Supporting Information

Machine learning-guided protein engineering to improve the catalytic activity of transaminases under neutral pH conditions

Si-Han Liu,[a] Ligai Bai,*[a] Xu-Dong Wang,[b,c] Qi-Qiang Wang,[b,c] De-Xian Wang,[b,c] Uwe T. Bornscheuer[d] and Yu-Fei Ao*[b,c,d]

[a]*Key Laboratory of Medicinal Chemistry and Molecular Diagnosis of Ministry of Education, Key Laboratory of Pharmaceutical Quality Control of Hebei Province, College of Pharmaceutical Sciences, Hebei University, Wusi Dong Road No.180, Baoding, 071002, China.*

[b]*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*

[c]*University of Chinese Academy of Sciences, Yuquan Road 19(A), Beijing, 100049, China*

[d]*Department of Biotechnology and Enzyme Catalysis, Institute of Biochemistry, University of Greifswald, Greifswald, 17487, Germany*

Table of Contents

1. Docking experiments

The structure of the wild-type transaminase 3FCR was downloaded from the PDB database (PDB code: 3FCR). The geometrical optimizations of the quinonoid of substrate (*S*)-1-phenylethanamine were carried out with the *ChemBio3D Ultra 13.0* suite program using the MM2 basis set. Dockings were performed with the *AutoDock 4.2.6* suite.[S1] All water molecules were deleted in the docking simulations, and the grid box was centered on the carbon atom of the external aldimine of the cofactor PLP. Other parameters were kept at the default settings. The resulting ligand and original protein data files were used to generate Figure 1 using the *PyMOL 2.3.0* program.[S2]

2. Site-directed mutagenesis

All variants were prepared using the Q5® site-directed mutagenesis kit from New England BioLabs. Degenerate primers were designed non-overlapping by using the standard setting of the NEBaseChanger. For the PCR, 0.25 ng μL^{-1} template plasmid (carrying the 3FCR, 3HMU, or the 6SNU gene, respectively), 0.5 µM forward and reverse primers, Q5® hot start high-fidelity 2X master mix were used. The PCR was performed as follows: (i) 98 °C, 30 s; (ii) 30 cycles: 98 °C, 10 s; 50-72 °C, 30 s; 72 °C, 0.5 min/kbp; (iii) 72 °C, 2 min. The resulting PCR product was directly treated with the kinase, ligase & *Dpn*I (KLD enzyme mix) (100 µL mL^{-1}; NEB) at room temperature for 30 minutes and then used for the transformation of chemically competent *E. coli* TOP10 cells. After confirming the introduced mutation(s) by single colonies sequence detection, the plasmids were used for the transformation into chemical competent *E. coli* BL21(DE3) cells by the heat shock method. Primers used in this work are listed in Table S1.

Table S1. List of primers.

Primer	Sequence (5' to 3')
3FCR-S19I Fw	GTGATAACTTCTTCCACCCGATTACGCACCTGGCGC
3FCR-S19I Rv	CATGTTGCGCCAGGTGCGTAATCGGGTGGAAG
3FCR-S19H Fw	CGTGATAACTTCTTCCACCCGCATACGCACCTGGCGC
3FCR-S19H Rv	TTGTTGCGCCAGGTGCGTATGCGGGTGGAAGAAG
3FCR-L58M/Y59W Fw	GATGCTTTCGCGGGCATGTGGTGCGTTAAT
3FCR-L58M/Y59W Rv	CGACATTAACGCACCACATGCCCGCGAAAGCAT
3FCR-L58I/Y59W Fw	TGCTGGATGCTTTCGCGGGCATTTGGTGCGTTAATG
3FCR-L58I/Y59W Rv	TAGCCGACATTAACGCACCAAATGCCCGCGAAAGCAT
3FCR-L58K/Y59W Fw	CTGGATGCTTTCGCGGGCAAATGGTGCGTTAATG
3FCR-L58K/Y59W Rv	GTAGCCGACATTAACGCACCATTTGCCCGCGAAAG
3FCR-Y59L Fw	GGATGCTTTCGCGGGCCTGCTGTGCGTTAATGTC
3FCR-Y59L Rv	GTAGCCGACATTAACGCACAGCAGGCCCGCGAAAG
3FCR-Y59C Fw	GGATGCTTTCGCGGGCCTGTGTTGCGTTAATGTC
3FCR-Y59C Rv	GTAGCCGACATTAACGCAACACAGGCCCGC
3FCR-Y59F Fw	GATGCTTTCGCGGGCCTGTTTTGCGTTAATGTCGG
3FCR-Y59F Rv	GTAGCCGACATTAACGCAAAACAGGCCCGCGAAAG
3FCR-W59Y Fw	GGATGCTTTCGCGGGCCTGTATTGCGTTAATGTCG
3FCR-W59Y Rv	GTAGCCGACATTAACGCAATACAGGCCCGCGAAAG
3FCR-Y152N Fw	CAGTCGTTGGCGCGGTAATCATGGCAGTGGTC
3FCR-Y152N Rv	GTAACCAGACCACTGCCATGATTACCGCGCCAACGAC
3FCR-Y152F Fw	TATCAGTCGTTGGCGCGGTTTTCATGGCAGTG
3FCR-Y152F Rv	GTAACCAGACCACTGCCATGAAAACCGCGCC
3FCR-Y152H Fw	TTATCAGTCGTTGGCGCGGTCATCATGGCAGTGGTCT
3FCR-Y152H Rv	GTAACCAGACCACTGCCATGATGACCGCGCCAACGAC
3FCR-L165K Fw	CCGGCTCCCTGACGGGTAAAGAACTGTTTC
3FCR-L165K Rv	GAATTTTTTATGAAACAGTTCTTTACCCGTCAGGGAGCCG
3FCR-L165E Fw	CCGGCTCCCTGACGGGTGAAGAACTGTTTC
3FCR-L165E Rv	GAATTTTTTATGAAACAGTTCTTCACCCGTCAGGGAGCCG

3FCR-L167H Fw	GCTCCCTGACGGGTCTGGAACATTTTCATAAAAAATT
3FCR-L167H Rv	GCAGATCGAATTTTTTATGAAAATGTTCCAGACCCGTCAG
3FCR-L167F Fw	GGCTCCCTGACGGGTCTGGAATTCTTTCATAAAAAATTC
3FCR-L167F Rv	GCAGATCGAATTTTTTATGAAAGAATTCCAGACCCGTCAG
3FCR-L167Y Fw	GGCTCCCTGACGGGTCTGGAATATTTTCATAAAAAATTCG
3FCR-L167Y Rv	GGCAGATCGAATTTTTTATGAAAATATTCCAGACCCGTCAGG
3FCR-F168H Fw	CCTGACGGGTCTGGAACTGCATCATAAAAAATTC
3FCR-F168H Rv	GGCAGATCGAATTTTTTATGATGCAGTTCCAGACC
3FCR-F168Y Fw	CCTGACGGGTCTGGAACTGTATCATAAAAAATTCG
3FCR-F168Y Rv	GGCAGATCGAATTTTTTATGATACAGTTCCAGACCCG
3FCR-F168C Fw	CCTGACGGGTCTGGAACTGTGTCATAAAAAATTCG
3FCR-F168C Rv	GCAGATCGAATTTTTTATGACACAGTTCCAGACCCG
3FCR-F168M Fw	CCCTGACGGGTCTGGAACTGATGCATAAAAAATTCG
3FCR-F168M Rv	GGCAGATCGAATTTTTTATGCATCAGTTCCAGACCCG
3FCR-T231G Fw	CGAACCGATTCTGGGTGGTGGCGGTATTGTGC
3FCR-T231G Rv	CGGCACAATACCGCCACCACCCAGAATCGG
3FCR-V261T Fw	GCTGGTTGCGGACGAAACCGTTACCGGCTTT
3FCR-V261T Rv	CGACCAAAGCCGGTAACGGTTTCGTCCGCAAC
3FCR-V261A Fw	TGGTTGCGGACGAAGCAGTTACCGGCTTTG
3FCR-V261A Rv	GACCAAAGCCGGTAACTGCTTCGTCCGCAAC
3FCR-R420K Fw	ACAAGATAAAATTATCGCGAAAGCCATGCCGCAGGG
3FCR-R420K Rv	TCGCCCTGCGGCATGGCTTTCGCGATAATTTT

3. Gene expression and purification of the enzyme variants

For the protein biosynthesis of 3FCR and its variants, transformed *E. coli* BL21(DE3) cells were incubated overnight at 37 °C in a 5 mL LB-medium (Lysogeny Broth) preculture with ampicillin (100 µg mL^{-1}). 1 mL of the preculture was used for the inoculation of 100 mL TB-medium (supplemented with the corresponding antibiotic) and incubated at 37 °C, 180 rpm. The expression for enzyme production was induced at an optical density of approx. 0.6 (measured at 600 nm) with 0.2% L-rhamnose. The cells were incubated at 20 °C for 20 h

and then harvested by centrifugation (20 min, 4000 g, 4 °C).

For purification, the harvested cells were resuspended in 50 mM HEPES buffer pH 7.5 containing 0.1 mM PLP, 0.3 M NaCl, 0.01 M imidazole and then lysed via ultrasound (50% pulse, 50% power, 2x5 min; Sonoplus HD2070, Bandelin Electronic GmbH). The lysate was clarified by centrifugation (0.5 h, 10000 g, 4 °C) and purified by immobilized metal affinity chromatography with the following buffers: washing buffer (50 mM HEPES buffer pH 7.5 containing 0.1 mM PLP, 0.3 M NaCl, 0.02 M imidazole), elution buffer (50 mM HEPES buffer pH 7.5 containing 0.1 mM PLP, 0.3 M NaCl, 0.3 M imidazole). The proteins were desalted in 50 mM HEPES buffer pH 7.5, 0.1 mM PLP using PD-10 desalting columns (GE Healthcare). The purified and desalted proteins were stored at -20 °C in 30% glycerol.

4. Measurement of specific activity of the ATAs

All solvents and chemicals were obtained from standard commercial sources and used without further purification. The activities of the purified transaminase variants were studied using the conversion of substrates resulting in the formation of ketone products, which were quantified spectrophotometrically at 245-302 nm over time using the Infinite® 200 PRO (TECAN) plate reader in UV-transparent microtiter plates (UV-Star, Greiner Bio-One GmbH). The detection wavelengths used were determined by the optimal UV absorption wavelength of the substrates and products, and the standard curves were measured accordingly.[S3] The assay was performed with 2.5 mM amine substrates as amine donor and 2.5 mM pyruvate as amine acceptor in 1.25-2.5% DMSO, 100 mM phosphate buffer (K_2HPO_4 - KH_2PO_4) pH 9.0 - 6.5 at 30 °C. One unit (U) activity was defined as the formation of 1 µmol corresponding aromatic ketone (for example: acetophenone) per minute. All measurements were performed in triplicates.

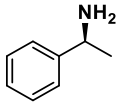
5. The descriptors

Each amino acid descriptors comprises three elements (Table S2): element A represents the volume of an amino acid; element B represents the hydrophobicity index of an amino acid: element C represents the isoelectric point of an amino acid.

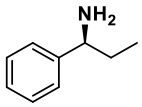
Table S2: Descriptors of amino acids

amino acid	element A	element B	element C	amino acid	element A	element B	element C
A	67	1.8	6.00	C	86	2.5	5.07
P	90	-1.6	6.30	T	93	-0.7	5.60
V	105	4.2	5.96	N	96	-3.5	5.41
I	124	4.5	6.02	Q	114	-3.5	5.65
L	124	3.8	5.98	H	118	-3.2	7.59
M	124	1.9	5.74	Y	141	-1.3	5.66
F	135	2.8	5.48	K	135	-3.9	9.74
W	163	-0.9	5.89	R	148	-4.5	10.76
G	48	-0.4	5.97	D	91	-3.5	2.77
S	73	-0.8	5.68	E	109	-3.5	3.22

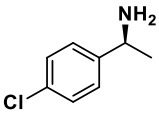
Table S3: Descriptors of the substrates



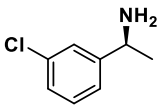
1a



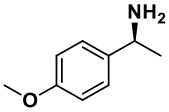
1b



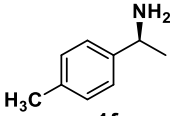
1c



1d



1e



1f

substrate	Element A	Element B	Element C
1a	0	0	0
1b	0	0	1
1c	1.2	0	0
1d	0	1.2	0
1e	1.9	0	0
1f	1	0	0

Each substrate comprises three elements (Table S3): elements A, B and C represent steric properties of the substituent group. Among them, element A represents the atom (exception: hydrogen) number of the para-position substitute of the phenyl group, element B represents the atom number of the meta-position substitute of the phenyl group; element C represents the atom number of the methyl group. In order to differentiate the carbon-atom and

heteroatom substitutes in the phenyl ring, a bond length corrective was used. For example, the carbon-chlorine bond length is 1.2 times larger than the carbon-carbon single bond.

6. Machine learning

Machine learning code was modified according our previous research.[S5] The dataset was built according to the specific activity data for a given reaction and the designed descriptors, and it was used to search the best hyperparameters. In this article, only variance filtering has been used for feature selection. It removes features whose variance does not meet a certain threshold. Here, we removed the features with zero variance, that is, the features whose values are the same in all samples. The hyperparameters of four regression algorithms have been: random forest (RF), support vector machine (SVM), kernel ridge regression (KRR), and gradient boosting regression tree (GBRT). These were tuned with a ten-fold cross-validated grid-search on the training set. Subsequently, these models were retrained on the training set with the above optimized hyperparameters and evaluated on the test set (Table S4). The GBRT was selected as the best algorithm.

The quality of ML predictors were evaluated by the coefficient of determination (R^2) and root mean squared error (RMSE). They are defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where y_i and $\hat{y}_i$ are the true and predicted values of the *i*th sample, respectively. All machine learning works were performed using Scikit-Learn[S4] package in this work.

Table S4: Details of the GBRT workflow approach with increased dataset

Model	Updated cycle	Database			Hyperparameters	Test set	
		N_{sample}	$N_{variants}$	$N_{substrates}$		R^2	RMSE
GBRT	1	810	27 3FCR variants	5	n_estimators = 450, learning_rate = 0.10, max_depth = 6, min_samples_split = 2, min_samples_leaf = 4, subsample = 0.60	0.952	0.485
	2	930	31 3FCR variants	5	n_estimators = 450, learning_rate = 0.10, max_depth = 6, min_samples_split = 2, min_samples_leaf = 5, subsample = 0.6	0.949	0.481

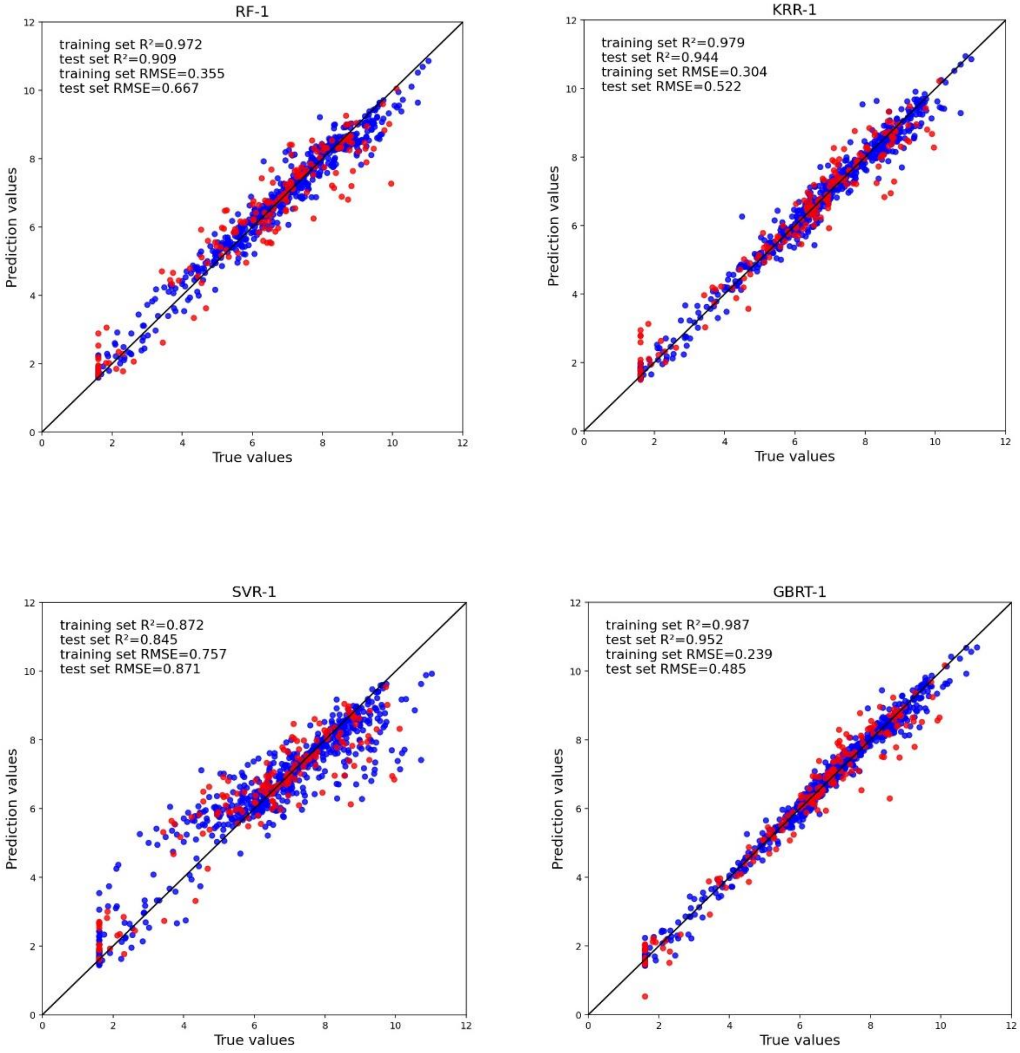


Fig. S1. Results of the screening using the ML regression algorithms (RF, KRR, SVR and GBRT) using the 810 dataset. The true value of each point corresponds to the natural logarithm of the mean specific activity values of three independent experiments. Blue and red dots in the figures represent data from training and test set, respectively.

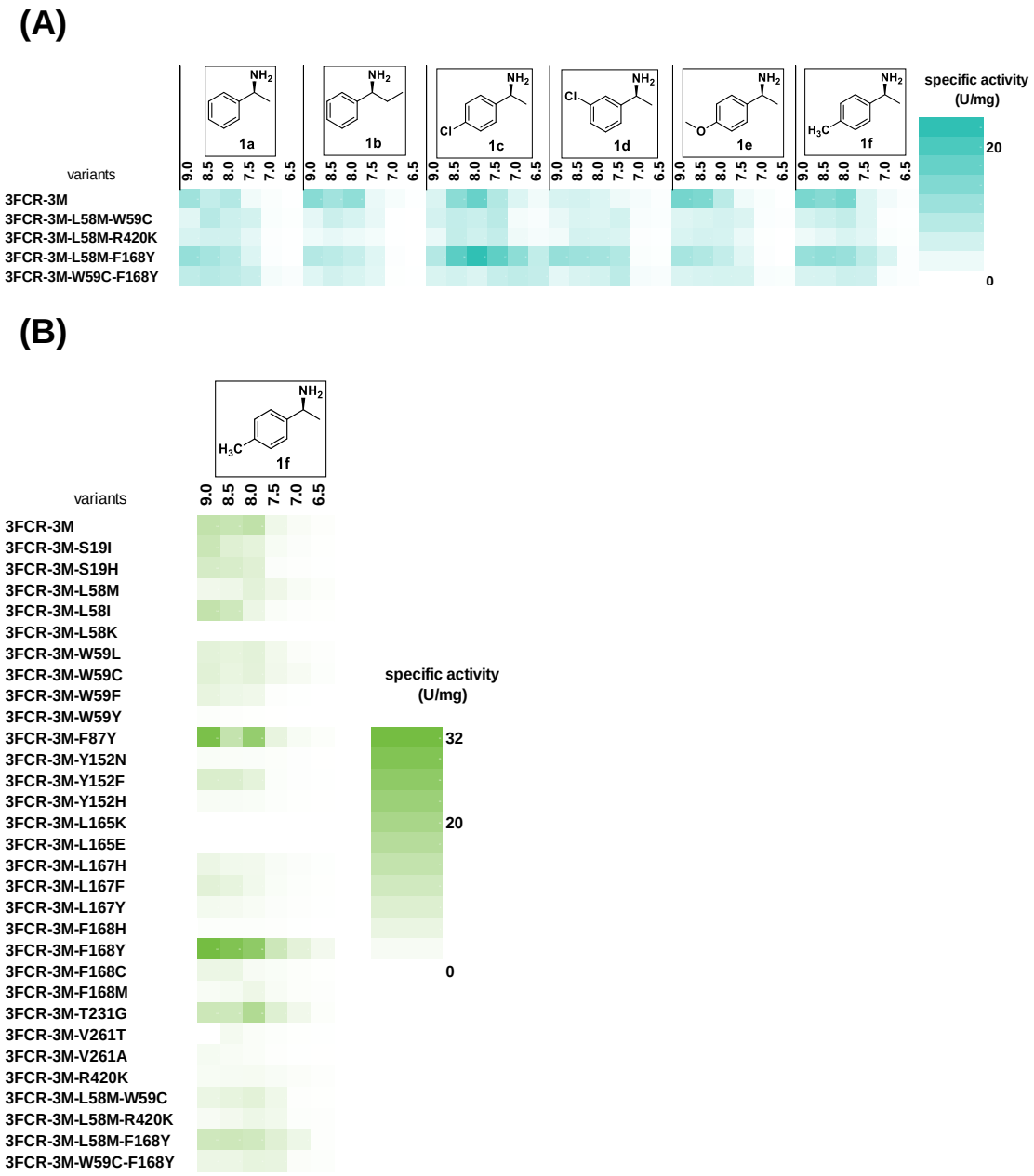


Fig. S2. Experimental validation of ML model's prediction. Specific activities measured using the predicted 4 variants of 3FCR (**A**), or toward new substrate **1f** (**B**). All measurements were performed in triplicates and the mean values are indicated by a color gradient. The detailed data for specific activity are given in the SI file named "SI-excel".

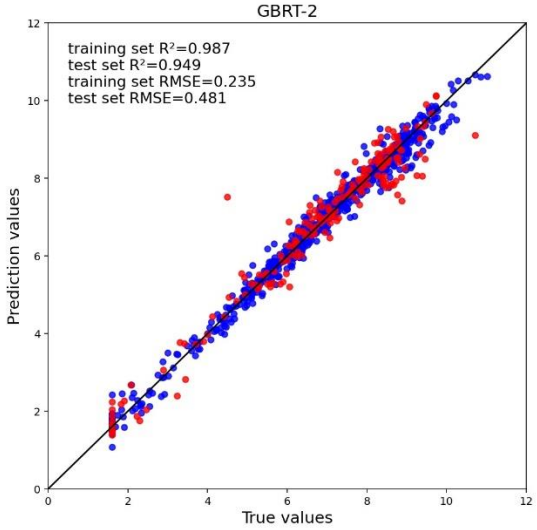


Fig. S3: Performances of the updated GBRT model. The dataset incorporates new experimental data for the four variants, bringing the total number of data points to 930. The true value of each point corresponds to the natural logarithm of the mean specific activity values of three independent experiments. Blue and red dots in the figures represent data from training and test set, respectively.

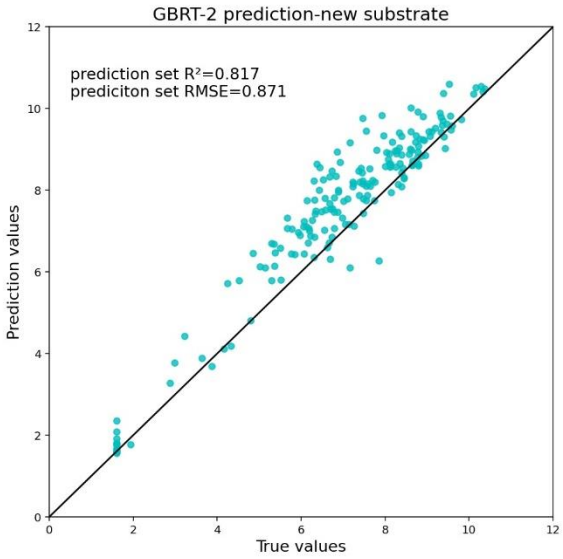


Fig. S4: Regression performances of the updated GBRT predictor toward substrate **1f**. The true values for each point correspond to the natural logarithm of the mean specific activity values of three independent experiments.

Table S5. Predicted and measured specific activities toward substrates **1f** at pH 7.5

Entry	Variants	Specific activity[a] [U mg^{-1}]		
		Predict results	Measured results	error
1	3FCR-3M-F168Y	32.103	11.936	0.012
2	3FCR-3M-F87Y	22.432	5.514	0.006
3	3FCR-3M-L58M-F168Y	18.210	7.320	0.007
4	3FCR-3M-L58I	9.607	1.286	0.001
5	3FCR-3M-W59C-F168Y	8.176	5.397	0.005
6	3FCR-3M	7.890	3.778	0.004
7	3FCR-3M-L167Y	7.608	0.957	0.001
8	3FCR-3M-L58M	7.308	3.908	0.004
9	3FCR-3M-T231G	7.034	7.728	0.008
10	3FCR-3M-W59L	6.392	3.197	0.003
11	3FCR-3M-W59C	5.365	3.406	0.003
12	3FCR-3M-L58M-W59C	5.288	3.839	0.004
13	3FCR-3M-L58M-R420K	5.263	3.339	0.003
14	3FCR-3M-L167H	5.147	1.664	0.002
15	3FCR-3M-Y152F	4.226	0.930	0.001
16	3FCR-3M-W59F	3.749	0.546	0.001
17	3FCR-3M-L167F	3.516	1.408	0.001
18	3FCR-3M-S19I	3.309	1.890	0.002
19	3FCR-3M-F168C	3.277	1.376	0.001
20	3FCR-3M-S19H	3.007	0.983	0.001
21	3FCR-3M-R420K	2.651	1.599	0.002
22	3FCR-3M-F168M	2.435	1.769	0.002
23	3FCR-3M-Y152H	2.173	0.796	0.001
24	3FCR-3M-V261A	1.796	0.567	0.001
25	3FCR-3M-Y152N	1.176	0.883	0.001
26	3FCR-3M-F168H	0.973	0.504	0.001
27	3FCR-3M-W59Y	0.807	0.200	0.000
28	3FCR-3M-V261T	0.552	0.800	0.001
29	3FCR-3M-L58K	0.049	0.038	0.000
30	3FCR-3M-L165K	0.007	0.005	0.000
31	3FCR-3M-L165E	0.005	0.005	0.000

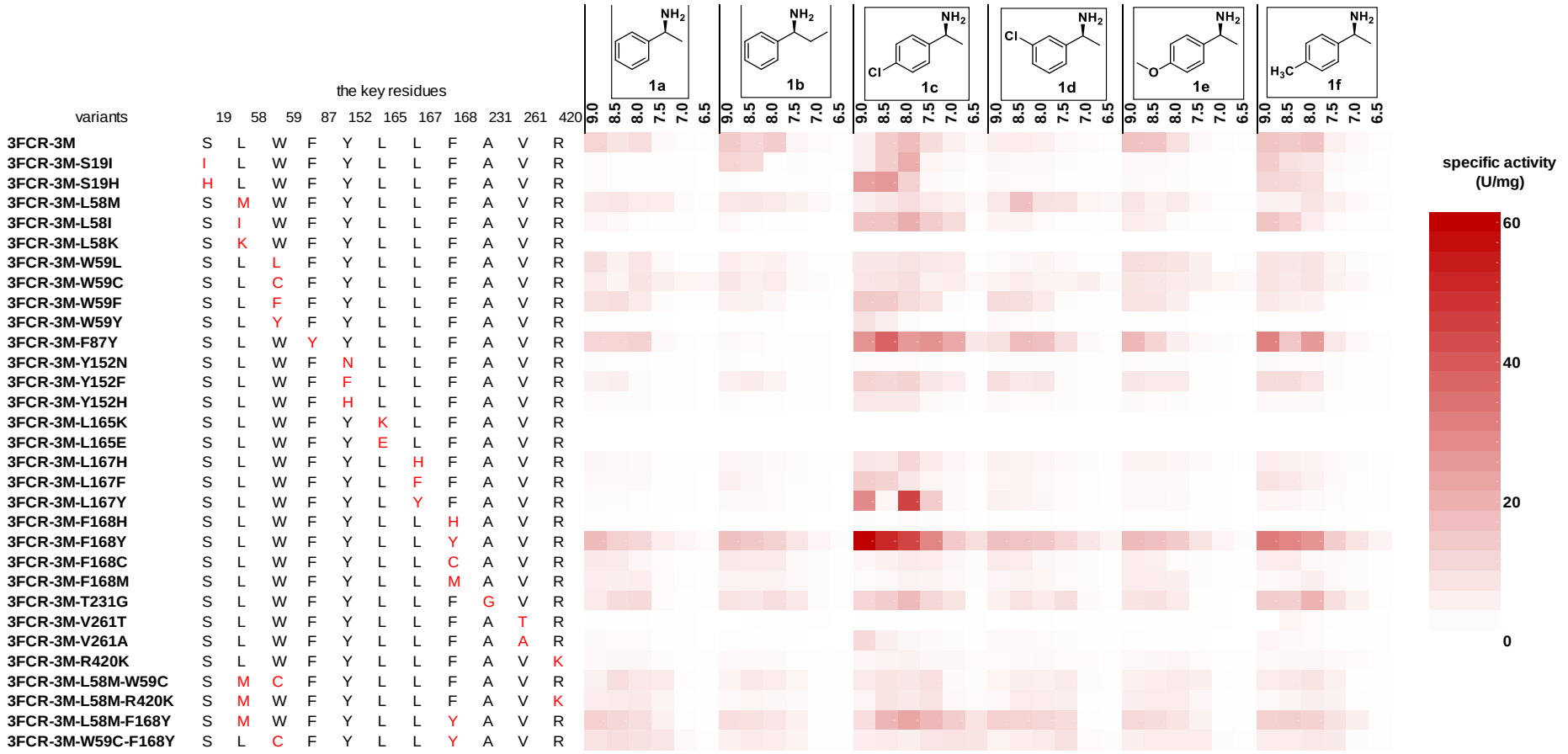


Fig. S5. Specific activities of 3FCR variants toward the substrate **1a**-**f**. All measurements were performed in triplicates and the mean values are indicated by a color gradient. Detailed data are given in the SI file named ”SI-excel”.

7. Protein and DNA sequences of the enzymes studied

>Protein sequence of 3FCR-3M:

MLKNDQLDQWDRDNFFHPSTHLAQHARGESANRVIKTASGVFIEDRDGTKLLDAFA
GLWCVNVGYGRQEIAEAIADQARELAYYHSFVGHGTEASITLAKMILDRAPKNMSK
VYFGLGGSDANETNVKLIWYYNNILGRPEKKKIISRWRGYHGSGLVTGSLTGLELFH
KKFDLPVEQVIHTEAPYYFRREDLNQTEEQFVAHCVAELEALIEREGADTIAAFIGEPI
LGAGGIVPPPAGYWEAIQTVLNKHDILLVADEVVTGFGRLGTMFGSDHYGLEPDIITI
AKGLTSAYAPLSGSIVSDKVWKVLEQGTDENGPIGHGWTYSAHPIGAAAGVANLKL
LDELNLVSNAGEVGAYLNATMAEALSQHANVGDVRGEGLLCAVEFVKDRDSRTFF
DAADKIGPQISAKLLEQDKIIARAMPQGDILGFAPPFCLTRAEADQVVEGTLRAVKAV
LGSLIGSDGGSGGGSTSRDHMVLHEYVNAAGITLIGSDGGSGGGSTSHHHHHH

>DNA sequence of 3FCR-3M:

ATGCTGAAAAACGACCAACTGGACCAATGGGACCGTGATAACTTCTTCCACCCG
TCAACGCACCTGGCGCAACATGCCCGTGGCGAATCAGCTAACCGTGTGATCAA
AACCGCGTCGGGCGTTTTTATTGAAGATCGCGACGGTACGAAACTGCTGGATG
CTTTCGCGGGCCTGTGGTGCGTTAATGTCGGCTACGGTCGTCAGGAAATTGCC
GAAGCAATCGCTGATCAAGCGCGCGAACTGGCCTATTACCATAGCTTCGTGGG
CCACGGTACCGAAGCTTCTATCACGCTGGCGAAAATGATTCTGGATCGTGCCCC
GAAAAACATGAGTAAAGTTTACTTTGGTCTGGGCGGTTCCGACGCAAACGAAAC
CAATGTCAAACTGATCTGGTATTACAACAATATTCTGGGCCGCCCGGAGAAAAA
GAAAATTATCAGTCGTTGGCGCGGTTATCATGGCAGTGGTCTGGTTACCGGCTC
CCTGACGGGTCTGGAACTGTTTCATAAAAAATTCGATCTGCCGGTGGAACAGGT
TATTCACACCGAAGCCCCGTATTACTTTCGTCGCGAAGACCTGAACCAGACGGA
AGAACAATTCGTCGCACACTGTGTGGCTGAACTGGAAGCGCTGATCGAACGTG
AAGGCGCGGATACCATTGCGGCCTTCATCGGCGAACCGATTCTGGGTGCGGGC
GGTATTGTGCCGCCGCCGGCCGGTTATTGGGAAGCAATCCAGACCGTCCTGAA
TAAACATGATATTCTGCTGGTTGCGGACGAAGTGGTTACCGGCTTTGGTCGCCT
GGGCACGATGTTCGGTTCTGATCACTATGGCCTGGAACCGGACATTATCACCAT
CGCGAAAGGTCTGACGTCAGCGTACGCCCCGCTGAGCGGTTCTATTGTGTCGG
ATAAAGTCTGGAAAGTGCTGGAACAGGGCACCGACGAAAACGGTCCGATCGGC
CATGGTTGGACGTATAGCGCACACCCGATTGGTGCAGCTGCAGGTGTTGCAAA
TCTGAAACTGCTGGATGAACTGAACCTGGTTAGCAATGCCGGCGAAGTCGGTG
CCTACCTGAACGCAACCATGGCAGAAGCTCTGTCCCAACATGCTAATGTTGGCG
ATGTCCGTGGCGAAGGTCTGCTGTGCGCGGTGGAATTTGTTAAAGATCGTGACA
GCCGCACGTTTTTCGATGCCGCAGACAAAATCGGTCCGCAGATTTCTGCGAAAC
TGCTGGAACAAGATAAAATTATCGCGCGTGCCATGCCGCAGGGCGACATTCTG
GGTTTTGCCCCGCCGTTCTGTCTGACCCGCGCAGAAGCTGATCAAGTCGTGGA
AGGTACGCTGCGCGCTGTCAAAGCCGTTCTGGGTTCACATCACCATCACCACCA
CTAA

8. References:

S1. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785-2791.

S2. Schrodinger, LLC. The PyMOL molecular graphics system, version 2.3.0. New York, NY.

S3. S. Schätzle, M. Höhne, E. Redestad, K. Robins, U. T. Bornscheuer, Rapid and sensitive kinetic assay for characterization of ω–transaminases. *Anal. Chem.*, 2009, **81**, 8244-8248.

S4. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn Res.*, 2011, **12**, 2825-2830.

S5. Y.-F. Ao, S. Pei, C. Xiang, M. J. Menke, L. Shen, C. Sun, M. Dörr, S. Born, M. Höhne, U. T. Bornscheuer, Structure- and Data-Driven Protein Engineering of Transaminases for Improving Activity and Stereoselectivity, *Angew. Chem. Int. Ed.*, 2023, **62**, e202301660.