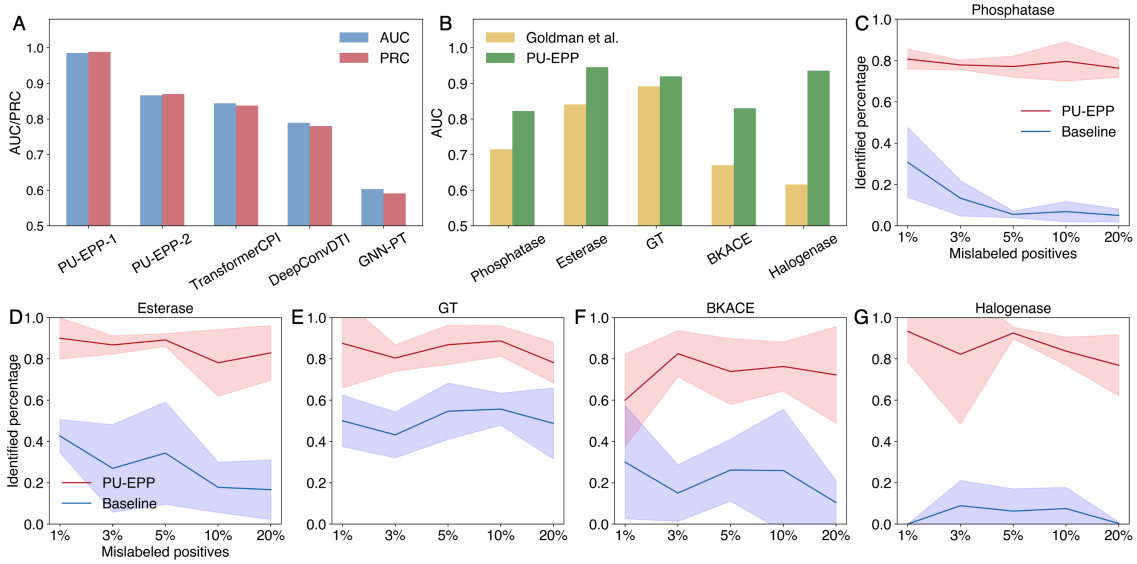

Figure 3. Evaluation of the models' performance. (A) Comparison between PU-EPP and previously published models. PU-EPP-1 is the model trained on whole datasets, while PU-EPP-2 is the model trained on the dataset consisting of 20,000 positive samples and corresponding unlabeled samples. (B) The performance of PU-EPP and previously published models on five independent datasets related to phosphatase, esterase, glycosyltransferase (GT), β-keto acid cleavage enzymes (BKACE), and halogenase. (C-G) Recovery experiments of mislabeled positive samples on these five datasets. The light area presents the ranges of the ± s.d of the five-fold cross-validation. Here, 1%, 3%, 5%, 10%, and 20% of positive samples were added to the unlabeled samples. PU-EPP successfully identified ~80% of mislabeled positive samples, which presented better robustness and anti-interference ability than the baseline models.

To assess whether the PU learning strategy we proposed can identify potential positive samples from unlabeled samples, thereby improving the model's robustness, we conducted a recovery experiment of hidden positive samples. To this end, 1%, 3%, 5%, 10%, and 20% of the positive samples were mislabeled as negative, and the recognition capability of PU-EPP and a

baseline model (models trained by conventional procedure) on these mislabeled samples was tested by five-fold cross-validation. Through the PU learning strategy, PU-EPP identified ~80% of mislabeled positive samples in the training process, while the baseline model trained through the conventional procedure (non-PU) only identified between ~40% to 65% (Fig 3C–G). When 20% of the positive samples were mislabeled, PU-EPP still had a strong recognition capability, demonstrating its high robustness. We also evaluated the performance of the PU-EPP and the baseline model under each condition and found that PU-EPP also achieved higher AUC scores than the baseline model (Fig S5). Thus, when there are biases and noise in the training datasets, the PU learning strategy employed herein could be considered an effective approach to improve the model's performance and robustness.

PU-EPP successfully identified 15 new enzymes for mycotoxin detoxification. Two mycotoxins with high contamination levels, ZEA and OTA, were selected as case studies. The complex structural features of ZEA and OTA lead to various biotransformation possibilities under the action of enzymes. Therefore, we comprehensively predicted the potential toxin biotransformations of OTA and ZEA based on reaction rules in ToxinDB (Fig S6) [13]. To rationally select the optimal detoxification reactions for ZEA and OTA detoxification, we evaluated the feasibility, safety, and economics of these enzymatic biotransformations (Table S2) and finally selected lactone bond hydrolyses (EC 3.1.1.-) and amide bond hydrolyses (EC 3.5.1.-) for the detoxification of ZEA and OTA, respectively (Tables S3 and S4).

According to the probes predicted by PU-EPP, we selected 10 potential ZEA hydrolases (ZH 1-10) and 10 potential OTA hydrolases (OH 1-10) for downstream experiments. The candidate genes were synthesized and then expressed with a CFPE system that only required a few hours to synthesize one protein. Then, the obtained enzymes were incubated with ZEA and OTA for 3 h. Overall, 75% of predicted enzymes exhibited expected activities. Nine enzymes exhibited prominent catalytic activities toward ZEA, of which ZH4 eliminated >90% and ZH9 eliminated ~100% of the ZEA content (Fig 4A). Six enzymes exhibited obvious catalytic activity toward OTA. Among them, OH1, OH4, and OH6 eliminated >90% and OH8 eliminated nearly 100% of the OTA content in the reaction system (Fig 4B). To confirm that the enzymes did catalyze the expected degradation reactions, the metabolic products of ZEA and OTA degradation in the reaction system were measured using high-resolution mass spectrometry. A catalytic product of ZEA with an *m/z* [M-H]- 291.16 was detected in samples of ZH1-ZH9 (Fig 4D-G), while the catalytic product of OTA, ochratoxin α, with an *m/z* [M-H]- 255.0067, was detected in samples of OH1, OH2, OH4, OH6, OH8, and OH9 (Fig 4E and F), verifying the predicted degradation reactions.

We also tested the degradation ratio of mycotoxins contaminated in two food matrices (wheat and maize flour) after incubation with four enzymes (ZH4, ZH9, OH1, and OH8) with relatively higher degradation activities (Fig 4C). Despite a certain decrease in activity owing to the influence of the complex matrices, all the enzymes exhibited obvious activity. ZH9 eliminated 67% of ZEA in wheat flour and 24 % in maize flour, while OH8 eliminated 62% of OTA in wheat flour and 33% in maize flour. These results indicate the feasibility of using enzymes to eliminate mycotoxins contaminating foodstuffs. It is worth noting that we only tested the degradation effects of low-dose natural enzymes. With the structural modification of enzymes and engineering methods, such as immobilization, the degradation ratio of mycotoxin contaminants in food matrices will be further improved.

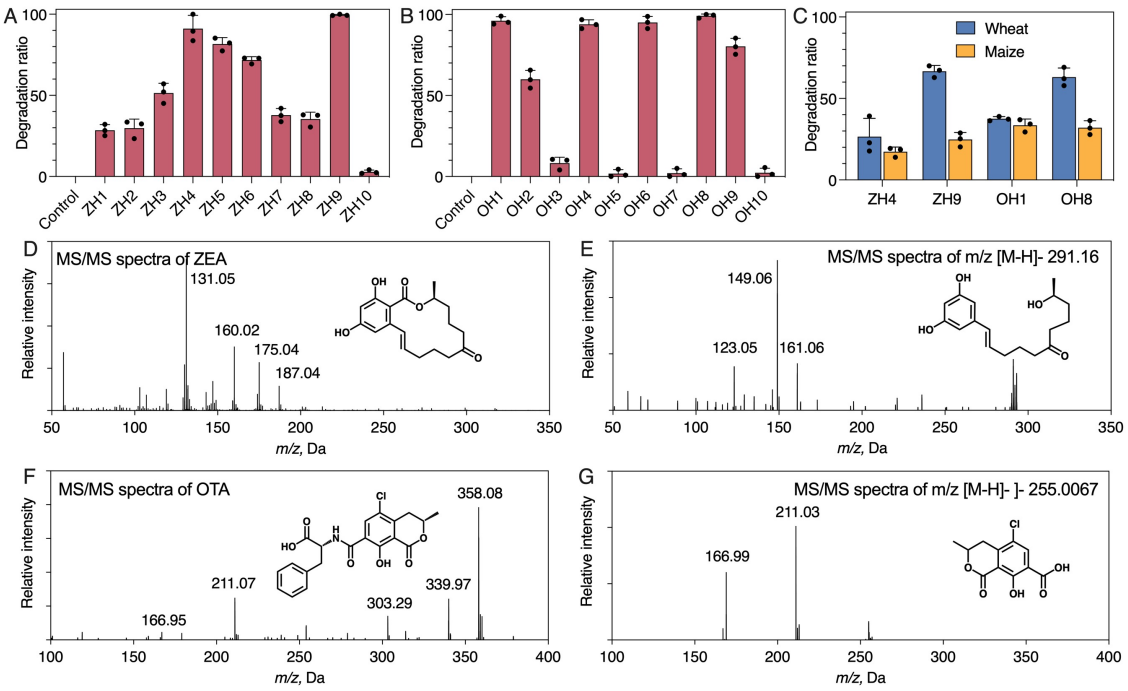

Figure 4. Experimental verification of the catalytic activity of candidate enzymes. (A, B) The degradation ratio of candidate enzymes on mycotoxins ZEA and OTA was assayed by measuring the residual concentration of mycotoxins in the reaction system. After incubation for 3 h at 37 °C, the content of mycotoxins and their metabolites were measured using liquid chromatography–mass spectrometry (LC-MS). (C) The degradation ratio of enzymes for mycotoxin contaminants in wheat and maize flour. Enzyme supernatant was mixed with mycotoxin-contaminated flour. After incubation under the same conditions, the content of mycotoxins was tested using LC-MS. (D–G) The MS/MS spectra of ZEA and OTA and their degradation products. Error bars represent ± s.d.

PU-EPP successfully identified critical sites on substrates and enzymes. Despite successful application in many areas, interpretability remains a challenge for deep learning-based methods [28]. PU-EPP uses attention mechanisms [31] to capture the importance and contributions of different input positions (e.g., atoms in substrates or residues in enzymes) to the final prediction, thereby inferring the knowledge the model has learned (Fig 5A and B). To verify whether PU-EPP was able to learn the catalytic mechanism behind the data, we mapped the attention weights to the substrates and enzymes and compared them with known data. Enzymatic catalysis lies in the enzyme's recognition of reaction sites on substrates. To evaluate the model's recognition ability for the reaction sites on substrates, we compared the attention weights with annotated reaction site data and found that ~60% of reaction sites were correctly identified, which was significantly higher than the results of random sampling (Fig 5G and H). For instance, in screening tasks, the model successfully identified the oxygen atom near the reaction sites in ZEA and the nitrogen atom in OTA, which both are top-ranked in all atoms in ZEA and OTA (Fig 5C and D).

Since enzymes are complexes composed of hundreds of amino acid residues, identifying critical residues in enzymes is more challenging. Hence, we took ZH4 and OH1 as examples and analyzed their critical residues using molecular dynamics (MD) simulations and compared them with the important residues inferred through attention mechanisms. The most likely enzyme–

substrate binding pose was predicted via systematic docking plus extensive MD simulations. Energetically critical residues that contribute significantly to substrate binding free energy were identified via the molecular mechanics-generalized Born and surface area (MM-GBSA) methods (Fig S7–S9). We found that the attention weights of the interaction layer that has learned both enzyme and substrate features could map with the binding sites well, while the attention weights of the encoder that only learned enzyme features could not (Fig S10). This indicates that both enzyme and substrate features are crucial for prediction, which validated our initial methodological assumptions. It also demonstrates that although the model only takes enzyme sequences as input, it learned the hidden three-dimensional structural information (i.e., sites at which the substrate might bind). Furthermore, allosterically critical residues were calculated using root mean square fluctuation (RMSF) and dynamics cross-correlation map analysis (Fig S11 and S12). The consistency of important residues identified via attention mechanisms and MD (~90% on ZH4 and ~40% on OH1 were consistent) was significantly higher than that between MD and randomly selected residues (Fig 5E, 5F, and Fig S13–S15), indicating that PU-EPP learned the meaningful enzymatic knowledge behind the data.

To further test the reliability of identified residue sites, we selected eight high-attention sites on enzymes and performed single-site mutation verification. We mutated two proximal sites on each enzyme to alanine and found that, in most cases, enzyme activity significantly decreased after the mutation, indicating that these residues play important roles in catalysis (Fig 5I and J). Furthermore, the OH1_348 site was only identified by PU-EPP, suggesting that PU-EPP could be considered a supplement to MD for the identification of potentially important residues. Subsequently, we mutated two top-ranked distal sites on each enzyme to alanine and found that the enzymes' activity also decreased significantly, indicating that PU-EPP can identify distal residues that play important roles during catalytic processes, although most of them are generally considered to have no significant effect on catalytic activity (Fig 5I and J).

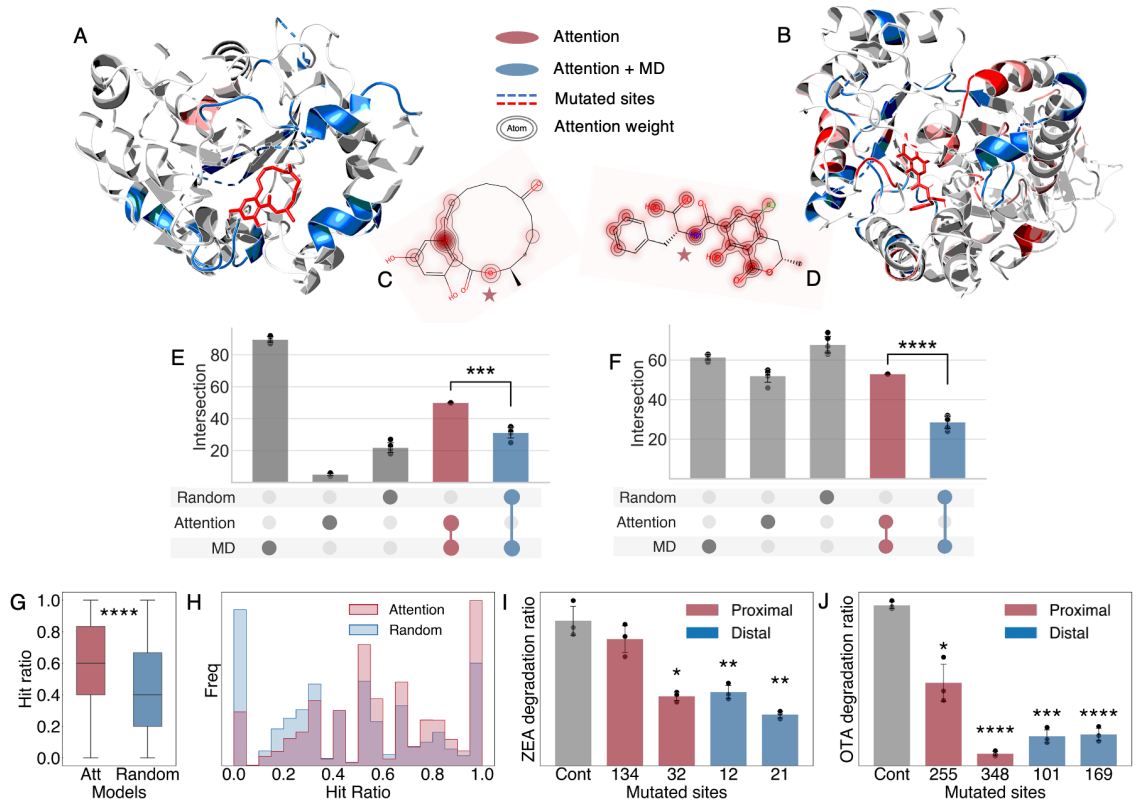

Figure 5. Evaluation of the interpretability of PU-EPP. (A, B) The attention weights on each residue in the ZH4/OH1 enzymes. The red and blue regions on enzymes present regions that were identified by PU-EPP's attention mechanism and those identified by both PU-EPP and molecular dynamics (MD), respectively. Residues marked as dotted lines were selected for subsequent mutation experiments to verify their functionality. (C, D) The attention weights on each atom in the ZEA/OTA molecules. The contour lines around the atoms represent the attention weights: the more contour lines, the higher the attention weight. Atoms near the reaction sites, such as the oxygen atom in the lactone bond of ZEA and nitrogen atom in the amido bond of OTA, both ranked 2nd in all the 45 and 46 atoms in ZEA and OTA, indicating these atoms contribute more to the final prediction. (E, F) The simplified UpSet plots of randomly selected residues and residues identified by PU-EPP and MD simulations, respectively. We evaluated the consistency of predicted residues and found ~90% of residues identified by PU-EPP on ZH4 ($n = 56$ for attention-identified residues and random residues, and $n = 161$ for MD-identified residues) and ~40% on OH1 ($n = 124$ for attention-identified residues and random residues, $n = 131$ for MD-identified residues) are consistent with MD, and both are significantly higher than randomly selected residues. Error bars represent ± s.d. (G, H) The boxplot and distribution of the hit ratio of atoms near reaction centers. It shows the median (horizontal line), 25th and 75th percentiles (lower and upper boundaries, respectively). Whiskers extend to data points that lie within 1.5 interquartile ranges of the 25th and 75th quartiles; and observations that fall outside this range are not displayed. The hit ratio of PU-EPP is significantly higher than that of random selection. (I, J) The enzyme activity before and after site mutation. The red bars indicate proximal sites, while the blue bars indicate distal sites. Error bars represent ± s.d. $P < 0.05$ (*), $P < 0.01$ (**), $P < 0.001$ (***), and $P < 0.0001$ (****), two-tailed Student's t-test.

Discussion

Identifying functional enzymes for a specific substrate and determining their interaction has been a major bottleneck in the fields of synthetic biology, metabolic engineering, and pollutant biodegradation. Here, we constructed a comprehensive enzyme promiscuity dataset containing over 600,000 known enzyme–substrate pairs and developed PU-EPP for robust enzyme promiscuity prediction based on weighted random sampling and PU learning strategies. PU-EPP achieved good performance on test sets and presented significantly better robustness and anti-interference ability than that of baseline models in cases of highly biased data. PU-EPP was combined with a rapid CFPE system to establish a rapid screening framework, which shortened the time needed for new enzyme discovery from several years to 29 d. Using the framework, we successfully identified nine carboxylic-ester hydrolases that can specifically open the macrocyclic structure of ZEA and six amidohydrolases that can efficiently degrade OTA into OTA α, by causing fundamental alterations to their conformation and eliminating most of their toxicity. The attention mechanism of PU-EPP enabled us to identify important enzyme residues, which can assist in downstream enzyme engineering and modification to optimize catalytic efficiency and stability, and ultimately promote the practical application of enzymes.

Compared with traditional strategies based on microbial screening and functional metagenomics, the proposed data-driven method can better elucidate the biotransformation potential of living organisms. For instance, of the screened enzymes, one OTA-degrading enzyme (OH9) was derived from *Exilibacterium tricleocarpae,* a marine bacterium isolated from coralline algae in the Beibu Gulf [36]. To our knowledge, this is the first reported enzyme from a marine microbe with OTA-degrading activity. Researchers intending to use the PU-EPP model are encouraged to fine-tune PU-EPP in their datasets comprising specific sub-categories of enzymes and substrates. This will further reduce the time required for model training while improving model performance.

Several aspects of the current research could be further expanded on. First, the application of enzyme-based methods to remove mycotoxin contaminants from food remains limited by high economic costs. Although degradation pathways comprising multi-step reactions can achieve complete toxin degradation, the present study used an optimal single-step reaction to degrade mycotoxins into products with low toxicity, considering economic factors. However, the screening framework can also be combined with pathway design tools, such as novePathFinder [37], to develop more comprehensive degradation pathways for pollutants or synthetic pathways for high-value chemicals. Second, similar to lead identification in drug discovery, the discovery of lead enzymes (natural enzymes with moderate activity) is the initial and rate-limiting step of the research and development cycle in enzyme engineering and, in most cases, a prerequisite for the application of directed evolution and rational design. This study focused on the screening of lead enzymes rather than the improvement of enzyme activities via structural modification. Third, although we successfully synthesized all the tested enzymes, the potential for protein expression failure remains because the correct folding of some proteins may not be observed in the CFPE system; researchers will need to consider the applicability of CFPE systems when using the screening framework. Nevertheless, we believe that the present study provides an indispensable tool for the efficient identification of new enzymes with expected catalytic activity.

Materials and Methods

Dataset construction. Enzyme–substrate interaction data were collected from the Rhea [22], KEGG [23], MetaCyc [24], Brenda [25], and RxnFinder [26,27] databases. We collected biochemical reactions with known enzymes from these databases and organized them into relational pairs composed of the enzymes and substrates. Since most of the enzymatic reactions are bidirectional, both reactants

and products of reactions were considered substrates of an enzyme. The molecular structure of substrates and the sequences of enzymes were collected from PubChem [38] and UniProt [14], respectively. Subsequently, redundancy caused by data integration was eliminated based on the SMILES of substrates and UniProt IDs of enzymes. To avoid the excessive consumption of resources and interference for model training, data pairs involving enzymes with sequence lengths >1500 amino acids and some parts of substrates, such as water, cupric ions, and carbon dioxide, were removed from the data set. Unlike the traditional random sampling methods, we adopted a weighted random sampling strategy, which assumes that enzymes with fewer reported substrates have higher catalytic specificity and enzymes with many reported substrates have higher catalytic promiscuity. Therefore, to reflect both enzyme promiscuity and specificity, more corresponding "negative" enzyme–substrate pairs should be included in the dataset (Fig 2A). The weight of the "negative" enzyme–substrate pairs was calculated as follows:

$$Weight = -\log_{10}\left(\frac{10}{X_{max}-X_{min}} \times \frac{X-X_{min}}{X_{max}-X_{min}}\right) \quad (1)$$

where X is the number of corresponding substrates of an enzyme, X_{max} is the maximum of corresponding substrates of an enzyme, and X_{min} is the minimum of corresponding substrates of an enzyme. In this way, we generated ~10 times as many "negative" enzyme–substrate pairs as known (positive) enzyme–substrate pairs. To save valuable data and computing resources, we adopted experience-based manual hyperparameter selection according to previously published studies [39]. Thus, we randomly selected 10,000 positive datapoints and 10,000 unlabeled datapoints as the test set and used the rest as the training set. The data in the test set were not "seen" by the models during the training process.

Implementation of PU learning. The implementation of PU learning primarily consisted of three steps (Fig 2B). First, equal numbers of unlabeled and positive samples were sampled and input into the model. Second, the probability ranges of positive samples in each epoch were calculated, and were used for identifying and removing potential positive samples in the unlabeled dataset. Third, the potential positive samples were removed from the dataset while others were put back. To ensure that the model had acceptable prediction ability, the PU mechanism was conducted after the model achieved an AUC >0.80. This process was repeated until the end of the training. Furthermore, the 90th percentile of the probability distribution of positive samples was chosen as the threshold for the exclusion of potential positive samples in the unlabeled dataset (Fig S3 and S4).

Construction of the deep learning pipeline. Graph neural networks (GNNs) were used to extract substrate features [40]. Substrates were considered as molecular graphs G = (v, ϵ). Each atom $v_i \in v$ was represented by a 46-dimensional vector, which was the concatenation of one-hot encodings representing the atom types, degrees of the atom, chirality, hybridization types, number of radical electrons, number of hydrogen atoms attached, explicit valence, implicit valence, and aromaticity of the corresponding atoms. ϵ was the set of covalent bonds $(v_i, v_j) \in \epsilon$, in a molecule represented as an adjacency matrix $A \in R^{N\times N}$, in which N was the number of atoms in the molecule. We used the following layer-wise propagation rules to extract substrate features:

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (2)$$

where σ is the activation function, we used ReLU ($ReLU(\cdot) = max(0,\cdot)$). $\tilde{A} = A + I_n$ is the adjacency matrix of G with added self-connections while I_n is the identity matrix. For each atom (i), the degree of the atom is $\widetilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. $W^{(l)}$ is a layer-specific trainable weight matrix. $H^l \in R^{N*d}$(d = 46) is the matrix of activations in the l^{th} layer. The output of the GNN is a metric $Z \in R^{N*d}$.

The Continuous Bag-of-Words model, a classical unsupervised learning algorithm for learning semantic knowledge from a large amount of text, was used to encode enzyme sequences:

$$L(\theta) = \frac{1}{T}\sum_{n=1}^{T} log P(w_t | w_{t-n}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+n}) \quad (3)$$

where T is the length of the enzyme sequence, P is the probability of the amino acid residues, and w_t is given by the context of the amino acid residues w_{t-n} - w_{t+n}. N is set to 1. The feature dimension was set to 100.

The multi-head self-attention mechanism [31] was employed to interpret which sub-regions on enzymes and substrates contributed more to the final prediction (Fig 2C). We used a multi-head self-attention layer in the encoder to evaluate the contribution of each point of the input enzyme sequence. Concurrently, two multi-head attention layers were used in the decoder. The first one was used to evaluate the contribution of each atom of the input substrates, and the second one was used to extract interaction information between the substrates and the enzymes. The attention mechanism was defined as follows:

$$\begin{aligned} &\alpha(q,k) = q^T k/\sqrt{d} \\ &f(q,(k_1,v_1),(k_2,v_2),\ldots,(k_n,v_n)) = \sum_{i=1}^{n} \alpha\,(q,k_i)v_i \\ &\alpha(q,k_i) = \frac{exp(\alpha(q,k_i))}{\sum_{j=1}^{n} exp(\alpha(q,k_j))} \\ &Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \end{aligned} \quad (4)$$

where $q \in R^{d_q}$ is the query, $k \in R^{d_k}$ is the key, $v \in R^{d_v}$ is the value, and Q, K, V is the minibatch representation of q, k, v. The softmax can be described as:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}} \quad (5)$$

The multi-head in multi-head attention h_ι was defined as follows:

$$h_\iota = Attention(W_\iota^q q, W_\iota^k k, W_\iota^v v) \quad (6)$$

where $W_\iota^q \in R^{p_q \times d_q}, W_\iota^k \in R^{p_k \times d_k}, W_\iota^v \in R^{p_v \times d_v}$ are learnable parameters. The learnable parameter W_o was used to process the output of multi-head attention. For h-head attention, it can be described as: $W_o[h_1, \ldots, h_h] \in R^{p_o}$. p_o is specified based on the dimension of hidden layers. In code implementation, we set $(p_q = p_k = p_v = p_o/h)$, so $(p_q h = p_k h = p_v h = p_o)$, and then h-heads attention could be computed in parallel. We used the "mask-softmax" function to obtain the valid sequence length. Any value beyond the valid length was set to an extremely small value 10^{-10}, so that it would be 0 after softmax. Then, we used linear layers to achieve the final output.

We used the Leaky Rectified Linear Unit (Leaky ReLU) as an activation function. It has been proposed as a way to resolve the dying units problem of ReLU, by preventing the unit from saturating, thus enabling a small gradient to always flow through the unit, potentially recovering extreme values of the weights [41]; it can be described as follows:

$$leaky\_relu(\mathbf{z}) = \begin{cases} \mathbf{z} & \text{if } \mathbf{z} \geq 0 \\ \alpha\mathbf{z} & \text{if } \mathbf{z} < 0 \end{cases} \quad (7)$$

where α is a hyperparameter, which was set to 0.01.

We used Glorot initialization [42] as the weight initialization method. Glorot initialization is designed to keep the scale of the gradients roughly the same in all layers. It can be described as follows:

$$a = gain * \sqrt{\frac{6}{n_{in}+n_{out}}}$$
$$\mathrm{W} \sim \mathcal{N}(-a, a) \quad (8)$$

where $gain$ is an optional scaling factor that is set to 0.2, n_{in} is the number of input units in the weight tensor, and n_{out} is the number of output units in the weight tensor.

In the training process, RAdam was used as the optimizer because it can adjust the adaptive momentum in the original Adam optimizer according to the size of the variation, outperforming manual warmups under a variety of warmup lengths and various learning rates:

$$\begin{aligned} & g_t = \nabla_\theta f_t(\theta_{t-1}) \\ & v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ & m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ & \widehat{m_t} = m_t/(1-\beta_1^t) \\ & p_t = p_\infty - 2t\beta_2^t/(1-\beta_2^t) \\ if\, p_\infty > 4: \quad & \hat{v}_t = \sqrt{v_t/(1-\beta_2^t)} \\ & r_t = \sqrt{\frac{(p_t-4)(p_t-2)p_\infty}{(p_\infty-4)(p_\infty-2)p_t}} \\ & \theta_t = \theta_{t-1} - \alpha_t r_t \widehat{m_t}/\hat{v}_t \\ else: \quad & \theta_t = \theta_{t-1} - \alpha_t \widehat{m_t} \end{aligned} \quad (9)$$

where α_t is the step size, β_1, β_2 is the decay rate to calculate the moving average and moving 2nd moment, $f_t(\theta)$ is the stochastic objective function. θ_t is the resulting parameter, and $p_\infty = 2/(1-\beta_2) - 1$ is the maximum length of the approximated simple moving average.

To make the training process more robust and stable, we introduced the LookAhead [43] mechanism, which consists of three steps:

(1) Sync slow and fast weights. For slow weights (ϕ_{t+1}):

$$\phi_{t+1} = \phi_t + \alpha(\theta_{t,k} - \phi_t) \quad (10)$$

where the synchronization period (k) and step size of slow weights (α) are hyperparameters, which were set to 5 and 0.5, respectively. $\theta_{t,k}$ is the k step (the last step) of the t round of the fast weights update.

(2) Update fast weights. We chose RAdam to update the fast weights as follows:

$$\theta_{t,i+1} = \theta_{t,i} + RAdam(L, \theta_{t,i-1}, d) \quad (11)$$

where L is the objective function, d is a sample minibatch of data, and $i \in \{1,2,\dots,k\}$ is the fast weight update step.

(3) According to the update direction of the fast parameters, the slow weights were updated as follows:

$$\phi_t = \phi_{t-1} + \alpha\left(\theta_{t,k} - \phi_{t-1}\right) \quad (12)$$

Since there may be some mislabeled samples in the unlabeled dataset, maximizing the likelihood of $\log p\,(y \mid x)$ directly can be harmful. Considering this, we used label smoothing to introduce noise for the labels, making the model more robust so that it could generalize well:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[\,(1-\epsilon)log(\epsilon) + \epsilon log(1-\epsilon)] \qquad (13)$$

where N is the batch size and ϵ is the label smoothing rate, which was set to 0.1.

Owing to the large training dataset used in this study, it was impossible for us to use common parameter search approaches, such as grid search or random search, to obtain the best hyperparameters. According to previously published studies and experimental tests on small datasets, we set the batch size to 30, the learning rate to 1E-4, the weight decay to 1E-4, the layers of the encoder block and decoder block to 12, the hidden dim to 128, and the norm shape to 128. The models were built on PyTorch 1.10 (https://pytorch.org/). We used five NVIDIA Tesla V100 GPUs for model training, which required about 14 days.

Model evaluation. PU-EPP was compared with TransformerCPI [28], DeepConv-DTI [33], and GNN-PT [34], with their default hyperparameters. We also tested the performance of the PU-EPP with five-fold cross-validation on five independent datasets, including phosphatase, esterase, glycosyltransferase, β-keto acid cleavage enzymes, and halogenase (Table S1), which were compared with the performance of prediction models developed by Goldman et al. [35]. To demonstrate whether PU-EPP can identify potential positive samples from unlabeled samples, we conducted recovering experiments of mislabeled positive samples on these five datasets. To this end, 1%, 3%, 5%, 10%, and 20% of the positive samples were mislabeled as negative, and the recognition capability of PU-EPP and default models on these mislabeled samples was tested with five-fold cross-validation.

Prediction and evaluation of mycotoxin biotransformations. ToxinDB [13] was used to predict the potential biotransformations of mycotoxins. To rationally select the optimal biotransformation for detoxification of ZEA and OTA, we evaluated the feasibility, safety, and economics of predicted biotransformations at the molecular (e.g., the availability of co-reactants and toxicity and synthesis feasibility of mycotoxin metabolites) and reaction levels (e.g., reaction feasibility and biotransformation types) (Table S2). Because ToxinDB [13] can only predict the main metabolic products of the input mycotoxins, the co-reactants of the top-ranked reference reaction provided by ToxinDB were considered as the co-reactants of the predicted biotransformation and subsequently used for evaluation. Meanwhile, the enzyme classification (EC) number of the top-ranked reference reaction was considered as the EC number of the predicted biotransformation.

Screening and expression of candidate enzymes. UniProt [14] was used as the source of candidate enzymes. Because the EC numbers of most enzymes in UniProt [14] are unknown, we annotated the EC number of the enzymes with Bio2Rxn [44], a sequence-catalytic function model based on a consensus strategy. Subsequently, enzyme sequences belonging to the 3.1.1.- and 3.5.1.- categories were extracted, and the enzyme sequences and the molecular structure of ZEA and OTA were input into the PU-EPP model in pairs to obtain the predicted probability of whether a certain enzyme could catalyze a selected substrate. Then, according to the predicted probability, 10 enzymes for each mycotoxin with potential catalytic activity were selected for expression and activity verification (Table S5). The gene sequences for coding candidate enzymes were synthesized and then cloned into the pD2P vector from Healthcode (http://healthcode.com/). The constructed plasmids with candidate genes were expressed using the cell-free expression system provided by Healthcode according to their instructions. The control was a plasmid without candidate genes. After incubation at 30 °C for 4 h for targeted enzyme production, the supernatant of reaction mixtures containing enzymes was collected for subsequent mycotoxin degradation assays.

Evaluation of the catalytic activity of candidate enzymes. The mycotoxin standards of ZEA or OTA were added into 500 μL reaction supernatants at a final concentration of 1000 ng/mL. One major method to eliminate mycotoxins using enzymes is to fix enzymes with the feed. These enzymes start to perform detoxification when they reach the stomach of animals along with the feed. To simulate the activity of enzymes under practical conditions (e.g., animal intestines and stomach), we set the temperature of the experimental environment to 37 °C. After incubation at 37 °C for 3 h, 500 μL methanol was added to stop the reaction. The mixture was shaken vigorously and then centrifuged at 13,000 *g* for 10 min. Finally, the supernatant was filtered using a 0.22-μm filter before being subjected to liquid chromatography–mass spectrometry (LC-MS) analysis. Mycotoxin determination and metabolite analyses were achieved following the previously developed LC-MS method [45]. For the degradation assay under food matrices, mycotoxin standards were added to the wheat and maize flour at a final concentration of 1000 ng/g. Then, 500 μL enzyme supernatant was mixed with 500 mg mycotoxin-contaminated flour. After incubation at 37 °C for 3 h, the reaction was stopped by adding 500 μL methanol. Finally, the mycotoxin in the samples was extracted for analysis as described above. The experiments were independently performed three times (n=3 biologically independent samples), and at least two technical replicates were used in each experimental group.

Model interpretability evaluation and MD simulations. To interpret which residues on enzymes and which atoms on substrates are most important for the catalysis process, the multi-head self-attention mechanism [31] was employed by assigning attention weights to the residues and atoms. For instance, a higher attention weight for a residue means that the residue is more important for enzyme activity toward the specific substrate, a higher attention weight for an atom means that it is important for binding with enzymes or means it may be present at the reaction sites (i.e., the atoms and bonds changed in the substrate during reactions) [21]. To evaluate the model's recognition ability for the reaction sites on substrates, we compared attention weights with known reaction sites collected from EnzyMine [46]. Atoms less than one bond distance from bonds that changed in the reaction were considered reaction sites. We considered it to be identified correctly if an atom in a reaction site was ranked in the top 2N (N = the number of atoms in reaction sites) of all atoms because atoms with other functions (e.g., providing binding affinity with enzymes) may also be top ranked. We also randomly generated equal numbers of atoms on each substrate. Then, we compared the recognition ratio of atoms in reaction sites by PU-EPP and the recognition ratio of the randomly generated result. To evaluate the correctness of important residues inferred using the attention mechanism, we analyzed the important residues on ZH4 and OH1 enzymes using MD simulations (Fig S7). The first 25 amino acids in OH1 were deleted because they were likely to serve as a signal peptide. Then, the consistency of the results achieved by these two approaches was evaluated. To conduct MD simulations, an initial guess of the small molecule binding site in the enzyme was detected by BiteNet (https://sites.skoltech.ru/imolecule/tools/bitenet/) with the protein structure predicted by AlphaFold [47] as input. Additionally, four extra putative sites (e.g., pockets and grooves) were manually identified. The small molecule was then docked to each putative site via AutoDock Vina using the default scoring function. For docking, a 25 × 25 × 25 Å cubic box was set up to encompass the binding site, and the exhaustiveness value was set as 32. For each putative site, nine binding poses were obtained. Then, we selected the 10 top-ranked binding poses to conduct all-atom MD simulations.

MD simulations were performed using Amber (https://ambermd.org/). The force field parameters for small molecules were adapted from the Amber GAFF force field, with point charges of atoms calculated via AM1-BCC. The protein was described by the Amber FF19SB force field and solvated in an OPC water box with a 12-Å buffer. KCl ions were added into the system to

neutralize the charge. The system was first energy-minimized using the steepest descent algorithm for 200 steps followed by the conjugate gradient algorithm for 9800 steps. The system was then heated from 0 to 100 K in the NVT ensemble over 800 ps and from 100 to 300 K in the NPT ensemble over 2 ns. During the heating stage, the backbone of the protein was restrained by a harmonic constraint with a force constant of 5 kcal/mol/Å^2. A 1-ns equilibrium simulation with force constant reduced to 1 kcal/mol/Å^2 was then performed to relax the system. The equilibrated system was used to initiate a 200-ns production run. During the simulations, the SHAKE algorithm [48] was used to restrain the lengths of bonds that involve hydrogen. The hydrogen mass repartition operation, which redistributes the masses of heavy atoms to their bonded hydrogens, was performed to enable a 4-fs time step for MD simulations. The Langevin thermostat was used during the simulations, and the equilibration and production runs were performed in the NPT ensemble at 300 K. The non-bonded interaction cutoff was 10 Å. For each of the selected binding poses, its 200-ns MD trajectory with the protein was subjected to MM-GBSA [49] to estimate the binding free energy. The ionic strength was set at 0.15 M. The binding poses were re-ranked according to their MM-GBSA binding free energies. Finally, we identified and selected the three top-ranked poses and extended their simulations to 1 μs.

Based on the extended MD trajectories, the RMSF of the protein's heavy atoms and the dynamic cross-correlation matrix (DCCM) of the protein's C_α atoms were calculated via the CPPTRAJ program [50]. The alignment of trajectories to the first frame was conducted prior to the RMSF and DCCM calculations. Residues that made a < -1 kcal/mol contribution to MM-GBSA-binding free energy were categorized as energetically important residues. Residues that showed clear RMSF changes (ΔRMSF > ~0.5 Å) and significant correlation changes (as evidenced by DCCM analyses) upon small molecule binding were identified as potential allosterically important residues. Collectively, these energetically and allosterically important residues were identified as key residues for enzyme catalysis.

We used RDKit (https://www.rdkit.org/) to visualize attention weights on the molecular structure of substrates. We used ProDy (http://prody.csb.pitt.edu/) and PyMol (https://pymol.org/) to visualize attention weights and MD-identified important residues on the structure of enzymes. The 90th percentile of the probability distribution of attention weights was considered the threshold to identify important residues. According to the Continuous Bag-of-Words mechanism, the number of attention weights extracted is two less than the number of amino acids; therefore, we set the attention weight of an amino acid as the highest attention weight in three neighboring amino acids and then selected the important amino acids according to the threshold.

Gene site mutagenesis. The residue sites for mutation were selected according to attention weights, including four proximal sites (distance to substrate <10 Å) and four distal sites (distance to substrate >10 Å). Gene site mutagenesis was achieved using a ClonExpress Ultra One Step Cloning Kit (Vazyme Biotech, Nanjing) as per the manufacturer's protocol. The pD2P plasmids with genes ZH4 and OH1 were used as PCR templates to generate point mutations in selected residues of ZH4 and OH1. The sequence of each variant was confirmed through sequencing. Finally, the pD2P plasmids with point mutation on genes ZH4 (index of amino acid residues: 12, 21, 32, and 134) and OH1 (index of amino acid residues: 101, 169, 255, and 348) were expressed and assayed for their mycotoxin degradation activity, as described in the sections on the evaluation of the catalytic activity of candidate enzymes. The experiments were independently performed three times (n=3 biologically independent samples), and at least two technical replicates were used in each experimental group.

Acknowledgments

We would like to thank S. Pfister and Y. Yu for providing feedback to improve this manuscript. This project received funding from the National Key Research and Development Program of China [Grant numbers: 2021YFC2103001 and 2019YFA0904300], the International Partnership Program of the Chinese Academy of Sciences (CAS) [Grant number: 153D31KYSB20170121], and the CAS Science and Technology Service Network Initiative Program [Grant number: QYZDB-SSW-SMC012].

Data Availability

All data are publicly available. However, in some cases, user licenses are required to access the underlying data. The dataset of enzyme–substrate pairs for model training was collected from Rhea (https://www.rhea-db.org/), KEGG (https://www.genome.jp/kegg/), MetaCyc (https://metacyc.org/), RxnFinder (http://www.rxnfinder.org/rxnfinder/), and Brenda (https://www.brenda-enzymes.org/). The molecular structures of substrates were collected from PubChem (https://pubchem.ncbi.nlm.nih.gov/). The amino acid sequences of enzymes were collected from UniProt (https://www.uniprot.org/). Reaction site data was collected from EnzyMine (http://www.rxnfinder.org/enzymine/). Datasets of five types of enzymes for model evaluation were collected from the GitHub repository (https://github.com/samgoldman97/enzyme-datasets). An example dataset of enzyme–substrate pairs for training and testing is provided in the GitHub repository: https://github.com/xinghd142857/PU-EPP.

To facilitate further usage, detailed instructions and all codes for model training and testing are provided in a zenodo repository: https://doi.org/10.5281/zenodo.7813738 and a GitHub repository: https://github.com/xinghd142857/PU-EPP/. User-friendly examples of enzyme promiscuity prediction and fine-tuning PU-EPP on new datasets are also included in the repository. Any additional information required to reanalyze the data reported in this paper is available from the corresponding author upon reasonable request.

References

1 L. E. Johns, D. P. Bebber, S. J. Gurr, N. A. Brown, Emerging health threat and cost of Fusarium mycotoxins in European wheat. *Nat. Food* 3, 1014–1019 (2022). https://doi.org:10.1038/s43016-022-00655-z

2 Y. Tao, S. Xie, F. Xu, A. Liu, Y. Wang, D. Chen, Y. Pan, L. Huang, D. Peng, X. Wang, Z. Yuan, Ochratoxin A: toxicity, oxidative stress and metabolism. *Food Chem. Toxicol.* 112, 320–331 (2018). https://doi.org/10.1016/j.fct.2018.01.002

3 C. Gruber-Dorninger, T. Jenkins, G. Schatzmayr, Global mycotoxin occurrence in feed: A ten-year survey. *Toxins (Basel)* 11 (2019). https://doi.org:10.3390/toxins11070375

4 G. Schatzmayr, E. Streit, Global occurrence of mycotoxins in the food and feed chain: facts and figures. *World Mycotoxin J.* 6, 213–222 (2013). https://doi.org:10.3920/wmj2013.1572

5 M. Eskola, G. Kos, C.T. Elliott, J. Hajšlová, S. Mayar, R. Krska, Worldwide contamination of food-crops with mycotoxins: validity of the widely cited 'FAO estimate' of 25. *Crit. Rev. Food Sci. Nutr.* 60, 2773–2789 (2020). https://doi.org:10.1080/10408398.2019.1658570

6 D. Milicevic, R. Petronijević, Z. Petrović, J. Đinović-Stojanović, J. Jovanović, T. Baltić, S. Janković, Impact of climate change on aflatoxin M1 contamination of raw milk with special focus on climate conditions in Serbia. *J. Sci. Food Agric.* 99, 5202–5210 (2019). https://doi.org:10.1002/jsfa.9768

7 S. Mishra, S. Srivastava, J. Dewangan, A. Divakar, S. Kumar Rath, Global occurrence of deoxynivalenol in food commodities and exposure risk assessment in humans in the last decade: a survey. *Crit. Rev. Food Sci. Nutr.* 60, 1346–1374 (2020). https://doi.org:10.1080/10408398.2019.1571479

8 R. Colovic, N. Puvača, F. Cheli, G. Avantaggiato, D. Greco, O. Đuragić, J. Kos, L. Pinotti, Decontamination of mycotoxin-contaminated feedstuffs and compound feed. *Toxins (Basel)* 11 (2019). https://doi.org:10.3390/toxins11110617
9 P. Karlovsky, M. Suman, F. Berthiller, J. De Meester, G. Eisenbrand, I. Perrin, I.P. Oswald, G. Speijers, A. Chiodini, T. Recker, P. Dussort, Impact of food processing and detoxification treatments on mycotoxin contamination. *Mycotoxin Res.* 32, 179–205 (2016). https://doi.org:10.1007/s12550-016-0257-7
10 H. Xu, L. Wang, J. Sun, L. Wang, H. Guo, Y. Ye, X. Sun, Microbial detoxification of mycotoxins in food and feed. *Crit. Rev. Food Sci. Nutr.* 62, 4951–4969 (2022). https://doi.org:10.1080/10408398.2021.1879730
11 Y. Tian, D. Zhang, P. Cai, H. Lin, H. Ying, Q. N. Hu, A. Wu, Elimination of Fusarium mycotoxin deoxynivalenol (DON) via microbial and enzymatic strategies: current status and future perspectives. *Trends Food Sci. Technol.* 124, 96–107 (2022). https://doi.org:10.1016/j.tifs.2022.04.002
12 S. W. Gratz, Do plant-bound masked mycotoxins contribute to toxicity? *Toxins (Basel)* 9 (2017). https://doi.org:10.3390/toxins9030085
13 D. Zhang, Y. Tian, Y. Tian, H. Xing, S. Liu, H. Zhang, S. Ding, P. Cai, D. Sun, T. Zhang, Y. Hong, A data-driven integrative platform for computational prediction of toxin biotransformation with a case study. *J. Hazard Mater.* 408, 124810 (2021). https://doi.org:10.1016/j.jhazmat.2020.124810
14 UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489 (2021). https://doi.org:10.1093/nar/gkaa1100
15 UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506-D515 (2018). https://doi.org:10.1093/nar/gky1049
16 W. M. Atkins, Mechanisms of promiscuity among drug metabolizing enzymes and drug transporters. *FEBS J.* 287, 1306–1322 (2020). https://doi.org:10.1111/febs.15116
17 G. M. Visani, M. C. Hughes, S. Hassoun, Enzyme promiscuity prediction using hierarchy-informed multi-label classification. *Bioinformatics* 37, 2017–2024 (2021). https://doi.org:10.1093/bioinformatics/btab054
18 H. Nam, N. E. Lewis, J. A. Lerman, D. H. Lee, R. L. Chang, D. Kim, B. O. Palsson, Network context and selection in the evolution to enzyme specificity. *Science* 337, 1101–1104 (2012). https://doi.org:10.1126/science.1216861
19 W. Finnigan, L. J. Hepworth, S. L. Flitsch, N. J. Turner, RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* 4, 98–104 (2021). https://doi.org:10.1038/s41929-020-00556-z
20 Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. Science 379, 1358–1363. https://doi.org:10.1126/science.adf2465
21 F. Li, L. Yuan, H. Lu, G. Li, Y. Chen, M. K. M. Engqvist, E. J. Kerkhoven, J. Nielsen, Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal.* 5, 662–672 (2022). https://doi.org:10.1038/s41929-022-00798-z
22 T. Lombardot, A. Morgat, K. B. Axelsen, L. Aimo, N. Hyka-Nouspikel, A. Niknejad, A. Ignatchenko, I. Xenarios, E. Coudert, N. Redaschi, A. Bridge, Updates in Rhea: SPARQLing biochemical reaction data. *Nucleic Acids Res.* 47, D596–D600 (2019). https://doi.org:10.1093/nar/gky876
23 M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361 (2017). https://doi.org:10.1093/nar/gkw1092
24 R. Caspi, R. Billington, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P. E. Midford, Q. Ong, W. K. Ong, S. Paley, P. Subhraveti, P. D. Karp, The MetaCyc database of

metabolic pathways and enzymes. *Nucleic Acids Res.* 46, D633–D639 (2018). https://doi.org:10.1093/nar/gkx935
25 S. Placzek, I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, D. Schomburg, BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res*. 45, D380-D388 (2017). https://doi.org:10.1093/nar/gkw952
26 Q. N. Hu, Z. Deng, H. Hu, D. S. Cao, Y. Z. Liang, RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. *Bioinformatics* 27, 2465–2467 (2011). https://doi.org:10.1093/bioinformatics/btr413
27 Y. Tian, L. Wu, L. Yuan, S. Ding, F. Chen, T. Zhang, A. Ren, D. Zhang, W. Tu, J. Chen, Q. N. Hu, BCSExplorer: a customized biosynthetic chemical space explorer with multifunctional objective function analysis. *Bioinformatics* 36, 1642–1643 (2020). https://doi.org:10.1093/bioinformatics/btz755
28 L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, M. Zheng, TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* 36, 4406–4414 (2020). https://doi.org:10.1093/bioinformatics/btaa524
29 S. M. H. Mahmud, W. Chen, H. Meng, H. Jahan, Y. Liu, S. M. M. Hasan, Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Anal. Biochem.* 589, 113507 (2020). https://doi.org:10.1016/j.ab.2019.113507
30 S. Jain, M. White, P. Radivojac, Recovering true classifier performance in positive-unlabeled learning. *Proc. AAAI Conf. Artif. Intell.* 31 (2017). https://doi.org:10.1609/aaai.v31i1.10937
31 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017).
32 T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013)
33 I. Lee, J. Keum, H. Nam, DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* 15, e1007129 (2019). https://doi.org:10.1371/journal.pcbi.1007129
34 J. Wang, X. Li, H. Zhang, GNN-PT: enhanced prediction of compound-protein interactions by integrating protein transformer. Preprint at https://arxiv.org/abs/2009.00805 (2020)
35 S. Goldman, R. Das, K. K. Yang, C. W. Coley, Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS Comput. Biol.* 18, e1009853 (2022). https://doi.org:10.1371/journal.pcbi.1009853
36 G. Wang, X. Zheng, S. Xu, G. Dang, H. Su, Z. Liao, B. Chen, W. Huang, J. Liang, K. Yu, Exilibacterium tricleocarpae gen. nov., sp. nov., a marine bacterium from coralline algae Tricleocarpa sp. *Int. J. Syst. Evol. Microbiol.* 70, 3427–3432 (2020). https://doi.org:10.1099/ijsem.0.004189
37 S. Ding, Y. Tian, P. Cai, D. Zhang, X. Cheng, D. Sun, L. Yuan, J. Chen, W. Tu, D. Q. Wei, Q. N. Hu, novoPathFinder: a webserver of designing novel-pathway with integrating GEM-model. *Nucleic Acids Res.* 48, W477–W487 (2020). https://doi.org:10.1093/nar/gkaa230
38 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. Q. He, B. A. Shoemaker, J. Y. Wang, B. Yu, J. Zhang, S. H. Bryant, PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213 (2016). https://doi.org:10.1093/nar/gkv951
39 S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning. Preprint at https://arxiv.org/abs/1811.12808 (2018)

40 J. Chen, S. Zheng, H. Zhao, Y. Yang, Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *J. Cheminform.* 13, 7 (2021). https://doi.org:10.1186/s13321-021-00488-1
41 T. Kipf, M. Welling, Semi-Supervised classification with graph convolutional networks. Preprint at https://arxiv.org/abs/1609.02907 (2016)
42 X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* 9, 249–256 (2010).
43 M. R. Zhang, J. Lucas, G. Hinton, J. Ba, Lookahead optimizer: k steps forward, 1 step back. *Adv. Neural Inf. Process Syst. 32 (NIPS 2019)* 32 (2019).
44 T. Zhang, Y. Tian, L. Yuan, F. Chen, A. Ren, Q. N. Hu, Bio2Rxn: sequence-based enzymatic reaction predictions by a consensus strategy. *Bioinformatics* 36, 3600–3601 (2020). https://doi.org:10.1093/bioinformatics/btaa135
45 M. S. Azam, D. Yu, N. Liu, A. Wu, Degrading ochratoxin A and zearalenone mycotoxins using a multifunctional recombinant enzyme. *Toxins (Basel)* 11 (2019). https://doi.org:10.3390/toxins11050301
46 D. Sun, X. Cheng, Y. Tian, S. Ding, D. Zhang, P. Cai, Q. N. Hu, EnzyMine: a comprehensive database for enzyme function annotation with enzymatic reaction chemical feature. *Database* (2020). https://doi.org/10.1093/database/baaa065
47 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org:10.1038/s41586-021-03819-2
48 J. Ryckaert, G. Ciccotti, H. J. C. Berendsen, Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341 (1997). https://doi.org:10.1016/0021-9991(77)90098-5
49 A. Onufriev, D. Bashford, D. A. Case, Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55, 383–394 (2004). https://doi.org:10.1002/prot.20033
50 D. R. Roe, T. E. Cheatham, 3rd. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9, 3084–3095 (2013). https://doi.org:10.1021/ct400341p

Supplementary Information

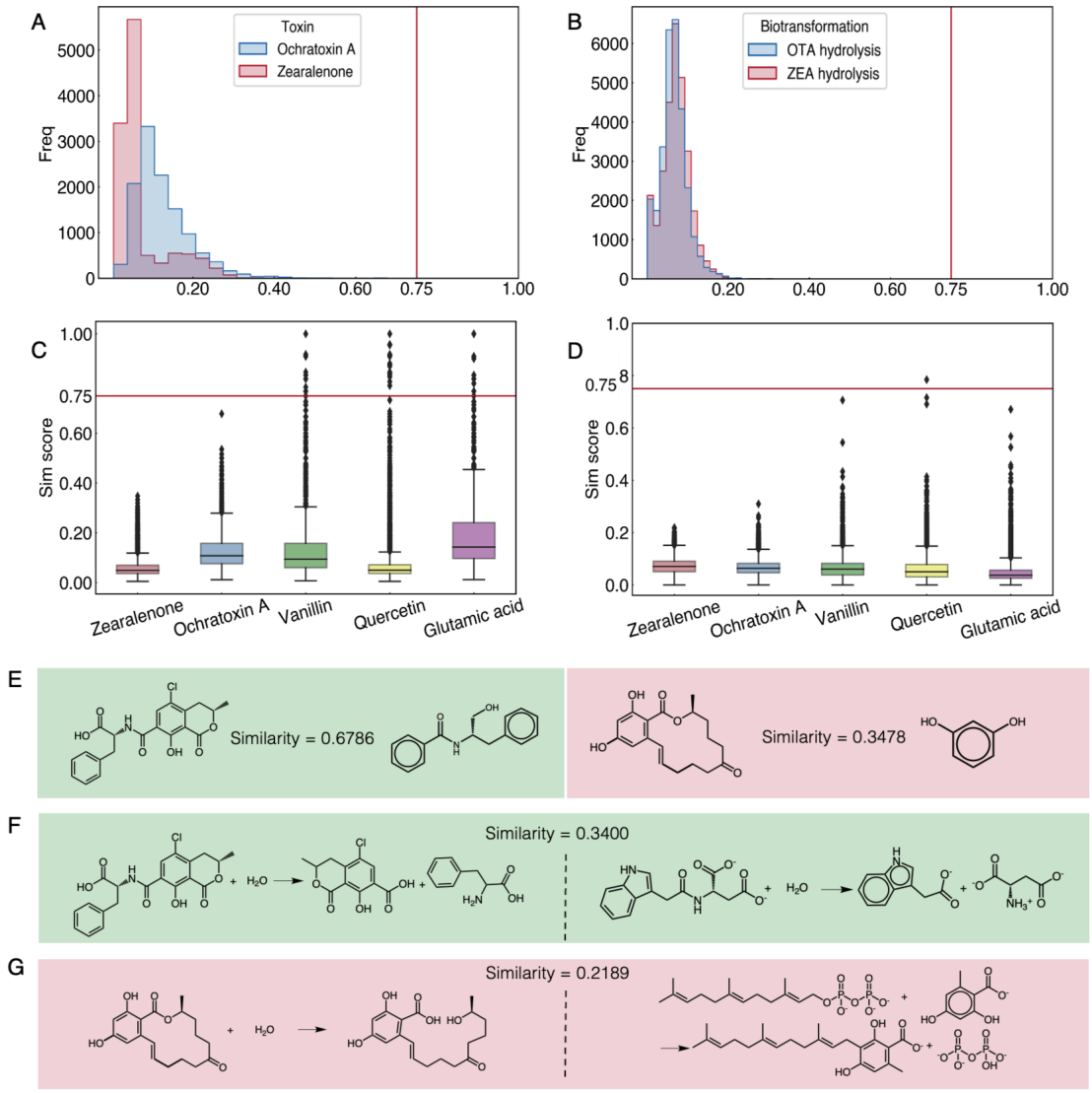

Fig. S1. Prediction of mycotoxin-degrading enzymes based on molecular and reaction similarity. (A) Frequency distribution of similarity scores between substrates in Rhea [1] (N = 11,780) and zearalenone (ZEA) and ochratoxin A (OTA). (B) Frequency distribution of similarity scores between biochemical reactions in Rhea (N = 29,101) and ZEA and OTA hydrolysis reactions. Molecules or reactions are treated as similar if they exceed similarity score threshold of 0.75. These results show that similarity-based methods cannot provide reliable predictions for ZEA and OTA. (C) Molecular similarity scores (N = 11,780) of ZEA, OTA, and three well-studied metabolites (vanillin, quercetin, and glutamic acid) with substrates in Rhea. (D) Reaction-similarity scores (N = 11,780) between biochemical reactions in Rhea and the hydrolysis reactions of ZEA, OTA, vanillin, quercetin, and glutamic acid. Compared to well-studied molecules, molecules with higher structural specificity, such as ZEA and OTA, have fewer similar molecules and reactions. This indicates that traditional similarity-based methods and conventional fingerprints are not eligible to assist in the identification of enzymes for complex substrates (e.g., ZEA and OTA). (E) The most similar molecules with OTA

and ZEA in the Rhea database. (F, G) The most similar reactions with OTA and ZEA hydrolysis reactions in the Rhea database. Molecular similarity was calculated using the maximum common substructure algorithm [2]. Reaction similarity was calculated based on the reaction difference fingerprint [3] and the Tanimoto algorithm [4]. Both hydrolysis reactions of ZEA and OTA are not included in the International Union of Biochemistry and Molecular Biology Enzyme Commission (EC) system [5], indicating EC-based prediction methods (e.g., CLEAN [6]) do not apply to these cases.

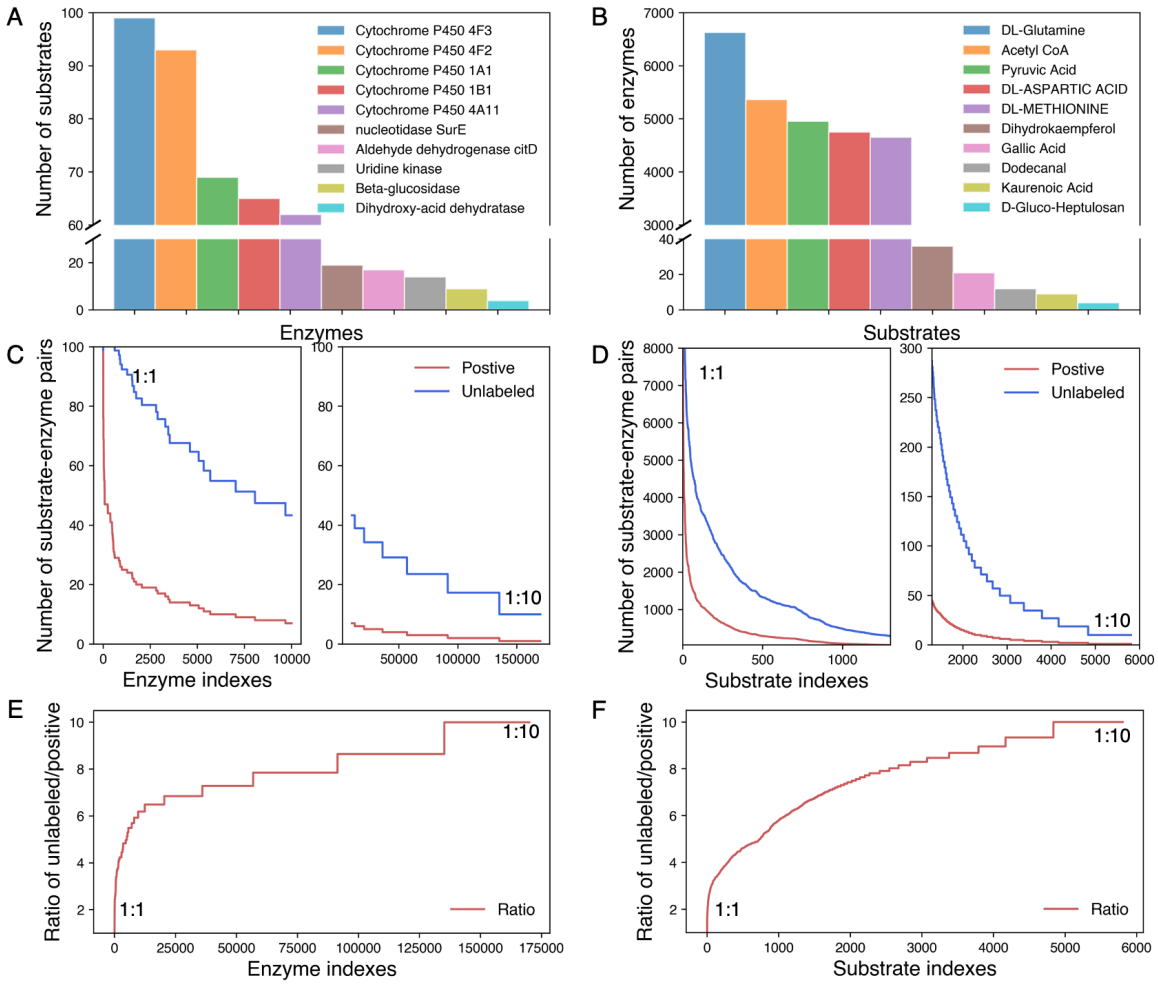

Fig. S2. Weighted random sampling strategy reflecting both enzyme specificity and promiscuity. (A) Distribution of the number of catalyzed substrates of enzymes in the positive data. (A) P450 metabolic enzymes catalyze several substrates, indicating that they have high promiscuity. Synthases of secondary metabolites can only catalyze a few substrates. (B) Distribution of corresponding enzymes of substrates in the positive data. Common primary metabolites such as glutamine, acetyl CoA, and pyruvic acid can be catalyzed by numerous enzymes, while secondary metabolites such as dihydrokaempferol, gallic acid, and kaurenoic acid can only be catalyzed by several specific enzymes. (C, D) Distributions of positive (N = 606,555) and negative samples (N = 6,488,914) were generated based on the weighted random sampling strategy. For enzymes/substrates with high specificity, more corresponding unlabeled (considered as negative) enzyme–substrate pairs were included in the dataset. For enzymes/substrates with high promiscuity, fewer corresponding unlabeled enzyme–substrate pairs were included in the dataset. (E, F) Ratios of unlabeled and positive enzyme–substrate pairs. The weighted random sampling strategy assumes that enzymes with fewer reported substrates have a higher catalytic specificity and enzymes with many reported substrates have a higher catalytic promiscuity.

Therefore, to reflect both enzyme promiscuity and specificity, 1–10 times as many corresponding unlabeled (considered as negative) enzyme–substrate pairs were included in the dataset.

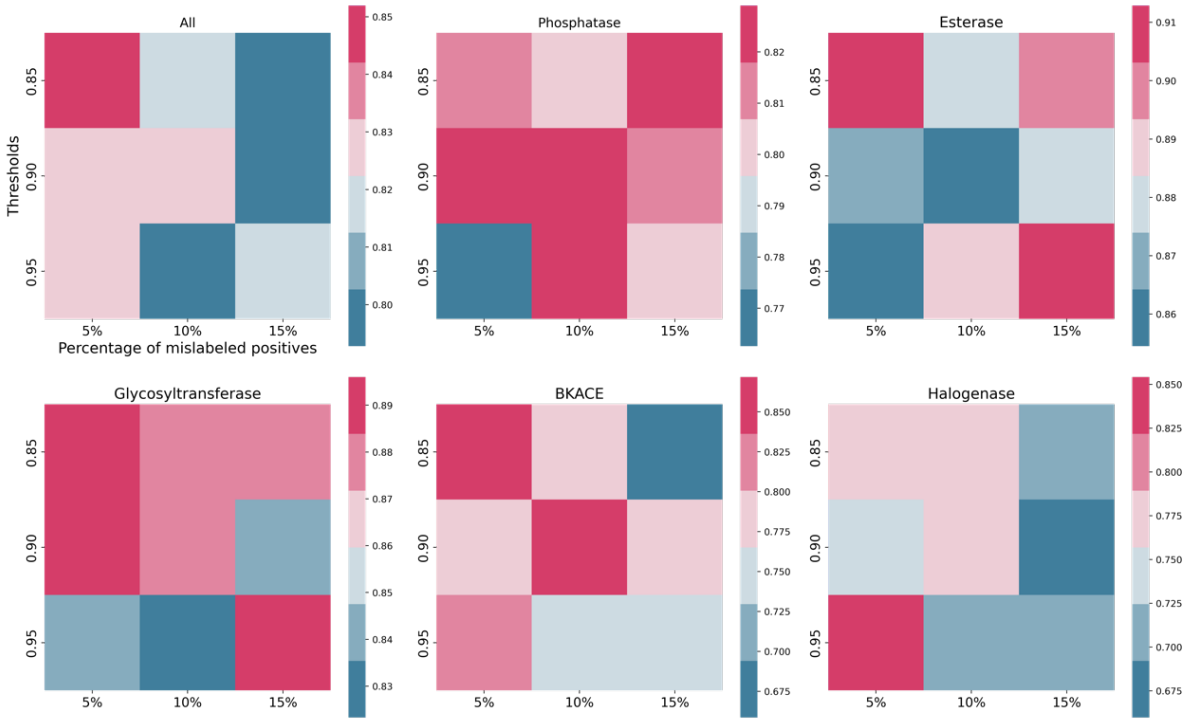

Fig. S3. Recovery experiments on five datasets consisting of real positive and negative samples. This was done to evaluate the model's hit ratios of mislabeled positives. The thresholds (0.85, 0.90, 0.95) stand for the probability ranges of positive samples for identifying and removing potential positive samples in the unlabeled dataset. The model achieved relatively stable hit ratios when the threshold was set to 0.90.

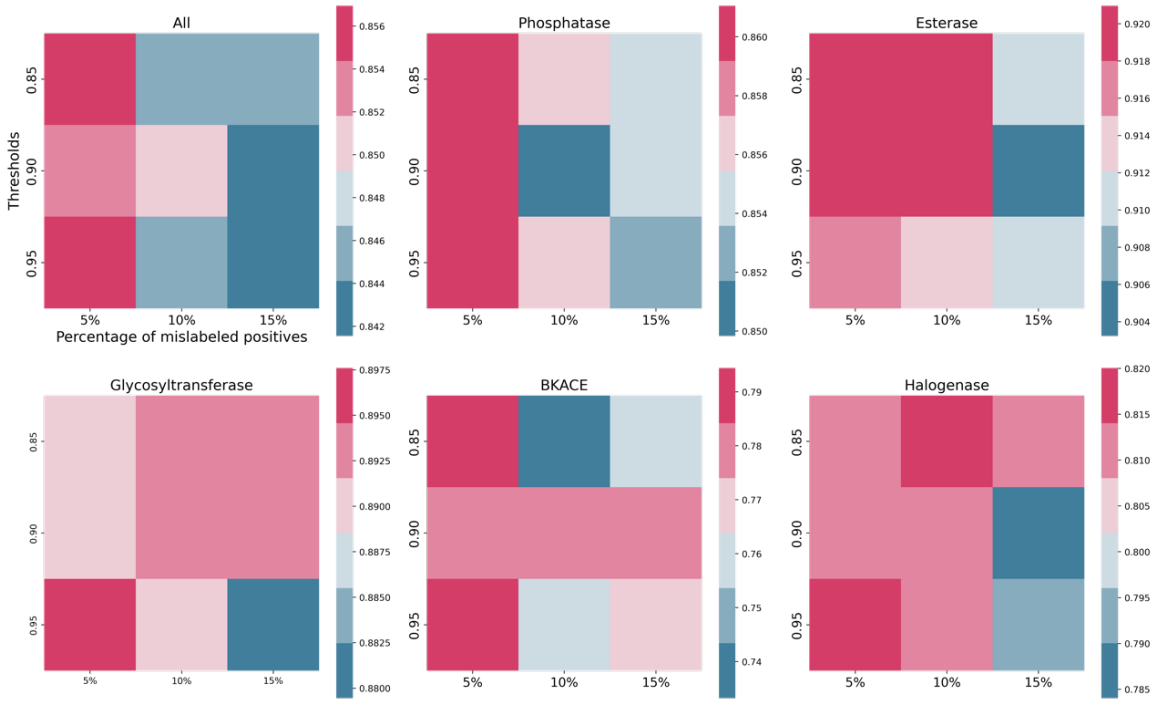

Fig. S4. Models' area under the curve (AUC) scores under different thresholds (0.85, 0.90, 0.95). The model achieved a relatively stable AUC when the threshold was set to 0.90.

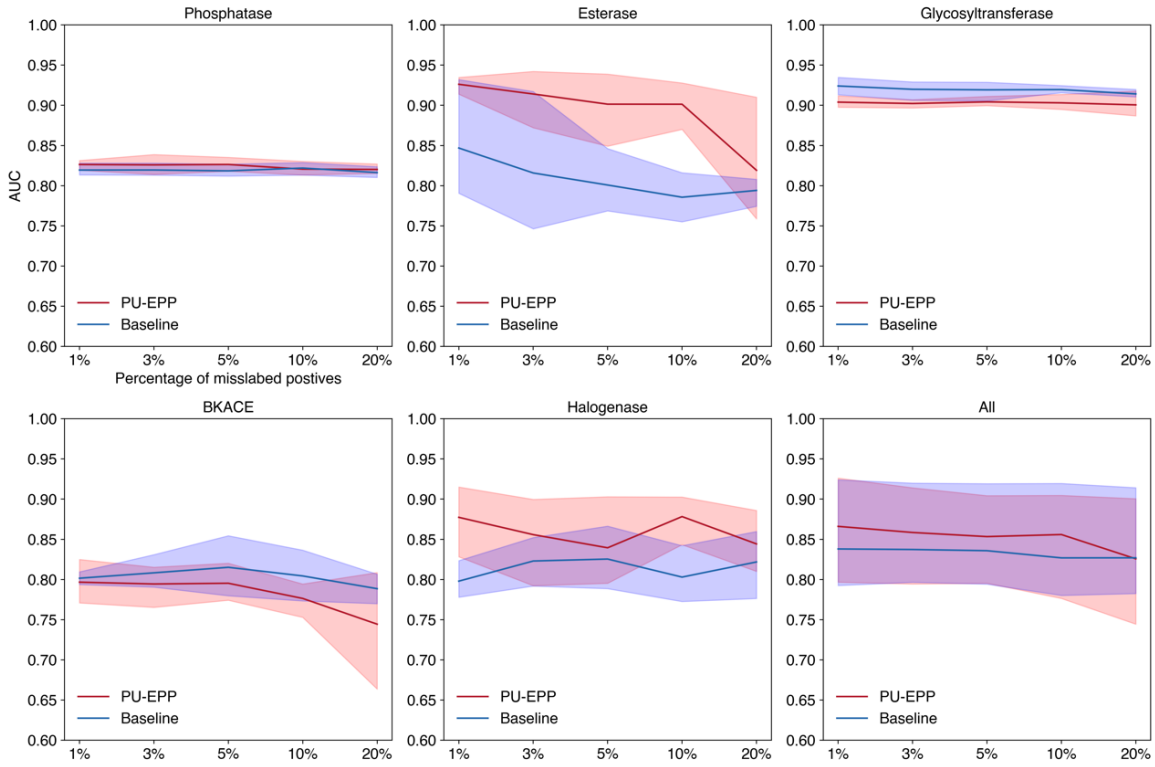

Fig. S5. Models' performance on five datasets related to phosphatase, esterase, glycosyltransferase (GT), β-keto acid cleavage enzyme (BKACE), and halogenase. The performance of the positive-unlabeled learning-based enzyme promiscuity prediction (PU-EPP) model is better than that of the baseline model.

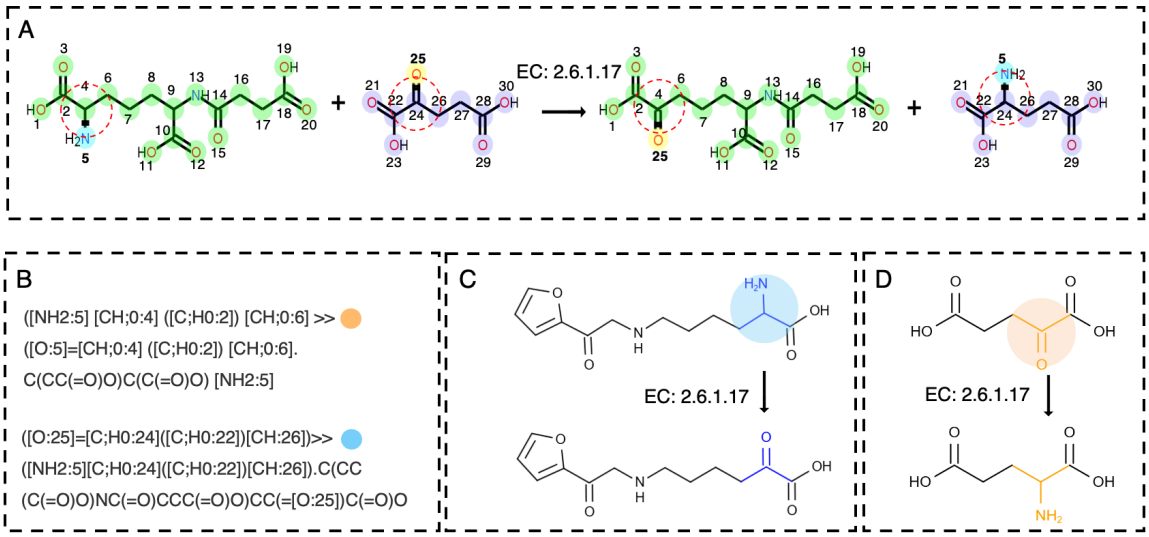

Fig. S6. Prediction of biotransformation reactions based on reaction rules. From the metabolic reaction (A) catalyzed by enzymes (EC number: 2.6.1.17), two reaction rules were extracted and encoded into the simplified molecular input line entry system (SMILES) arbitrary target specification format (B). These reaction rules were used to predict a molecule's potential biotransformations and metabolic products if the molecule shared a common substructure (blue and orange parts in C and D) with the reaction rule.

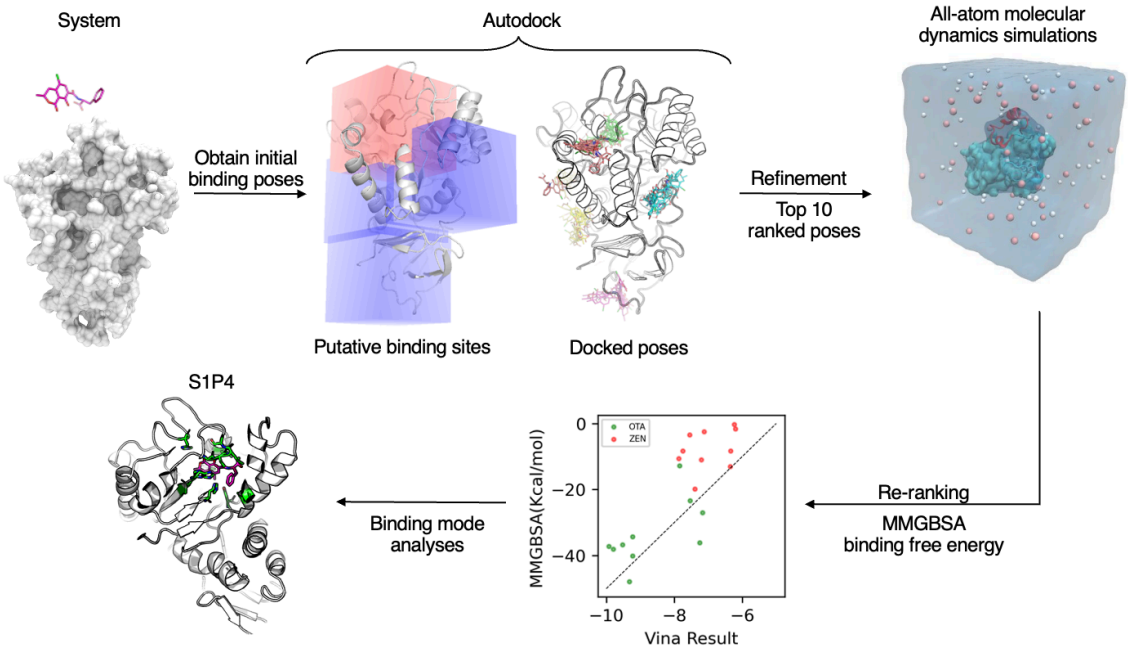

Fig. S7. Schematic of the computational methodology used for exploring the binding mode between substrates and enzymes. First, five putative binding sites were used for docking calculations. Second, the top-ranked docking poses were refined by all-atom molecular dynamics simulations. Finally, the binding poses were re-ranked using molecular mechanics-generalized Born and surface area (MM-GBSA)-calculated binding free energies. The most probable binding pose was selected for binding mode analyses.

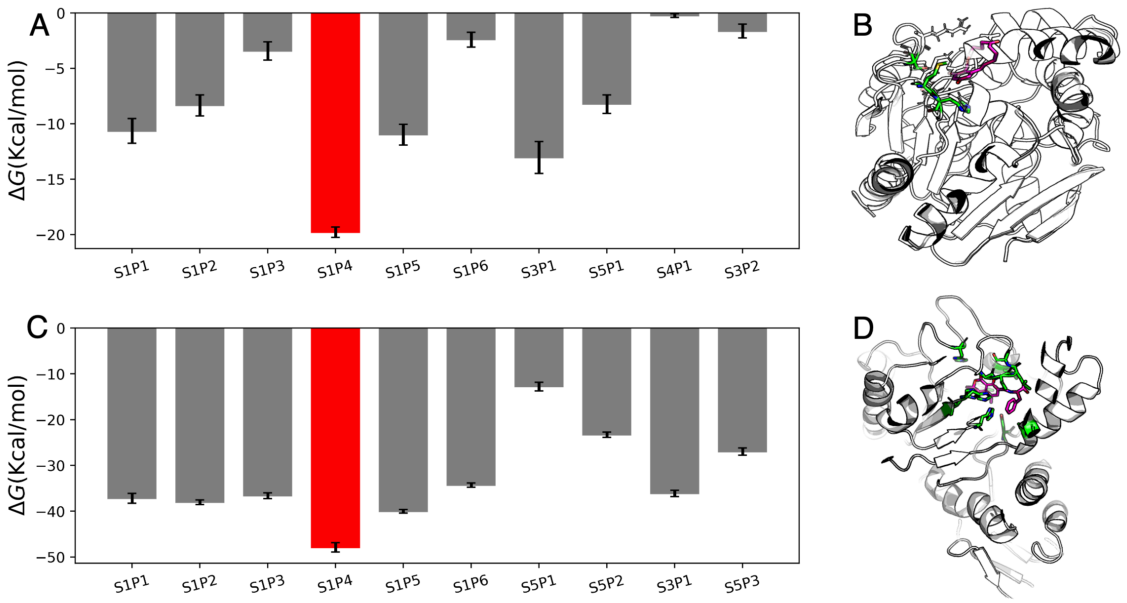

Fig. S8. Predicted enzyme–substrate binding sites and binding conformations. (A) A list of candidate binding sites and binding conformations of the ZH4 enzyme and zearalenone. (B) The most likely binding sites (S1P4) and conformations of ZH4. (C) A list of candidate binding sites and binding conformations of OH1 and OTA. (D) The most likely binding sites (S1P4) and conformations of the OH1 enzyme.

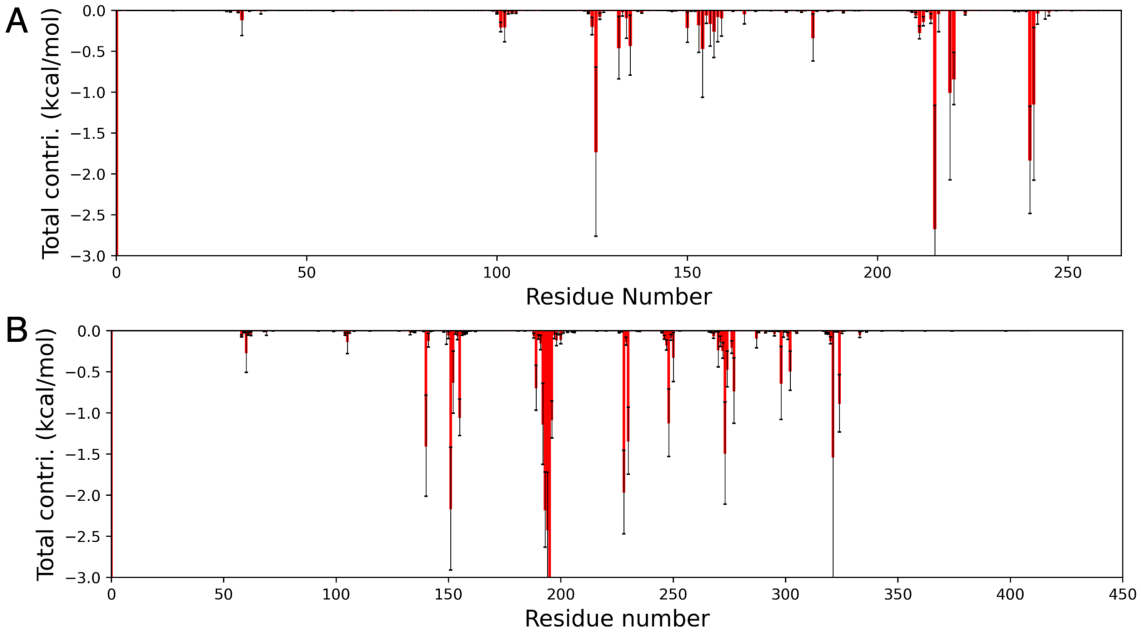

Fig. S9. Decomposition of the binding free energy using MM-GBSA to identify the residues contributing most to substrate binding. (A) Decomposition of the substrate-binding free energy of the ZH4 enzyme. (B) Decomposition of the substrate-binding free energy of the OH1 enzyme.

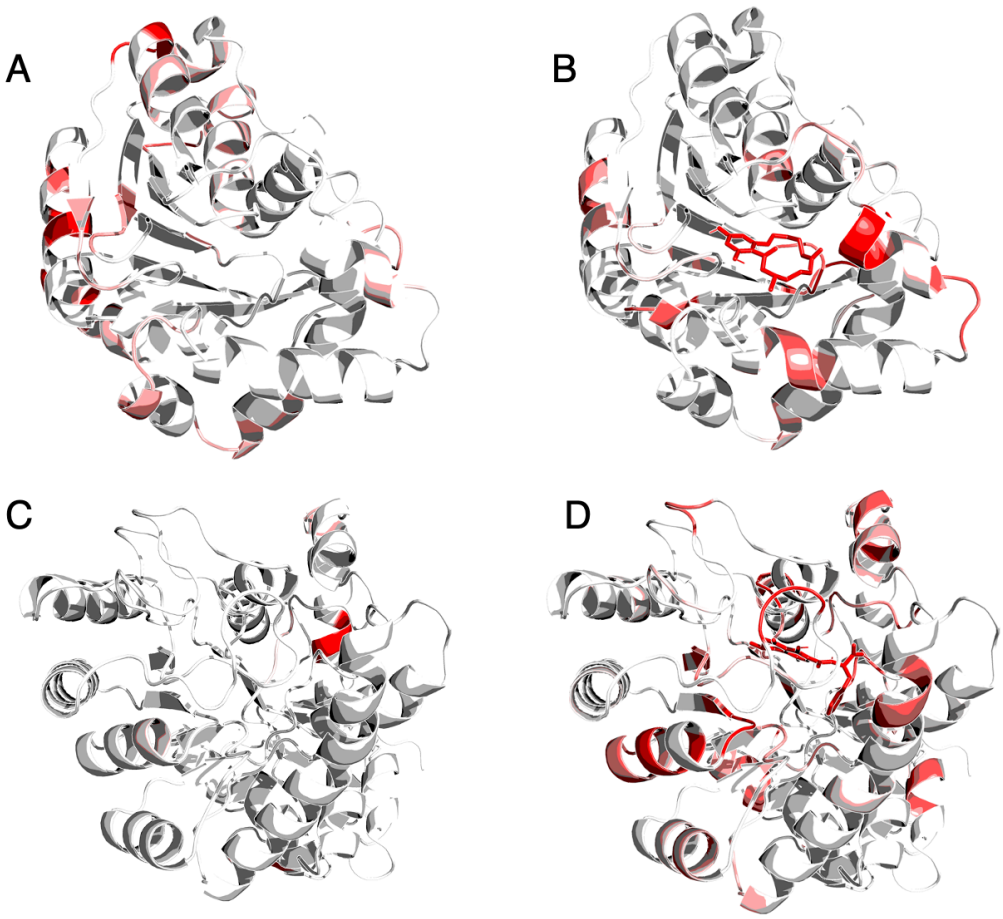

Fig. S10. Important regions on the enzymes identified by PU-EPP. (A) Attention weights of the coding layer (only based on features from enzymes) of ZH4, which cannot identify the enzyme's substrate-binding sites. (B) Attention weights of the interaction layer (based on features from both enzymes and substrates) of ZH4, which identified the regions around the binding sites. (C) Attention weights of the coding layer of OH1, which could not identify the enzyme's substrate-binding sites. (D) Attention weights of the interaction layer of OH1, which identified the regions around the binding sites.

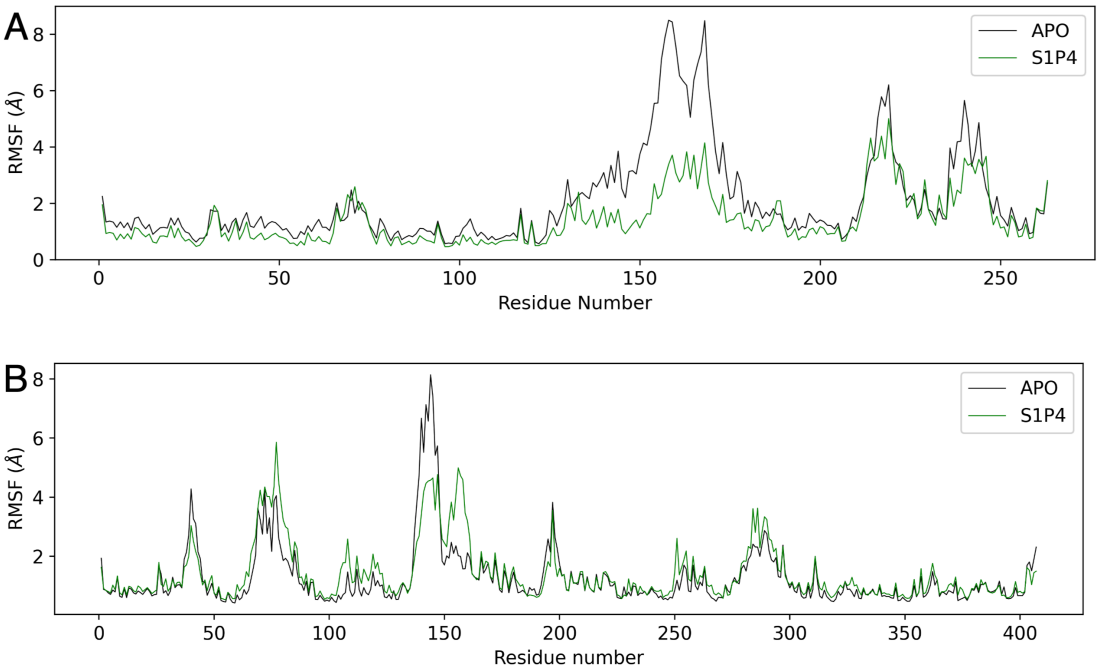

Fig. S11. Root mean square fluctuations (RMSF) of protein-heavy atoms before and after substrate binding. (A) RMSF of the ZH4 enzyme. (B) RMSF of the OH1 enzyme. The black line indicates the RMSF of enzymes before substrate binding. The green line indicates the RMSF of enzymes after substrate binding. The x-axis denotes the amino acids in enzymes. The y-axis denotes the RMSF scores of each amino acid.

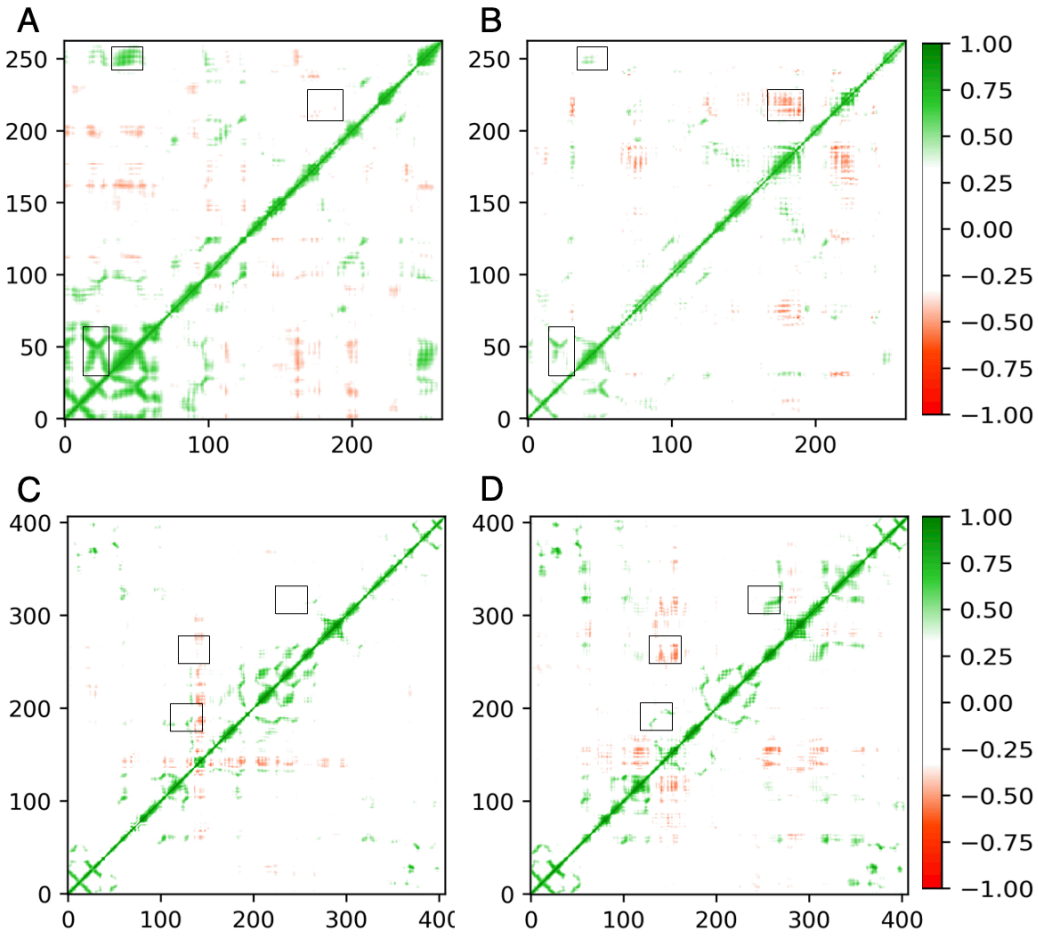

Fig. S12. Dynamic cross-correlation matrix (DCCM) analyses of proteins before (left panel) and after (right panel) substrate binding. The binding modes with the lowest MM-GBSA-binding free energy were selected for analysis. (A, B) DCCM of the ZH4 enzyme. Residues 30–60 vs 10–30, 244–260 vs 33–57, and 170–190 vs 210–227 have high correlations. (C, D) DCCM of the OH1 enzyme. Residues 170–190 vs 125–150, 135–160 vs 250–270, and 250–270 vs 300–320 have high correlations. The boxes highlight regions that show significant correlation changes upon substrate binding.

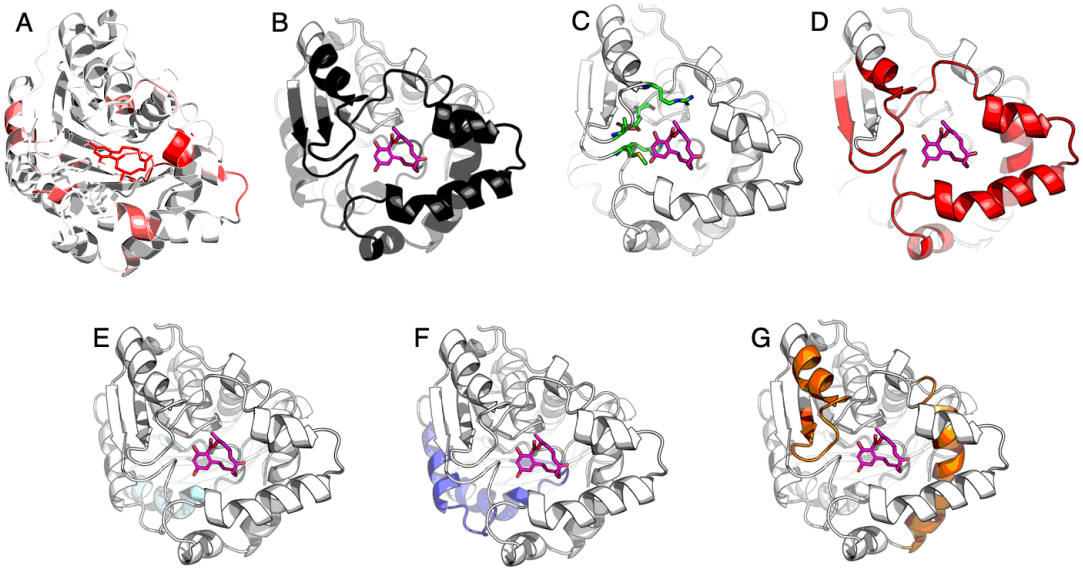

Fig. S13. Important regions on the ZH4 enzyme inferred via PU-EPP and molecular simulation. (A) Important regions identified by PU-EPP. (B) Important regions identified through molecular simulation. (C) Energetically important residues inferred through MM-GBSA analysis. (D) Allosterically important residues inferred through RMSF analysis. (E–G) Allosterically important residues inferred using DCCM analysis.

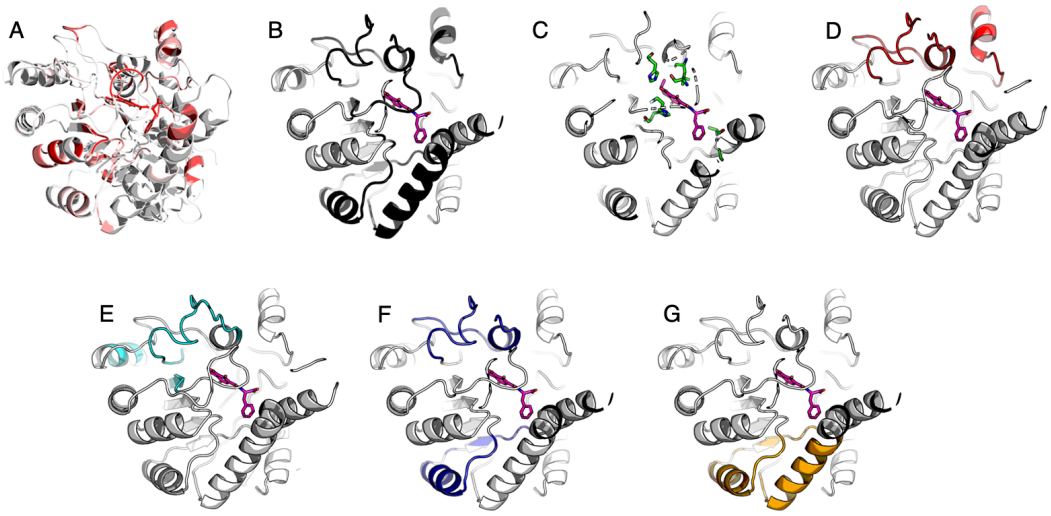

Fig. S14. Important regions on the OH1 enzyme inferred via PU-EPP and molecular simulation. (A) Important regions identified using PU-EPP. (B) Important regions identified through molecular simulation. (C) Energetically important residues inferred using MM-GBSA analysis. (D) Allosterically important residues inferred using RMSF analysis. (E–G) Allosterically important residues inferred through DCCM analysis.

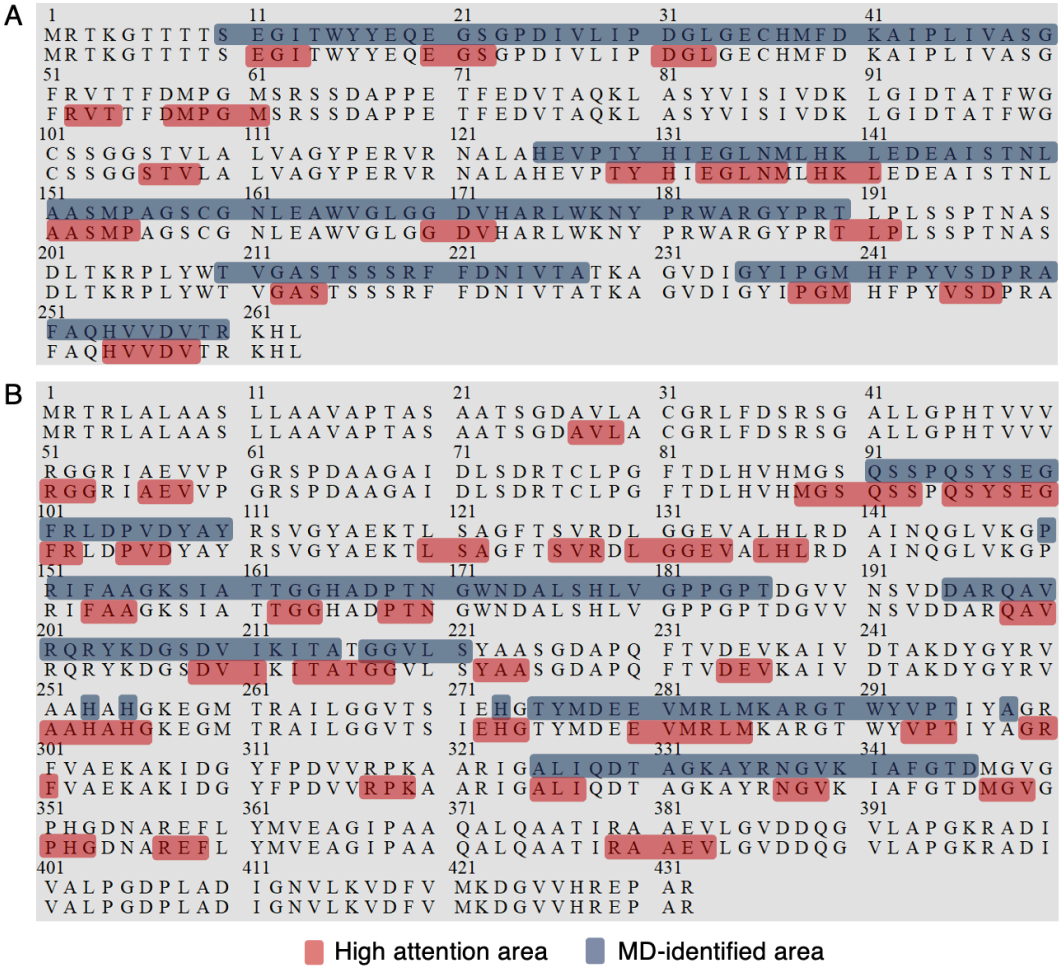

Fig. S15. Overview of important amino acid residues in the ZH4 and OH1 enzymes inferred through PU-EPP and molecular simulation. Red indicates the important residues inferred via PU-EPP, while blue indicates the important residues inferred via molecular simulation.

Table S1. Datasets used in this study and their data sizes

Dataset	Number of enzymes	Number of substrates	Number of pairs
Dataset for training PU-EPP	170,179	5,837	7,095,469
Phosphatase	218	108	23,544
Esterase	146	96	14,016
Glycosyltransferase	54	90	4,298
β-keto acid cleavage enzymes	161	17	2,737
Halogenase	42	62	2,604

Table S2. Rules for evaluating biotransformation reactions for toxin detoxification

Type	Category	Description	Software	Score
Molecular level	Availability of co-reactants	The availability of co-reactants was calculated according to the normalized frequency of co-reactants appearing in reactions. The more common the co-reactants, the higher the score. For reactions with multiple co-reactants, the lowest score of co-reactants was summarized into the final score.	Rhea [1]	0–1 (2 non-zero digits)
	Predicted lethal concentration 50% of the metabolite	> the parent form of toxins	ADMETLab [6]	0
		< the parent form of toxins		+1
	Synthetic accessibility of the metabolite	> the parent form of toxins	MOSES [7]	0
		< the parent form of toxins		+1
Reaction level	Biotransformation types	Occurred on the side chains of the molecular structure of mycotoxins	RDKit [8]	0
		Occurred on the scaffold of the molecular structure of mycotoxins		+1
	Reaction feasibility	Unfeasible	DeepRFC [9]	Exclude
		Feasible		Retain

Table S3. Representative ochratoxin A (OTA) biotransformation reactions and their feasibility scores

EC number	Description	Metabolic reaction	Score
3.5.1.-	Hydrolases acting on carbon-nitrogen bonds (except peptide bonds) in linear amides	OTA + H_2O >> Ochratoxin α + DL-phenylalanine	3.59
2.4.1.-	Hexosyltransferases	OTA + UDP-glucose + H_2O >> OTA-11-glucoside + UDP	1.02
2.7.2.-	Phosphotransferases transferring phosphorus-containing groups with a carboxy group as an acceptor	OTA + ATP >> OTA-27-phosphate + ADP + H+	0.12
1.14.16.-	Oxidoreductases acting on paired donors with reduced pteridine as one donor, and incorporation of one atom of oxygen into the other donor	OTA + O_2 + L-phenylalanine >> OTA-21-OH + 3-hydroxy-L-phenylalanine	0.01
3.1.1.-	Carboxylic-ester hydrolases	OTA + ethanol >> Ochratoxin C + H_2O	0

Table S4. Representative zearalenone (ZEA) biotransformation reactions and their feasibility scores

EC number	Description	Metabolic reaction	Score
3.1.1.-	Carboxylic-ester hydrolases	ZEA + H_2O >> hydrolyzed-ZEA	3.59
1.14.13.-	Oxidoreductases acting on paired donors with the incorporation of NADH or NADPH as one donor and one atom of oxygen as the other donor	ZEA + NADPH + O_2 + H+ >> ZEA-14-carboxyl + CO_2 + H_2O + NADP+	2.13
2.4.1.-	Hexosyltransferases	ZEA + UDP-glucose + H_2O >> ZEA-14-glucoside + UDP	1.02
1.1.1.-	Oxidoreductases acting on the CH-OH group with NAD+ or NADP+ as an acceptor	ZEA + NADPH + H+ >>Zearalenol + NADP+	0.13
2.7.1.-	Phosphotransferases with an alcohol group as an acceptor	ZEA + ATP >> ZEA-14-phosphate + ADP + H+	0.12

SI References:

1 A. Morgat, T. Lombardot, K. B. Axelsen, L. Aimo, A. Niknejad, N. Hyka-Nouspikel, E. Coudert, M. Pozzato, M. Pagni, S. Moretti, S. Rosanoff, J. Onwubiko, L. Bougueleret, I. Xenarios, N. Redaschi, A. Bridge, Updates in Rhea—an expert curated resource of biochemical reactions. *Nucleic Acids Res.* 45, D415–D418 (2017). https://doi.org:10.1093/nar/gkw990

2 D. Zhang, S. Ouyang, M. Cai, H. Zhang, S. Ding, D. Liu, P. Cai, Y. Le, Q. N. Hu, FADB-China: a molecular-level food adulteration database in China based on molecular fingerprints and similarity algorithms prediction expansion. *Food Chem.* 327, 127010 (2020). https://doi.org:10.1016/j.foodchem.2020.127010

3 Q. N. Hu, H. Zhu, X. Li, M. Zhang, Z. Deng, X. Yang, Z. Deng, Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. *PLoS One* 7, e52901 (2012). https://doi.org:10.1371/journal.pone.0052901

4 D. Bajusz, A. Racz, K. Heberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7, 20 (2015). https://doi.org:10.1186/s13321-015-0069-3

5 A. G. McDonald, S. Boyce, K. F. Tipton, ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.* 37, D593-597 (2009). https://doi.org:10.1093/nar/gkn582

6 J. Dong, N.N. Wang, Z.J. Yao, L. Zhang, Y. Cheng, D. Ouyang, A.P. Lu, D.S. Cao, ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J. Cheminform.* 10 (2018). https://doi.org:10.1186/s13321-018-0283-x

7 D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, A. Zhavoronkov, Molecular Sets (MOSES): a benchmarking platform for molecular generation models. *Front. Pharmacol.* 11, 565644 (2020). https://doi.org:10.3389/fphar.2020.565644

8 RDKit: Open-Source Cheminformatics Software. https://rdkit.org/ (2022).

9 Y. Kim, J. Y. Ryu, H. U. Kim, W. D. Jang, S. Y. Lee, A deep learning approach to evaluate the feasibility of enzymatic reactions generated by retrobiosynthesis. *Biotechnol. J.* 16, e2000605 (2021). https://doi.org:10.1002/biot.202000605