How to cite: *Angew. Chem. Int. Ed.* **2023,** *62,* e202301660
doi.org/10.1002/anie.202301660

Machine Learning Hot Paper

# Structure- and Data-Driven Protein Engineering of Transaminases for Improving Activity and Stereoselectivity

*Yu-Fei Ao,* Shuxin Pei, Chao Xiang, Marian J. Menke, Lin Shen,* Chenghai Sun, Mark Dörr, Stefan Born, Matthias Höhne, and Uwe T. Bornscheuer**

**Abstract:** Amine transaminases (ATAs) are powerful biocatalysts for the stereoselective synthesis of chiral amines. Machine learning provides a promising approach for protein engineering, but activity prediction models for ATAs remain elusive due to the difficulty of obtaining high-quality training data. Thus, we first created variants of the ATA from *Ruegeria* sp. (3FCR) with improved catalytic activity (up to 2000-fold) as well as reversed stereoselectivity by a structure-dependent rational design and collected a high-quality dataset in this process. Subsequently, we designed a modified one-hot code to describe steric and electronic effects of substrates and residues within ATAs. Finally, we built a gradient boosting regression tree predictor for catalytic activity and stereoselectivity, and applied this for the data-driven design of optimized variants which then showed improved activity (up to 3-fold compared to the best variants previously identified). We also demonstrated that the model can predict the catalytic activity for ATA variants of another origin by retraining with a small set of additional data.

## Introduction

Chiral amines are frequently used as key chiral building blocks for the synthesis of bioactive pharmaceuticals and agrochemicals, and therefore have attracted particular attention by synthetic chemists.[1] In the last few decades, several environmentally friendly biocatalytic strategies have been developed to access chiral amines with high selectivity, and various enzymes such as hydrolases, oxidoreductases and transferases have been employed.[2] Amine transaminases (ATAs), a subgroup of pyridoxal-5′-phosphate (PLP)-dependent enzymes that catalyze the asymmetric amination of a ketone to the corresponding amine, usually exhibit high enantioselectivity and broad substrate tolerance, and are thus widely applied for the preparation of optically pure amines.[3] The most popular example is the (*R*)-ATA-catalyzed asymmetric synthesis of the antidiabetic drug (*R*)-sitagliptin with >99.95% optical purity developed by researchers from Merck & Co., Inc. and Codexis, Inc., which replaced the previously developed asymmetric chemical hydrogenation.[4] Based on their enantiopreference, transaminases are classified into two types, (*S*)- and (*R*)-selective enzymes, which belong to the fold types I and IV, respectively, and differ substantially in their protein structure. So far, the majority of the identified ATAs are (*S*)-selective and their substrate binding site is defined by a large and a small binding pocket, which thus can accommodate

[*] Dr. Y.-F. Ao, Dr. C. Xiang, M. J. Menke, Dr. C. Sun, Dr. M. Dörr, Prof. M. Höhne, Prof. U. T. Bornscheuer
Department of Biotechnology and Enzyme Catalysis, Institute of Biochemistry, University of Greifswald
Felix-Hausdorff-Str. 4, 17487 Greifswald (Germany)
E-mail: uwe.bornscheuer@uni-greifswald.de

Dr. Y.-F. Ao
Beijing National Laboratory for Molecular Sciences, CAS Key Laboratory of Molecular Recognition and Function, Institute of Chemistry, Chinese Academy of Sciences
Zhongguancun North First Street 2, Beijing, 100190 (China)
and
University of Chinese Academy of Sciences
Yuquan Road 19(A), Beijing, 100049 (China)
E-mail: aoyufe@iccas.ac.cn

S. Pei, Prof. L. Shen
Key Laboratory of Theoretical and Computational Photochemistry of Ministry of Education, College of Chemistry, Beijing Normal University
Xinjiekouwai Street 19, Beijing, 100875 (China)
E-mail: lshen@bnu.edu.cn

Dr. S. Born
Technische Universität Berlin, Chair of Bioprocess Engineering
Ackerstraße 76, 13355 Berlin (Germany)

Prof. L. Shen
Yantai-Jingshi Institute of Material Genome Engineering
Nanchang Road 48, Yantai, Shandong, 265505 (China)

Prof. M. Höhne
Technische Universität Berlin, Department of Chemistry/ Biocatalysis
Müller-Breslau-Str. 10, 10623 Berlin (Germany)

ketones bearing a large and a small substituent at the carbonyl carbon.[3]

Protein engineering is powerful in improving various enzymatic catalytic features such as catalytic activity, selectivity, substrate promiscuity, and enzymatic stability.[5] So far, the most effective method of modifying substrate promiscuity, catalytic activity, and stereoselectivity of ATAs is to mutate crucial residues adjacent to the two binding pockets by protein engineering to alter their steric and electronic environment.[6] Recently, a series of rational-design or screening methods, such as structure-dependent rational design,[7] mechanism-guided computational design[8] and a growth selection[9] method, have been successfully applied to engineer ATAs for higher activity, broader substrate scope and improved stereoselectivity towards particular substrates. However, these strategies require a significant amount of computational and/or experimental effort to optimize the biocatalyst.

Machine learning (ML) has been shown to be a new powerful and competitive approach also suitable for protein engineering.[10] It has been applied to modulate the catalytic activity and/or stereoselectivity of a range of enzymes such as epoxide hydrolase,[11] a nitric oxide dioxygenase which was evolved for enantioselective carbene Si–H insertion,[12] an imine reductase,[13] an acyl-ACP reductase[14] and others.[15] This data-driven strategy can identify catalytic patterns in the collected data to predict previously unnoticed but promising variants such as new combinations of substitutions. Therefore ML is expected to significantly reduce the computational and experimental efforts required by traditional strategies.[7,8] Generally speaking, the success of an ML predictor crucially depends on the quality of the data used for training. However, the efficient acquisition of high-quality data has become a major challenge limiting the application of ML to biocatalyst design, due to the lack of diversity, sufficient and well-prepared samples in the existing datasets for model training.[10] Commonly, data from rational protein design experiments are used to provide a high-quality training set for ML, while ML predictive models can help to quickly obtain new and promising variants. Obviously, it is attractive to combine these two complementary strategies for protein engineering, however, such a combination method has rarely been published.[12]

For transaminases, although a ML model for their stability prediction has been reported recently,[16] a ML predictor for catalytic activity and stereoselectivity has not been reported. This is partly due to the time-consuming and labor-intensive measurement of catalytic activity and stereoselectivity towards different substrates, which makes data collection difficult. This applies especially for stereoselectivity, which results in low quality ML training data and thus limits the predictive capability of ML predictor for new data. Therefore, an assay for rapid measurement of catalytic activity and stereoselectivity is required to create high-quality datasets.

Our group had developed a rapid and sensitive photometric acetophenone assay for transaminase-catalyzed reactions, which enables to determine catalytic activity[17] and as well stereoselectivity by comparing the respective specific activity using enantiopure (*R*)- or (*S*)-substrates (see insert in Figure 1). In previous work, we already successfully improved the substrate scope and catalytic activity of several different ATAs by protein engineering,[18] which identified a series of crucial residues as potentially influential for the enzyme-substrate association. This paved the way to rationally design variants and to obtain high-quality data for this study.

Here we report the structure-guided rational design of ATA variants to obtain high quality catalytic activity and stereoselectivity data in the first step, followed by the building of the ML predictor, which uses a representation of the enzyme and substrate by descriptors developed from structural and biochemical considerations. This information then trained standard ML models for the prediction task. Finally, we demonstrate its application for the design of new variants that display higher activity and desired stereoselectivity towards substrates that have never been used for the construction of ML models (see flow chart in Figure S1 and Table S4, Supporting Information).

## Results and Discussion

To comprehensively evaluate the impact of protein engineering on biocatalytic reactions and to obtain a diverse set of data, we focused on a series of ATA variants with increased or reduced catalytic activity and increased, reduced or even reversed stereoselectivity. To identify a suitable scaffold that meets these requirements, several wild-type ATAs from fold class I and some interesting variants discovered previously by our group were evaluated in the model reaction using enantiopure (*R*)- or (*S*)-1-phenylethylamine (PEA) as amine donors and pyruvate as the acceptor (Figure 1, Figure S2). Among all the candidates, the ATA from *Ruegeria* sp. TM1040 (abbreviated as 3FCR, according to its PDB code) turned out to be the most suitable scaffold, which displayed a minor specific activity towards (*S*)-PEA, but structure-function analyses revealed important key residues that modulate activity.[18] Some of its previously designed variants,[19] such as 3FCR-Y59W/Y87F/T231A (3FCR-3M) and 3FCR-Y59W/Y87F/T231A/Y152F (3FCR-4M), showed much higher specific activity towards (*S*)-PEA, which demonstrates its potential for altering its properties through protein engineering. However, none of the enzymes and variants studied so far displayed any activity toward (*R*)-PEA, as training data for a stereoselectivity predictor were missing. Therefore, we aimed at first to reverse the (*S*)-selectivity to (*R*)-selectivity of these enzymes by protein engineering to diversify the stereoselectivity data, and secondly to further investigate the activity of different variants towards (*S*)- and (*R*)-PEA to obtain a diversified activity dataset.

According to our previous work on protein engineering of (*S*)-selective ATAs,[18,19] in combination with molecular docking simulations of wild-type 3FCR and (*S*)-PEA (Figure 2), we selected seven crucial residues that are likely to affect its activity and stereoselectivity. Four of these residues (Y87, S19, Y152 and S155) are located in the small binding

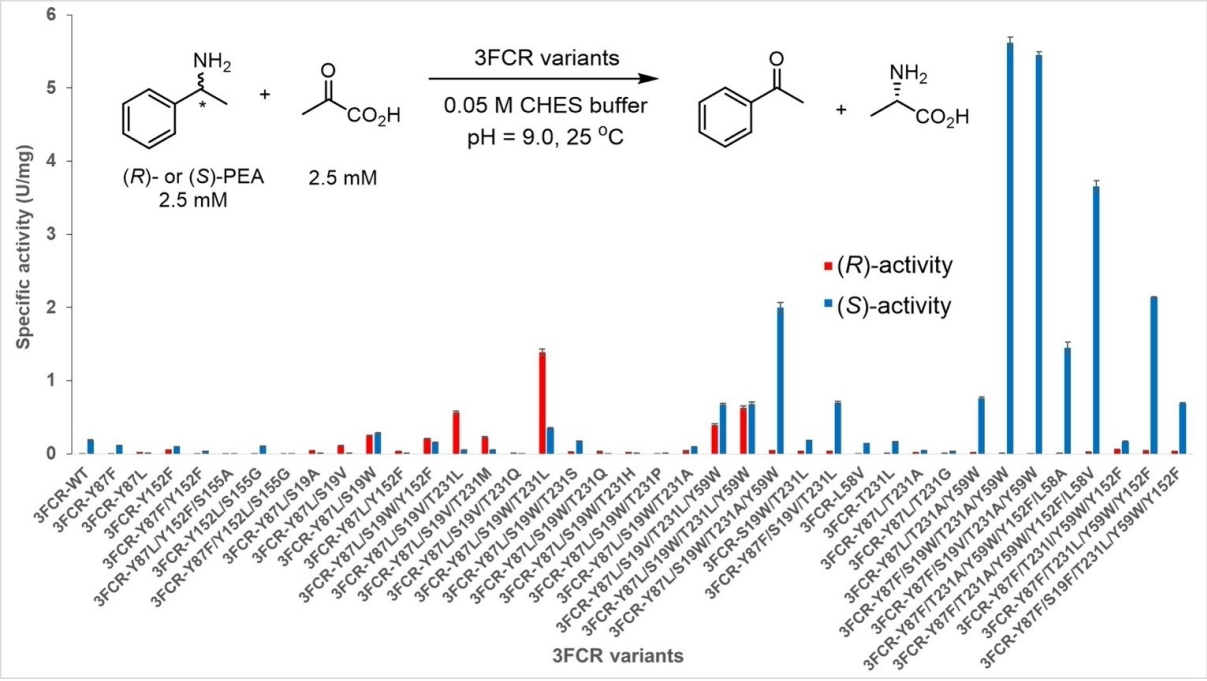


***Figure 1.*** Specific activity of 3FCR variants determined using the acetophenone assay.[17] For this, both (*S*)- and (*R*)-PEA were assayed with pyruvate as amino acceptor. One unit (U) activity was defined as the formation of 1 μmol acetophenone per minute. All measurements were performed in triplicates and the mean values are indicated in the figure as error bars. The specific activity toward (*R*)- and (*S*)-PEA is shown in red and blue bars, respectively. Detailed data for specific activity measurements are given in Table S6.

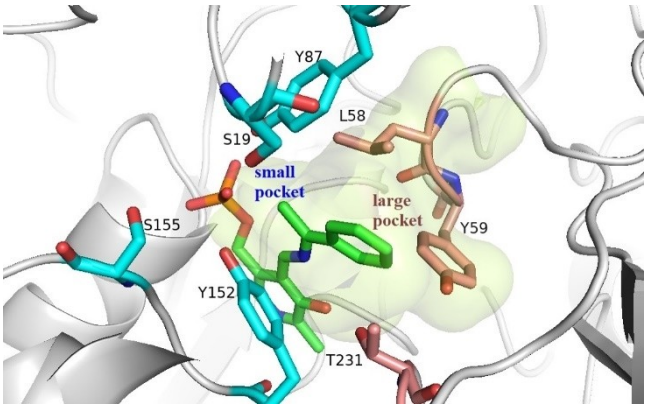


***Figure 2.*** External quinonoid intermediate formed by substrate (*S*)-PEA and the cofactor PLP located in the active site of 3FCR. The quinonoid (green sticks) of (*S*)-PEA was modelled into the substrate binding pocket of 3FCR. The protein is shown as white cartoon. The carbon atoms of the residues in the small and large binding pocket are colored cyan and wheat, respectively. The oxygen and nitrogen atoms of different residues are colored red and blue.

pocket and the remaining three residues (T231, Y59 and L58) are in the large binding pocket. Based on our previously developed rational-design strategy for reversing the enantioselectivity,[20] we planned to enlarge the small binding pocket of wild-type 3FCR and subsequently to shrink the large pocket. As shown in Figure 2, among the seven variants screened in the first round (Table S6, entries 4–10), the variants 3FCR-Y87L and 3FCR-Y152F showed an initial activity toward (*R*)-PEA and 3FCR-F87L even displays slight (*R*)-preference. The next two rounds of saturation mutagenesis targeted S19 and T231 using 3FCR-Y87L as the template and this resulted in a series of (*R*)-selective or highly active variants (Table S6, entries 11–24). Compared to the initial activity of the 3FCR-WT, the specific activity of the variants was increased up to 2000-fold, as shown for (*R*)-PEA using the variant 3FCR-Y87L/S19W/T231L (Table S6, entry 19). The specific activity ratio towards (*R*)-PEA and (*S*)-PEA could be increased up to 3000-fold for variant 3FCR-Y87L/S19V/T231L (Table S6, entry 16). Further replacing of the seven key residues using 3FCR-Y87L/S19W/T231L and 3FCR-Y87L/S19V/T231L as templates did not obtain variants with higher (*R*)-activity and selectivity (Table S6, entries 25–34). Variants 3FCR-3M and 3FCR-4M displayed much higher activity towards (*S*)-PEA[19c] and hence a series of variants were designed by substitution of the key residues from these two scaffolds to further evaluate the effect of them on the catalytic activity (Table S6, entries 35–41). The results show that the substitution of T231L or Y152F had a negative effect on (*S*)-activity.

Among the above mentioned 40 3FCR variants, eight variants displayed weak activity toward (*R*)- and (*S*)-PEA and were thus excluded. We chose the other 32 variants and wild-type 3FCR to investigate their specific activity toward 13 pairs of enantiopure substrates (13 (*R*)- and 13 (*S*)-compounds). These substrates (Figure 3) contain different substituents leading to different steric and electronic effects, and thus their analysis should lead to a clearer perception of

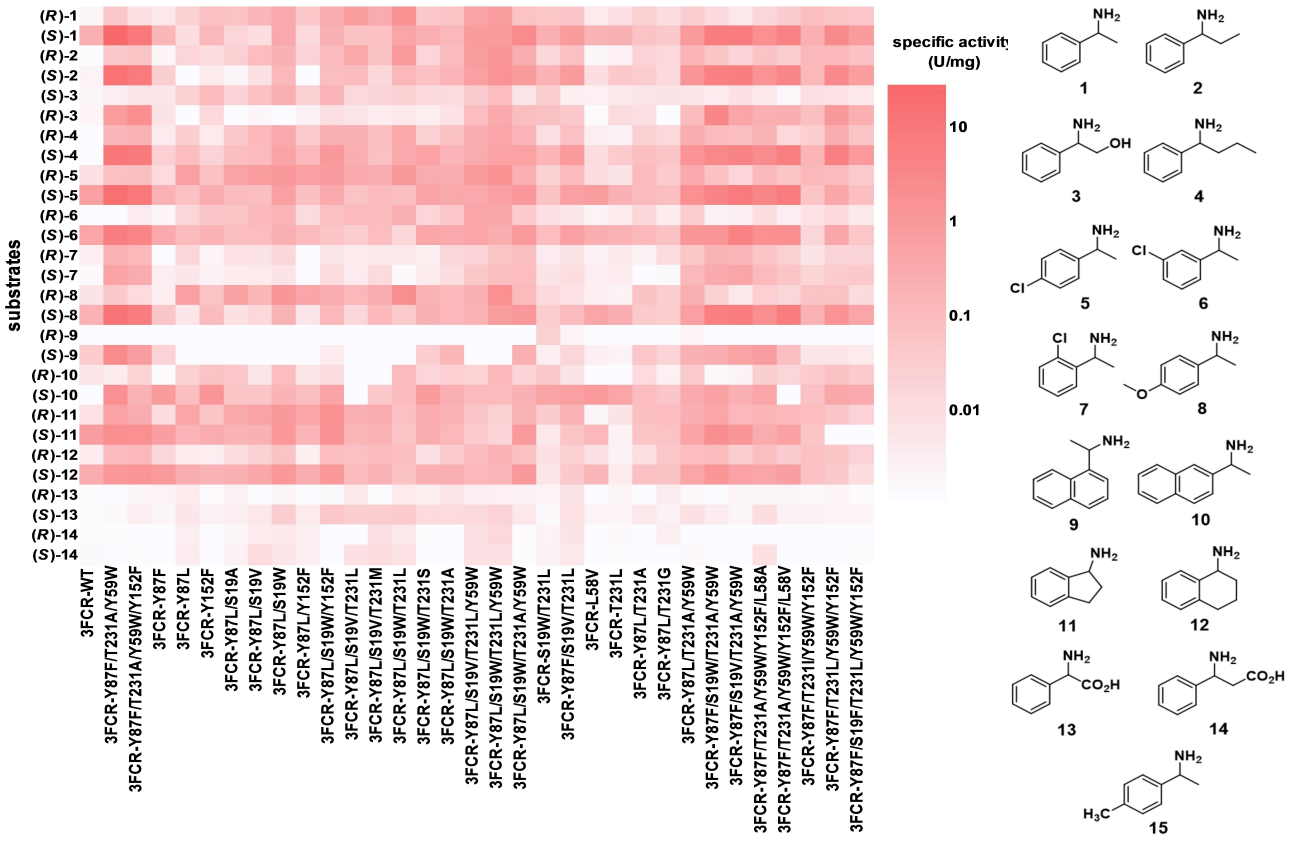


***Figure 3.*** Specific activity of 3FCR variants toward (*S*)- and (*R*)-enantiomers of amines **1**–**14**. All measurements were performed in triplicates and the mean values are indicated by a color gradient. Note that compound (*R*)-**3** has the same binding mode as all other (*S*)-compounds in the active site, but due to the hydroxyl substituent, its absolute configuration is (*R*) according to the CIP nomenclature priority rules. The detailed data for specific activity are given in the Supporting Information file named "SI-excel".

the structure-function relationships of the ATAs with their substrates.

As shown in Figure 3, most of these substrates were converted by the 3FCR variants, but when the substrate contains a carboxyl and a phenyl group (substrates **13** and **14**), the activity is rather low. Presumably this is because the carboxylic acid function of the substrate is usually bound in the large binding pocket by a flexible Arg side chain. However, if the large binding pocket is occupied, the phenyl group of the substrate has to be accommodated into the small binding pocket. When the substrate contains a hydroxyl group (substrate **3**), the activity is generally lower than for substrate **2**, even though both have almost the same steric effect but differ in their hydrophobic properties. Therefore, a hydrophilic group is not well accepted by 3FCR variants in comparison to hydrophobic groups in the substrates. In addition to the effect of hydrophobic or acidic groups mentioned above, the steric effect of substrates also has an important influence on the specific activity. The presence of a large substituent such as a naphthyl group in the substrate usually means that the activity is low, especially for the (*R*)-substrates that require the binding of the large substituent within the small binding pocket (substrates **9** and **10**). Regarding the substituent pattern on the benzene ring, an *ortho*-substitution (substrate **7**) usually causes a more significant reduction in activity than *para*- and *meta*-substitutions (substrates **5** and **6**), implying that the catalytic activity is more sensitive to the steric hindrance changes of *ortho*-positioned substituents.

Among these 3FCR variants, 3FCR-Y59W/Y87F/T231A and its single-mutation variants such as 3FCR-Y59W/Y87F/T231A/S19W and 3FCR-Y59W/Y87F/T231A/Y152F displayed the highest specific activity toward (*S*)-substrates, which probably results from the larger and/or more hydrophobic large-binding pocket that can accept a wider range of substrates. On the contrary, 3FCR-F91L/S19W and its variants such as 3FCR-F91L/S19W/T231L, 3FCR-F91L/S19V/T231L and 3FCR-F91L/S19W/T231L/Y59W probably have a larger small-binding pocket and thus show higher activity toward (*R*)-substrates. When the above variants also have a shrinked large-pocket, such as 3FCR-F91L/S19W/T231L and 3FCR-F91L/S19V/T231L, then they usually displayed stronger (*R*)-selectivity, which is consistent with our previous findings.

After collecting the above biocatalytic reaction data, we next built the ML prediction model that utilizes substrate structures and transaminase sequences as input variates and predicts specific activities as output results. Although a series of 3D structure- or 1D sequence-based

descriptors have been successfully applied in biocatalytic reactions,[21,22] their performance strongly depends on the quantity of training data, which is still an obstacle in present research. This is due to the fact that the activity of a system is usually dominated by a few functional groups of the substrate and a small active site of the enzyme. In other words, a lot of information included in the 3D structure- or 1D sequence-based descriptors is not relevant to the properties we wish to create, which depresses the performance of ML without a large number of data. Considering the important role of steric and electronic effects of the substrate on the enzyme, we encoded the most relevant information of substrates and amino acids residues within the ATA with a modified one-hot code. As shown in Table S1, each amino acid is encoded into two elements (denoted as A and B), which respectively represents the electronic and steric properties of a given amino acid residue. Each amine substrate can be divided into a large-pocket binding group and a small-pocket binding group, and each binding group is encoded into four elements (A, B, C and D, see Table S2) to describe its electronic and steric properties. Both 1D and 3D information of the substrates and amino acid residues of the enzymes are embedded in the code. Each element of the code acts as one descriptor of ML models.

After building the descriptors for amino acids and substrates, we started to create the input file for machine learning. 3FCR contains 456 amino acids, but only 7 of them had been changed in the 40 variants studied, and thus the other 449 amino acids were deleted in the feature selection step. Finally, 14 amino acid descriptors (7 amino acids × 2 elements) and 8 substrate descriptors (4 elements for the large-pocket binding group and 4 elements for the small-pocket binding group) were used to form the feature set. The label of data is the natural logarithm of the specific activity values. Then we started to build a statistical model to correlate the specific activity with 22 descriptors. There are several machine learning algorithms that are used to solve biocatalytic-related problems.[10] Here we applied four regression models, these are: random forest (RF), support vector regression (SVR), kernel ridge regression (KRR), and gradient boosting regression tree (GBRT) to predict the reaction activity. Different data split percentages have little effect on the prediction performance (Figure S9), so we decided to randomly split the data into training (90 %) and test (10 %) sets. Each of these models is in fact given as a family of models parametrized by so-called "hyper-parameters". Once the hyper-parameters are chosen, one can fit the models to data. However, unlike first-principles models with few parameters, such ML models can overfit and become useless for the prediction of new case studies. The hyper-parameters have to be tuned so that the predictive performance is optimized (Table S3). For this purpose, the expected prediction error on new cases is estimated by 10-fold cross-validation, which means to fit the model on training subsets of the data and validate the predictions on validation subsets. The details can be found in the Supporting Information. RF and GBRT performed better than SVR and KRR (Figure S3). The GBRT predictor provided the best $R^2$- (0.803) and RMSD-values (1.083), which was denoted as GBRT-1 (Figure 4A) and was applied in this research.

Feature importance analysis based on GBRT-1 (an assessment of the contribution of each descriptor to this model) reveals the key role of residues 87, 231 and 19 (Figure S4). Therefore, we focused the virtual screening target on these substitutions and obtained predicted data for 8192 variants. According to the prediction results, the change of residue 87 had a small effect on the predicted value, which is inconsistent with the results of the mutation experiments. This is most likely due to the lack of samples with a mutation at position 87. We therefore focused on the mutation at this position and thus selected nine variants with higher predicted activity toward (*R*)-substrates or (*R*)-selectivity. According to the experimental results (Figure S5A), 3FCR-Y87C/S19W/T231L displayed higher (*R*)-selectivity toward 1-phenyl-butan-1-amine (**4**), 1-(4-chlorophenyl)-ethan-1-amine (**5**) and 1-(naphthalen-2-yl)-ethan-1-amine (**10**) than the other 49 variants.

The GBRT-1 was rebuilt based on the increased dataset containing all previous data and the above new data (9 predicted variants × 28 substrates). The updated predictor was used to design new variants to predict the activity toward (*S*)- or (*R*)-substrates. Among the variants suggested by virtual screening, we selected 13 variants with predicted higher activity toward (*S*)- or (*R*)-substrates and studied their activity by experiments (Figure S5B). To our delight, 3FCR-Y87C/S19W/T231L/Y59W displayed 3.5-fold higher activity toward (*R*)-1-phenylpropan-1-amine ((*R*)-**2**) and 8 % higher activity toward (*R*)-1-phenyl-butan-1-amine ((*R*)-**4**) than the best one among the other 62 variants. The variant (Y87W/T231A/Y59W) also displayed higher activity toward (*S*)-1-(3-chlorophenyl)-ethan-1-amine ((*S*)-**7**) and (*S*)-1-(naphthalen-2-yl)-ethan-1-amine ((*S*)-**10**) than the other 62 variants, especially toward (*S*)-**10**, which was improved three times compared to 3FCR-Y87F/T231A/Y59W. These results demonstrate the successful application of the ML predictor to improve variants.

We further added new experimental data with the selected 13 variants to the current dataset and retrained GBRT-1 again. The newly updated model was applied to design 3FCR variants with higher activity for (*R*)- and (*S*)-1-(4-methylphenyl)-ethan-1-amine (**15**) by virtual screening. Among these 8124 screening variants, 3FCR-F91L/S19W/T231L/Y59W and 3FCR-F91F/T231A/Y59W displayed the highest activity toward (*R*)- and (*S*)-**15**, respectively. To our delight, the measured activities of these two variants were similar to the predicted values (Figure 4B). To further evaluate the updated GBRT-1 system for the prediction of activity toward (*R*)- and (*S*)-**15**, we measured the reactions of all variants and the results achieved a nice regression with $R^2$ 0.846 (Figure S6).

Finally, the GBRT-1 predictor was reconstructed with all available data, resulting in an $R^2$ up to 0.905 (Figure 4C). Its feature importance result (Figure 4D) displays more weight on amino acid features compared to the results shown in Figure S4, which probably results from the greater variation in amino acids from additional experiments. Element B of

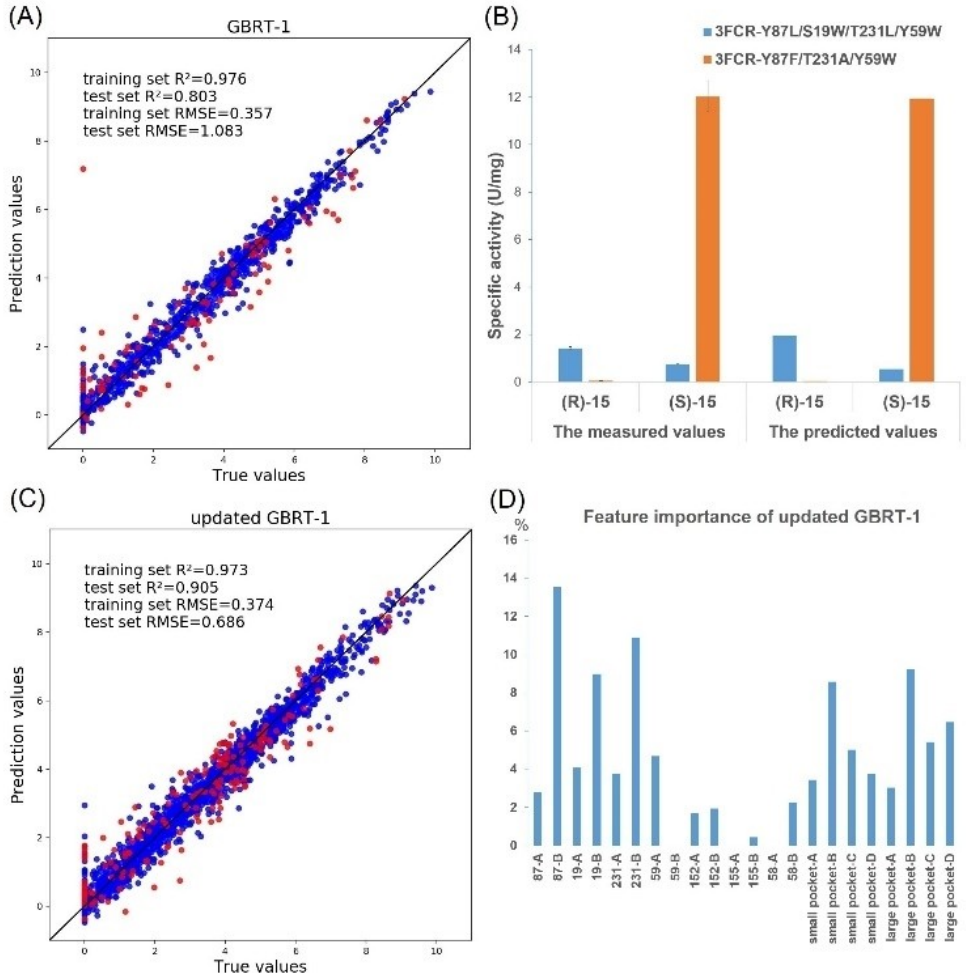


***Figure 4.*** Regression performances of the prediction values and true values from the initial dataset using GBRT-1. (A) Comparison of measured and predicted specific activity values of 3FCR-F91L/S19W/T231L and 3FCR-F91F/T231A/Y59W toward substrate (*R*)-**15** and (*S*)-**15**. (B) Regression performances (C) and feature importance (D) of the updated GBRT-1 predictor trained by all existing data. The true value of each point corresponds to the natural logarithm of the mean specific activity values (mU/mg) of three independent experiments. Blue and red dots in the figures represent data from training and test set, respectively. A detailed explanation of the descriptors in (D) is given in Tables S1 and S2.

residues 87, 231 and 19 were extracted as important features, which implies their steric effects on the predicted activity values. The insignificance of residues 155 and 58 may result from the small amount of data available for mutations at these residues. Different from amino acid features, the impacts of all substrate features on this model were observed. Especially, the large importance of element B of both large- and small-pocket demonstrates that the *ortho* substituent in the benzene ring plays an essential role on catalytic activity of 3FCR.

The ATA from *Ruegeria pomeroyi* (abbreviated as 3HMU, according to its PDB code) is another important (*S*)-selective transaminase which is widely used in chiral amine synthesis.[18,19] Although it shows totally different catalytic activity toward (*S*)-PEA compared to 3FCR-WT (Figure S2), it has a similar 3D structure, especially in the active site region. Therefore, we hoped to extend the application of GBRT-1 to predict the catalytic activity of 3HMU. Based on a protein structure alignment, seven key residues were identified (Table S7). To compare the catalytic activity and to calibrate the predictor GBRT-1, we designed three 3FCR variants (3FCR-Y87F/S19F/T231A/Y59W, 3FCR-Y87F/S19F/T231A/ Y59W/Y152F and 3FCR-Y87L/S19F/T231A/Y59W, Table S7, entries 1–3) as they have key residues similar to 3HMU-WT and 3HMU-F91L (Table S7, entries 4–5), respectively. Combining the activity data of these three 3FCR variants with previous 3FCR data, the ML model denoted as GBRT-2 was obtained (Figure S7A). In order to assess the predictive ability of GBRT-2, we designed several 3HMU variants containing the corresponding mutations of these key residues (Table S7, entries 6–15), and measured their specific activities toward 30 substrates. Unfortunately, the results showed that the measured values and predicted values show only low correlation (Figure S7B). This is due to the fact that the catalytic activity of the 3FCR and 3HMU variants are usually quite different, which indicates that the activity of different ATAs is influenced not

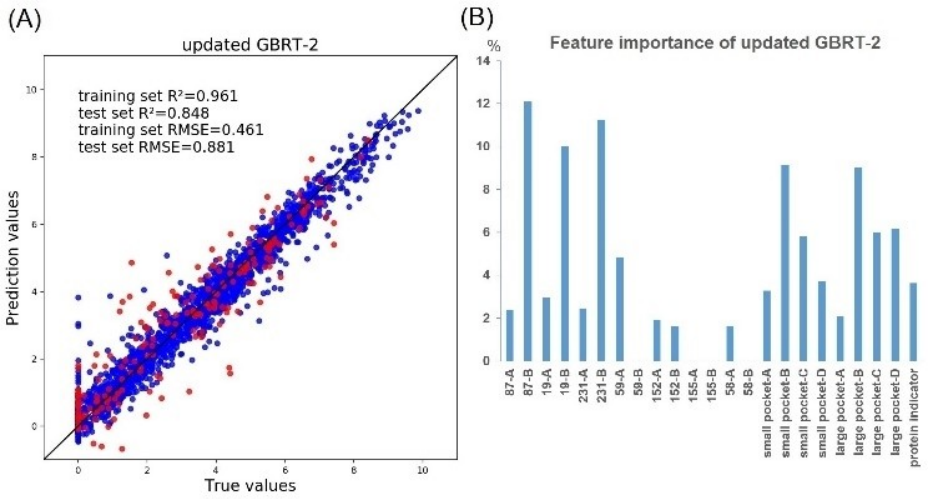


***Figure 5.*** Regression performances of the prediction values and true values for 3FCR and 3HMU variants using the updated GBRT-2 predictor (A). Feature importance of the updated GBRT-2 (B). The true value of each point corresponds to the natural logarithm of the mean specific activity values (mU/mg) of three independent experiments. Blue and red dots in the figure represent data from training and test sets, respectively. All available specific activity data are shown in Figure S8. The detailed measured specific activity data are shown in the Supporting Information file named "SI-excel".

only by the seven key residues shared in the active site, but also by other complex factors such as PLP-protein interactions. Although the two transaminases are in the same superfamily and share a highly similar fold, there are slight but significant differences of the backbone structures, e.g., in loop regions near to the active site, which might explain the lower quality of the predictions. An indicator variable that distinguishes between 3HMU and 3FCR was appended to the representation, thus allowing models to interpret the descriptors conditional on the two ATAs. The GBRT-2 predictor was constructed based on the new set of descriptors. 90 % of 3HMU catalytic data were randomly selected and combined with 3FCR data to train the updated GBRT-2 (Figure 5A). The remaining 10 % of 3HMU data composed the test set. To our delight, the updated GBRT-2 resulted in an $R^2$-score of up to 0.764 on the 3HMU test set (Figure S7C). Although the test set of GBRT-2 and the updated GBRT-2 are of different size, the introduction of the protein identifier descriptor indeed improves the $R^2$ score for 3HMU (Figure 5B).

## Conclusion

A high-quality dataset obtained by structure-dependent protein engineering, as well as a modified one-hot code that represents the most relevant electronic and steric properties of the key amino acids and substrates, paved the way for building an activity predictor for the amine transaminase 3FCR. We could demonstrate that the model can predict the catalytic activity of 3FCR variants for various amine substrates by virtual screening, which was then confirmed by experimental data. Thus, we have demonstrated the application of this simple but practical predictor in the data-driven rational design of 3FCR variants toward different substrates to obtain higher activity and especially also inverted stereoselectivity. With the help of the protein category descriptor and the design of calibration variants, we also extended this ML model to the amine transaminase 3HMU. The performance was acceptable but still far from perfect. Considering the limitations of this model in predicting the catalytic activity of ATAs involving complex substrates or other origin of ATAs, we will continue to improve this ML predictor by adding more data and selecting appropriate descriptors and algorithms in the future. This will also address our observation that the engineering effort is much more complex than simply increasing or decreasing the pocket size of the enzyme.

## Acknowledgements

Financial support from the National Key Research and Development Program of China (2019YFA0709400) to LS, the German Research Foundation (NFDI4CAT) to UTB, the National Natural Science Foundation of China (21977098, 22193041) to YFA and LS, the German Federal Ministry of Education and Research Program (01DD20002A to SB and 01DD20002C to UTB/MD) are gratefully acknowledged. Open Access funding enabled and organized by Projekt DEAL.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

**Keywords:** Biocatalysis · Catalytic Activity · Machine Learning · Stereoselectivity · Transaminases


[1] T. C. Nugent, *Chiral amine synthesis: methods, developments and applications*, 1st ed., Wiley-VCH, Weinheim, **2010**, p. 523.

[2] a) E. Alfonzo, A. Das, F. H. Arnold, *Curr. Opin. Green Sustainable Chem.* **2022**, *38*, 100701; b) W. Zawodny, S. L. Montgomery, *Catalysts* **2022**, *12*, 595; c) S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, U. T. Bornscheuer, *Angew. Chem. Int. Ed.* **2021**, *60*, 88–119; d) C. Ortiz, M. L. Ferreira, O. Barbosa, J. C. S. dos Santos, R. C. Rodrigues, Á. Berenguer-Murcia, L. E. Briand, R. Fernandez-Lafuente, *Catal. Sci. Technol.* **2019**, *9*, 2380–2420; e) M. D. Patil, G. Grogan, A. Bommarius, H. Yun, *ACS Catal.* **2018**, *8*, 10985–11015; f) J. H. Schrittwieser, S. Velikogne, M. Hall, W. Kroutil, *Chem. Rev.* **2018**, *118*, 270–348; g) J. Mangas-Sanchez, S. P. France, S. L. Montgomery, G. A. Aleku, H. Man, M. Sharma, J. I. Ramsden, G. Grogan, N. J. Turner, *Curr. Opin. Chem. Biol.* **2017**, *37*, 19–25.

[3] a) L.-C. Yang, H. Deng, H. Renata, *Org. Process Res. Dev.* **2022**, *26*, 1925–1943; b) S. A. Kelly, S. Pohle, S. Wharry, S. Mix, C. C. R. Allen, T. S. Moody, B. F. Gilmore, *Chem. Rev.* **2018**, *118*, 349–367; c) F. Guo, P. Berglund, *Green Chem.* **2017**, *19*, 333–360; d) I. Slabu, J. L. Galman, R. C. Lloyd, N. J. Turner, *ACS Catal.* **2017**, *7*, 8263–8284; e) M. Fuchs, J. E. Farnberger, W. Kroutil, *Eur. J. Org. Chem.* **2015**, 6965–6982.

[4] C. K. Savile, J. M. Janey, E. C. Mundorff, J. C. Moore, S. Tam, W. R. Jarvis, J. C. Colbeck, A. Krebber, F. J. Fleitz, J. Brands, P. N. Devine, G. W. Huisman, G. J. Hughes, *Science* **2010**, *329*, 305–309.

[5] a) D. Yi, T. Bayer, C. P. S. Badenhorst, S. Wu, M. Dörr, M. Höhne, U. T. Bornscheuer, *Chem. Soc. Rev.* **2021**, *50*, 8003–8049; b) U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, K. Robins, *Nature* **2012**, *485*, 185–194; c) D. C. Miller, S. V. Athavale, F. H. Arnold, *Nat. Synth.* **2022**, *1*, 18–23.

[6] a) E. Y. Bezsudnova, V. O. Popov, K. M. Boyko, *Appl. Microbiol. Biotechnol.* **2020**, *104*, 2343–2357; b) S. Kwon, H. H. Park, *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1031–1039.

[7] a) C. Xiang, Y.-F. Ao, M. Höhne, U. T. Bornscheuer, *Int. J. Mol. Sci.* **2022**, *23*, 15347; b) H. Jeon, A. D. Pagar, H. Kang, P. Giri, S. P. Nadarajan, S. Sarak, T. P. Khobragade, S. Lim, M. D. Patil, S.-G. Lee, H. Yun, *ACS Catal.* **2022**, *12*, 13207–13214; c) S. J. Novick, N. Dellas, R. Garcia, C. Ching, A. Bautista, D. Homan, O. Alvizo, D. Entwistle, F. Kleinbeck, T. Schlama, T. Ruch, *ACS Catal.* **2021**, *11*, 3762–3770; d) Y. Wang, J. Feng, W. Dong, X. Chen, P. Yao, Q. Wu, D. Zhu, *ChemCatChem* **2021**, *13*, 3396–3400; e) M. Voss, D. Das, M. Genz, A. Kumar, N. Kulkarni, J. Kustosz, P. Kumar, U. T. Bornscheuer, M. Höhne, *ACS Catal.* **2018**, *8*, 11524–11533.

[8] a) L. Yang, K. Zhang, M. Xu, Y. Xie, X. Meng, H. Wang, D. Wei, *Angew. Chem. Int. Ed.* **2023**, *62*, e20221255; b) L. Cui, A. Cui, Q. Li, L. Yang, H. Liu, W. Shao, Y. Feng, *ACS Catal.* **2022**, *12*, 13703–13714; c) Q. Meng, C. Ramirez-Palacios, N. Capra, M. E. Hooghwinkel, S. Thallmair, H. J. Rozeboom, A.-M. W. H. Thunnissen, H. J. Wijma, S. J. Marrink, D. B. Janssen, *ACS Catal.* **2021**, *11*, 10733–10747.

[9] S. Wu, C. Xiang, Y. Zhou, M. S. H. Khan, W. Liu, C. G. Feiler, R. Wei, G. Weber, M. Höhne, U. T. Bornscheuer, *Nat. Commun.* **2022**, *13*, 7458.

[10] a) M. Wittmund, F. Cadet, M. D. Davari, *ACS Catal.* **2022**, *12*, 14243–14263; b) L. G. Somermeyer, A. Fleiss, A. S. Mishin, N. G. Bozhanova, A. A. Igolkina, J. Meiler, M.-E. A. Pujol, E. V. Putintseva, K. S. Sarkisyan, F. A. Kondrashov, *eLife* **2022**, *11*, e75842; c) N. Sapoval, A. Aghazadeh, M. G. Nute, D. A. Antunes, A. Balaji, R. Baraniuk, C. J. Barberan, R. Dannenfelser, C. Dun, M. Edrisi, R. A. L. Elworth, B. Kille, A. Kyrillidis, L. Nakhleh, C. R. Wolfe, Z. Yan, V. Yao, T. J. Treangen, *Nat. Commun.* **2022**, *13*, 1728; d) B. L. Hie, K. K. Yang, *Curr. Opin. Struct. Biol.* **2022**, *72*, 145–152; e) S. L. Lovelock, R. Crawshaw, S. Basler, C. Levy, D. Baker, D. Hilvert, A. P. Green, *Nature* **2022**, *606*, 49–58; f) Y. Cui, J. Sun, B. Wu, *Trends Chem.* **2022**, *4*, 409–419; g) N. Singh, S. Malik, A. Gupta, K. R. Srivastava, *Emerging Top. Life Sci.* **2021**, *5*, 113–125; h) S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, *Nat. Methods* **2021**, *18*, 389–396; i) Y. Xu, D. Verma, R. P. Sheridan, A. Liaw, J. Ma, N. M. Marshall, J. Mclntosh, E. C. Sherer, V. Svetnik, J. M. Johnston, *J. Chem. Inf. Model.* **2020**, *60*, 2773–2790; j) M. J. Volk, I. Lourentzou, S. Mishra, L. T. Vo, C. Zhai, H. Zhao, *ACS Synth. Biol.* **2020**, *9*, 1514–1533; k) E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, *Nat. Methods* **2019**, *16*, 1315–1322; l) K. K. Yang, Z. Wu, F. H. Arnold, *Nat. Methods* **2019**, *16*, 687–694.

[11] F. Cadet, N. Fontaine, G. Li, J. Sanchis, M. N. F. Chong, R. Pandjaitan, I. Vetrivel, B. Offmann, M. T. Reetz, *Sci. Rep.* **2018**, *8*, 16757.

[12] Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8852–8858.

[13] E. J. Ma, E. Siirola, C. Moore, A. Kummer, M. Stoeckli, M. Faller, C. Bouquet, F. Eggimann, M. Ligibel, D. Huynh, G. Cutler, L. Siegrist, R. A. Lewis, A.-C. Acker, E. Freund, E. Koch, M. Vogel, H. Schlingensiepen, E. J. Oakeley, R. Snajdrova, *ACS Catal.* **2021**, *11*, 12433–12445.

[14] J. C. Greenhalgh, S. A. Fahlberg, B. F. Pfleger, P. A. Romero, *Nat. Commun.* **2021**, *12*, 5825.

[15] a) S. P. Kelly, V. V. Shende, A. R. Flynn, Q. Dan, Y. Ye, J. L. Smith, S. Tsukamoto, M. S. Sigman, D. H. Sherman, *J. Am. Chem. Soc.* **2022**, *144*, 19326–19336; b) F. Li, L. Yuan, H. Lu, G. Li, Y. Chen, M. K. M. Engqvist, E. J. Kerkhoven, J. Nielsen, *Nat. Catal.* **2022**, *5*, 662–672; c) Y. Saito, M. Oikawa, T. Sato, H. Nakazawa, T. Ito, T. Kameda, K. Tsuda, M. Umetsu, *ACS Catal.* **2021**, *11*, 14615–14624; d) S. L. Robinson, M. D. Smith, J. E. Richman, K. G. Aukema, L. P. Wackett, *Synth. Biol.* **2020**, *5*, ysaa004; e) B. M. Bonk, J. W. Weis, B. Tidor, *J. Am. Chem. Soc.* **2019**, *141*, 4108–4118; f) M. Yang, C. Fehl, K. V. Lees, E.-K. Lim, W. A. Offen, G. J. Davies, D. J. Bowles, M. G. Davidson, S. J. Roberts, B. G. Davis, *Nat. Chem. Biol.* **2018**, *14*, 1109–1117; g) R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon, G. W. Huisman, *Nat. Biotechnol.* **2007**, *25*, 338–344.

[16] L. Jia, T. Sun, Y. Wang, Y. Shen, *BioMed Res. Int.* **2021**, *2021*, 2593748.

[17] S. Schätzle, M. Höhne, E. Redestad, K. Robins, U. T. Bornscheuer, *Anal. Chem.* **2009**, *81*, 8244–8248.

[18] a) M. Kollipara, P. Matzel, U. T. Bornscheuer, M. Höhne, *Chem. Ing. Tech.* **2022**, *94*, 1836–1844; b) F. Steffen-Munsberg, C. Vickers, A. Thontowi, S. Schätzle, T. Meinhardt, M. S. Humble, H. Land, P. Berglund, U. T. Bornscheuer, M. Höhne, *ChemCatChem* **2013**, *5*, 154–157.

[19] a) S. Calvelage, M. Dörr, M. Höhne, U. T. Bornscheuer, *Adv. Synth. Catal.* **2017**, *359*, 4235–4243; b) M. Genz, O. Melse, S. Schmidt, C. Vickers, M. Dorr, T. van den Bergh, H.-J. Joosten, U. T. Bornscheuer, *ChemCatChem* **2016**, *8*, 3199–3202; c) I. V. Pavlidis, M. S. Weiβ, M. Genz, P. Spurr, S. P. Hanlon, B. Wirz, H. Iding, U. T. Bornscheuer, *Nat. Chem.* **2016**, *8*, 1076–1082.

[20] a) Y.-F. Ao, H.-J. Hu, C.-X. Zhao, P. Chen, T. Huang, H. Chen, Q.-Q. Wang, D.-X. Wang, M.-X. Wang, *ACS Catal.* **2021**, *11*, 6900–6907; b) H.-J. Hu, Q.-Q. Wang, D.-X. Wang, Y.-F. Ao, *Adv. Synth. Catal.* **2021**, *363*, 4538–4543.
[21] H. Lim, H.-N. Jeon, S. Lim, Y. Jang, T. Kim, H. Cho, J.-G. Pan, K. T. No, *Comput. Struct. Biotechnol. J.* **2022**, *20*, 788–798.
[22] R. Feehan, D. Montezano, J. S. G. Slusky, *Protein Eng. Des. Sel.* **2021**, *34*, gzab019.

Manuscript received: February 2, 2023
Accepted manuscript online: April 6, 2023
Version of record online: May 3, 2023