Rational Design of Norcoclaurine Synthase Using Molecular Simulation and Deep Learning: Supplementary Information

CONTENTS

CONTENTS

1. PLASMID DETAILS

1.1 DNA sequence of $\Delta 33TfNCS$

ATGCATCATCACCACCATCATTCTAGCGGCGTCGACTTGGGTACT**GAGAATCTGTATTTTCAGAGC**
ATGGGTATCATTAACCAAGTTAGCACGGTCACCAAAGTGATTCACCACGAACTG
GAAGTTGCGGCCAGCGCGGACGACATCTGGACGGTGTACTCGTGGCCGGGTCTGGCTA
AGCACTTGCCGGATCTGTTGCCGGGTGCGTTCGAGAAACTGGAAATCATCGGCGATGG
TGGCGTCGGTACCATTCTGGATATGACCTTCGTGCCGGGTGAATTTCCTCACGAGTACA
AAGAAAAGTTTATCCTGGTTGATAACGAACATCGTCTGAAAAAAGTTCAGATGATTGA
GGGCGGCTATCTGGACCTGGGCGTCACGTATTACATGGACACCATTCACGTCGTTCCGA
CCGGTAAAGACAGCTGCGTGATCAAGAGCAGCACCGAGTACCACGTGAAGCCAGAGT
TCGTGAAGATTGTAGAGCCGCTGATCACGACCGGTCCGCTGGCCGCGATGGCAGATGC
AATTAGCAAACTGGTTCTGGAGCACAAGTCCTGA

The TEV cleavage site is given in **bold**.

1.2 DNA sequence of codon-optimised $\Delta 33TfNCS$

ATGCACCACCACCACCACCATAGCAGCGGCGTTGATCTGGGCACC**GAAAACCTGTACTTCCAGAGC**
ATGGGCATCATCAACCAGGTTAGCACCGTTACCAAAGTTATCCACCACGAACTG
GAAGTTGCGGCGAGCGCGGATGATATCTGGACCGTTTACAGCTGGCCGGGCCTGGCGA
AACACCTGCCGGATCTGCTGCCGGGTGCGTTCGAAAAACTGGAAATCATCGGCGATGG
CGGTGTTGGCACCATCCTGGATATGACCTTCGTTCCGGGTGAATTCCCGCACGAATACA
AAGAAAAATTCATCCTGGTTGATAACGAACACCGTCTGAAAAAAGTTCAGATGATCGA
AGGTGGCTACCTGGATCTGGGCGTTACCTACTACATGGATACCATCCACGTTGTTCCGA
CCGGTAAAGATAGCTGCGTTATCAAAAGCAGCACCGAATACCACGTTAAACCGGAATT
CGTTAAAATCGTTGAACCGCTGATCACCACCGGCCCGCTGGCGGCGATGGCGGATGCG
ATCAGCAAACTGGTTCTGGAACACAAAAGCTAA

The TEV cleavage site is given in **bold**. The sequence was codon optimized for *E. coli* expression by Jiangsu KeyGEN BioTECH Corp, Ltd.

1.3 Amino sequence of $\Delta 33TfNCS$

MHHHHHHSSGVDLGT**ENLYFQS**MGIINQVSTVTKVIHHELEVAASADDIWTVYSWPGLAK
HLPDLLPGAFEKLEIIGDGGVGTILDMTFVPGEFPHEYKEKFILVDNEHRLKKVQMIEGGYLD
LGVTYYMDTIHVVPTGKDSCVIKSSTEYHVKPEFVKIVEPLITTGPLAAMADAISKLVLEHKS

The TEV cleavage site is given in **bold**.

2. SITE MUTATION

2.1 Cloning of Mutated DNA Fragments

The site-specific mutagenesis primers were synthesized by Sangon Biotech Company. Solutions of upstream and downstream primers were obtained (10pmole/μL).

Prepare the PCR reaction system according to **Table S1**. Mix well, and carry out the PCR according to the program detailed in **Table S2**.

Table S1. PCR Reaction System

Component	Volume (μL)
dd H2O	20
Primer F (10 pmole/μL)	2
Primer R (10 pmole/μL)	2
Plasmid WT	1
High-fidelity DNA Polymerase	25

Table S2. PCR Reaction Protocol

Step	Temperature (°C)	Time (sec)
1	98	30
	Step2-Step4 in 34 cycles	
2	98	10
3	Tm (Calculated)	5
4	72	30
5	72	60

2.2 Agarose Gel Electrophoresis for Nucleic Acids

Based on the required gel volume, weigh the agarose powder. Add 1×TAE solution to the corresponding volume, add nucleic acid dye, and heat until dissolved while mixing. Pour the liquid into the gel tray and insert the comb. Once solidified, remove the comb.

Load 16-17^{-}L of the PCR product into each well of the gel. Perform electrophoresis at a constant voltage of 120V for 30 minutes.Observe the electrophoresis bands under a gel imaging system.

2.3 Gel DNA Recovery

After electrophoresis, place the gel under UV light, and excise the gel piece containing the DNA bands, placing them into 1.5mL eppendorf tubes. Use the Gel DNA Recovery Kit (FastPure Gel DNA Extraction Mini Kit, Vazyme) to recover the linear mutagenized DNA.

2.4 Determination of Plasmid Concentration (Linear)

Determine the concentration of the recovered DNA product using microspectrophotometer (Unano-1000,Hangzhou UMI instrument).

2.5 Linear Plasmid DNA Recombination

Prepare the PCR reaction system according to **Table S3**. Mix well and place in the PCR machine at 37°C for 30 minutes.

Table S3. Linear DNA Recombinant System

Component	Volume (μL)
Exnase2	2
5*CE2	4
Linear DNA	2
dd H2O	12

2.6 Plasmid Transformation

Add 10 μ L of the obtained plasmid to 100 μL of competent DH5α cells. Incubate on ice for 30 minutes, followed by a heat shock at 42°C for 60 seconds, and then place back on ice for 2-3 minutes. Add 600 μL of LB medium and shake at 37°C for 1 hour. Centrifuge at 5,000 r/min, resuspend in 100 μL of LB medium, and plate.

2.7 Picking Single Colonies

Pick single colonies from the plate and culture in LB medium. Send a sample of the bacterial culture to Sangon Biotech Company for Sanger sequencing.

2.8 Mutant Plasmid Extraction

Amplify the single clone *E. coli.* Use the Plasmid DNA Extraction Kit (FastPure EndoFree Plasmid Mini Kit-BOX 2, Vazyme) to extract the plasmid DNA and store at -80°C.

3. ENZYME EXPRESSION, LYSIS AND PURIFICATION

3.1 Enzyme Expression

Extract the plasmid from the transformed DH5α and clone the plasmid into BL21 (DE3) *E. coli.* Inoculate a single colony into 30 mL of LB medium and culture at 37°C for 8 hours. Inoculate into 300 mL of LB medium (1% v/v, with 0.1% v/v 1000×kanamycin) and culture overnight at 37°C for 2 hours. Add isopropyl -D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.1 mM and 1000×kanamycin at 0.1% v/v, and induce overnight at 20°C. Aliquot 45mL per tube and centrifuge (6000xg, 5 minutes) to separate *E. coli* cell pellets. Store at -20°C for further purification or lysis for enzymatic reactions.

3.2 Purification

To purify Δ33TfNCS and its mutants, the cell pellets were resuspended in lysis buffer (100 mM Hepes, 20 mM imidazole, 100 mM NaCl, pH 7.5, 15% v/v culture volume). Cells were lysed by sonication (300W, 35 minutes [3s ON, 3s OFF]), followed by centrifugation at 11,000rpm for 15 minutes at 4°C to collect the lysate. The 10mL His-trap HP column was pre-treated by discarding 20% ethanol(Nickel column storage solution) and washed twice with pure water, then equilibrated with lysis buffer. The lysate was then loaded onto the column. The column was washed with lysis buffer until no proteins were eluted. Subsequently, the column was subjected to step-gradient elution using a combination of lysis buffer and elution buffer (100 mM Hepes, 500 mM imidazole, 100 mM NaCl, pH 7.5) with increasing imidazole concentrations (8% elution buffer for 3 column volumes, 16/100 elution buffer for 3 column volumes followed by 100% elution buffer).

The fractions from different imidazole concentrations were analyzed using SDS-PAGE and visualized after staining. Fractions containing the pure protein were pooled and dialyzed against the buffer (20 mM Tris, 50 mM NaCl, pH 7.5). The buffer was changed every 6 hours, for a total of three changes. The protein solution obtained after dialysis was collected, its concentration was measured, and it was stored at -80°C[1].

4. AGAROSE GEL ELECTROPHORESIS

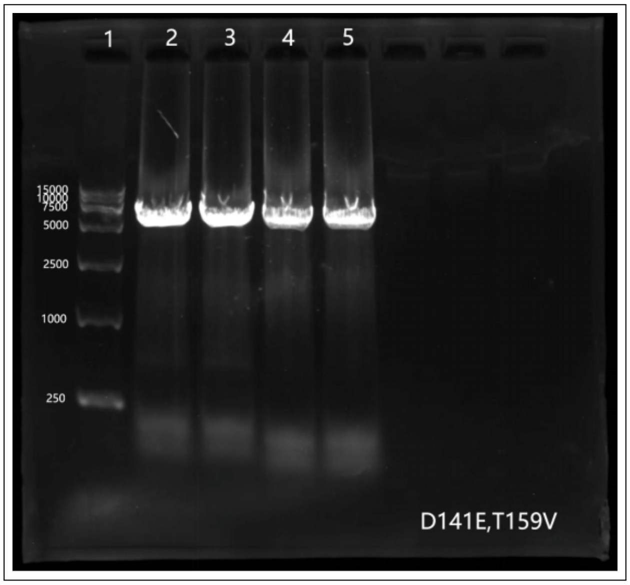


Fig. S1. Mutant plasmids of Δ33TfNCS in 1% Agarose gel. Lanes: 1, Benchmark™ Nucleic Acid Ladder masses given in bp. 2-3, mutant of D141E. 4-5, mutant of T159V.

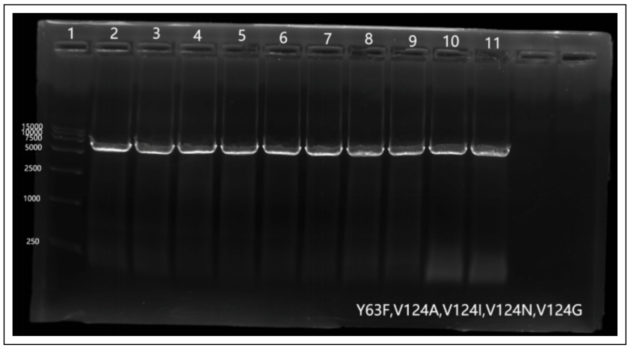


Fig. S2. Mutant plasmids of Δ33TfNCS in 1% Agarose gel. Lanes: 1, Benchmark™ Nucleic Acid Ladder masses given in bp. 2-4, mutant of Y108F.

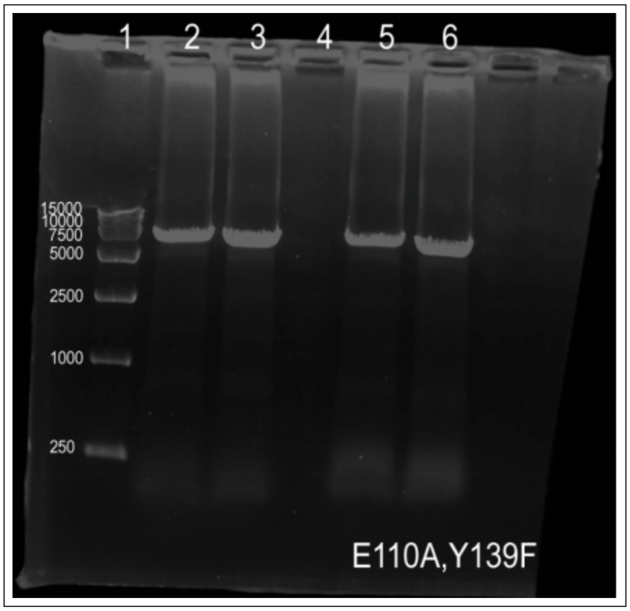


Fig. S3. Mutant plasmids of Δ33TfNCS in 1% Agarose gel. Lanes: 1, Benchmark™ Nucleic Acid Ladder masses given in bp. 2-3, mutant of E110A. 5-6, mutant of Y139F.

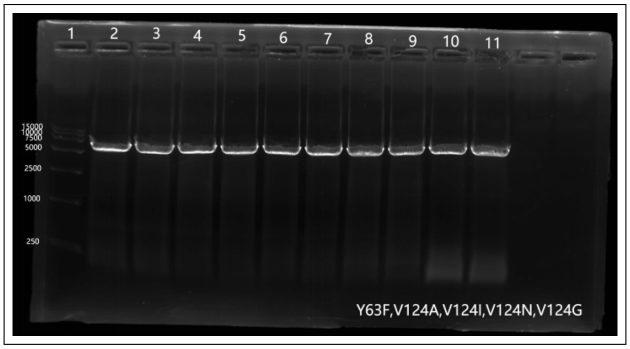


Fig. S4. Mutant plasmids of Δ33TfNCS in 1% Agarose gel. Lanes: 1, Benchmark™ Nucleic Acid Ladder masses given in bp. 2-3, mutant of Y63F. 4-5, mutant of V124A. 6-7, mutant of V124I. 8-9, mutant of V124N. 10-11, mutant of V124G.

5. SDS-PAGE ANALYSIS

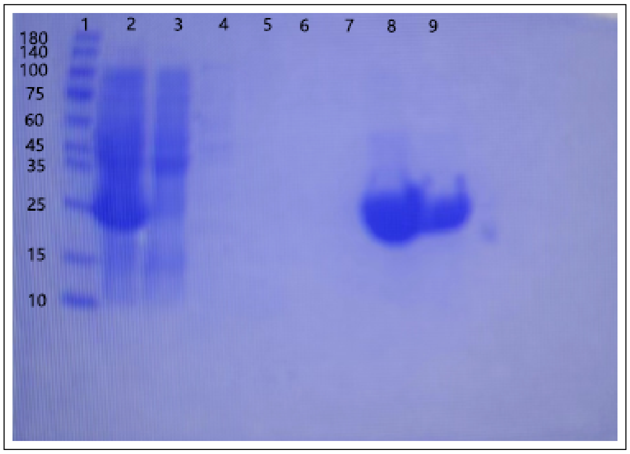


Fig. S5. His-trap purification of Δ33TfNCS. Lanes: 1, Benchmark™ Protein Ladder masses given in kDa. 2, clarified cell lysate. 3, flow through with lysis buffer. 4-5, wash with 3 CV 8% elution buffer. 6-7, wash with 3 CV 16% elution buffer. 8-9, wash with 100% elution buffer.

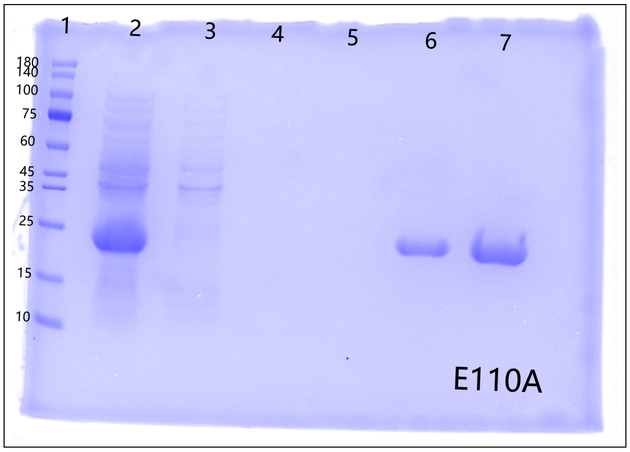


Fig. S6. His-trap purification of Δ33TfNCS-E110A. Lanes: 1, Benchmark™ Protein Ladder masses given in kDa. 2, clarified cell lysate. 3, flow through with lysis buffer. 4, wash with 3 CV 8% elution buffer. 5, wash with 3 CV 16% elution buffer. 6-7, wash with 100% elution buffer.

6. MOLECULAR SIMULATION AND MODELLING

6.1 Classical molecular dynamics

Molecular dynamics simulations were performed in GROMACS package (version 2023-gpu) on high performance computational center(BEIJING SUPER COLUD COMPUTING CENTER). The Amber14 force filed and TIP3P water model were employed for simulation. Protein was provided by PDBPDBID5N5O which had been demonstrated that simulation conducted by this structure can generate phenomena corresponding to the experiment. The structure of all the ligands were optimized by Gaussian16 with the M06-2X functional and def2-TZVP basis. Additionally, ligands were parametrized by GAFF model with the RESP2 charge method which has been fit well with GAFF model.

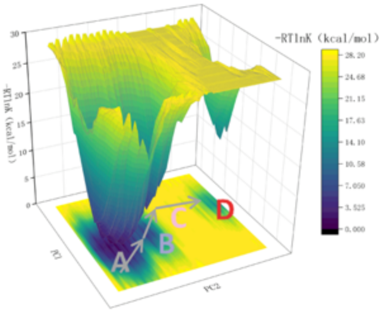
The entire workflow for classical molecular dynamics simulation: 1) Use Ledock to dock the ligand into catalytic pocket. 2) Utilize the cluster analysis to handle the structure generated by docking, then chose the most representative one as the initial structure. 3) Build gro files and positional restraints with gmx editconf (take the protein as the center, expand outward 1.2nm to get the whole box). 4) Use water, chloride and sodium ions to fill the spare space, keep the pH=7.0 and salt concentration is 0.15M (close to experiment). 5) Minimize the system by steepest descent method firstly then followed by conjugate gradient methods 6) Temperature coupling and pressure coupling were performed at the same time for quickly reaching equilibrium phase. 7) All data are collected from the 100ns simulation of equilibrium phase.

Minimization: Firstly use steepest descent method and Verlet integrator for 100000 steeps to reduce the maximum force less than 250 kJ*mol-1*nm-1. The protein and ligands were restrained by positional restraints but water is flexible. Followed by another separated minimization which utilize conjugate descent method for 100000 steeps to reduce the maximum force less than 100 kJ*mol-1*nm-1. In this steep all the molecules were flexible. This process had been proved successfully to avoid subsequent simulation crashes. Two minimization's electrostatic interactions were calculated by Fast smooth Particle-Mesh Ewald (SPME) and van der Waals interactions were calculated by cut-off method, additionally, the cut off range of both are 1.2nm. Dispersion correction was also employed.

Equilibrium phase simulation: Before simulation, the entire system is divided into complex group (protein and ligands) and environment group (all the rest),each part controls temperature independently. For reaching the equilibrium, simulation using the the structure optimized by minimization was performed under NPT conditions, the temperature coupling method is V-rescale (tauT=0.2, refT=310), the pressure coupling method is Berendsen with isotropic coupling (tauP=2.0). This stage needed 2ns with positional restrain on complex group. After equilibrium structure obtained, use Parrinello-Rahman pressure coupling method (tauP=2.0) replace the former, and perform 100ns simulation without any restraints.The electrostatic interactions were calculated by Fast smooth Particle-Mesh Ewald (SPME) and van der Waals interactions were calculated by cut-off method, additionally, the cut off range of both are 1.0nm. Dispersion correction was also employed.

6.2 Metadynamics simulation

MetaD. Dynamics was performed in GROMACS co-compilation with plumed (GROMCAS 2023 with Plumed 2.10). Before MetaD. Dynamics was performed, it also needs modelling, minimization, equilibrium simulation. Hence, the basic procedure was same as the classical molecular dynamics. After the 100ns simulation had been done, use the cluster analysis module in GROMACS to pick up the most representative conformation as the initial structure. Then use the Well-tempered metadynamics by switch on the BIASFACTOR key words. All the keyword parameters need hyperparameter optimization process which determined by different missions and problems. The parameter file for our problem has been uploaded.S7



(a)

Fig. S7. Through Principal Component Analysis (PCA), we transformed the three-dimensional coordinates of the path when water leaves the pocket into a distribution governed by two principal components' Boltzmann distribution.

6.3 QM/MM MD simulation

QM/MM MD simulation was performed in GROMACS co-compilation with CP2K (GROMACS 2022 with CP2K). Before QM/MM MD simulation was performed, it also needs modelling, minimization, equilibrium simulation. Hence, the basic procedure was same as the classical molecular dynamics. After the 100ns simulation had been done, use the cluster analysis module in GROMACS to pick up the most representative conformation as the initial structure. The QM region includes all ligands and part of particular residue (63 72 76 110 122 124 141 179 183), the side chain of these included and alpha-carbon was excluded, the cut between side chain and alpha-carbon use Hang Bond Methods, the all QM region calculated by GFN1-xTB method implemented by CP2K. The rest region are MM region parametrized by Amber14 force field. Total simulation time is 50ps. The electrostatic interactions were calculated by Fast smooth Particle-Mesh Ewald (SPME) and van der Waals interactions were calculated by cut-off method, additionally, the cut off range of both are 1.2nm. Dispersion correction was also employed.S8

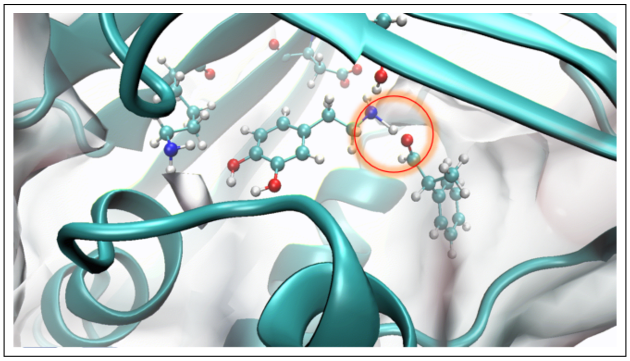


Fig. S8. Key intermediate (IM3, iminium intermediate) was captured by QMMM MD

6.4 Cluster model

Do classical MD for reactants, intermediate, products. And perform QM/MM MD simulation to obtain more stable structure. Then use cluster analysis module in GROMACS to pick up the most representative conformation as the initial structure. Then fetch all ligands and particular residue, the bonds that had previously connected other residue were now blocked by hydrogen atom. Structure optimization, transition state search and vibration analysis were calculated at M06-2X functional with def2-SVP basis and DFT-D3 dispersion correction level in chlorobenzene implicit solvent (IEFPCM implicit solvent model), the calculation is performed by Gaussian16. Single point energy was calculated at rev-DSD-PBEP86 functional with def2-TZVPP basis and DFT-D4 dispersion correction level in chlorobenzene implicit solvent (SMD implicit solvent model), the calculation is performed by ORCA (version 5.0.4). And free energy is conducted by single point energy corrected by vibration analysis, performed by Sherom package.S9

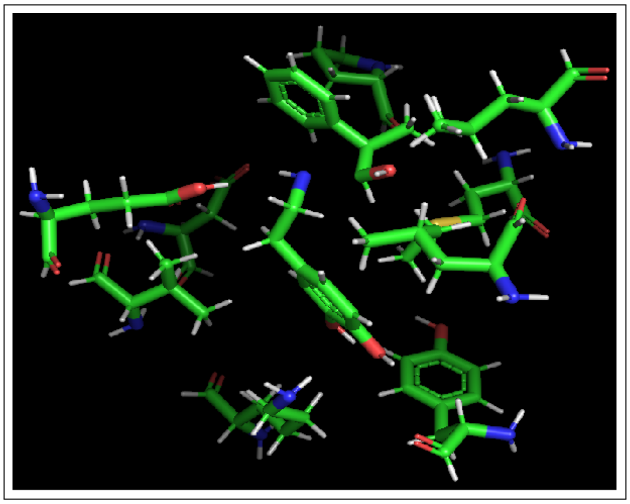


Fig. S9. We simulated the reaction of substrates within the protein pocket by constructing a residue pocket that surrounds the substrates. For other proteins not involved in the reaction, we replaced them with an implicit solvent model.

6.5 Fukui function and dual descriptors

The intermediate was optimized at M06-2X functional with def2-SVP basis and DFT-D3 dispersion correction level in chlorobenzene implicit solvent (SMD implicit solvent model), and by calculate the wavefunction of normal state (N state, charge=-1, multiplicity=1), one more electron state (N+1 state, charge=-2, multiplicity=2), one less electron state (N-1 state, charge=0, multiplicity=2). The 2(C) is the 1 position carbon, the 3(C) is the 2 position carbon, the 22(C) is the carbon in imino group.

Table S4. Atom Electrophilicity and Nucleophilicity Estimated by Fuki Function and Double Descriptors Method

Atom	Electrophilicity	Nucleophilicity
1(C)	0.01478	0.07174
2(C)	0.01176	0.16098
3(C)	0.00259	0.12291
4(C)	0.01132	0.06610
5(C)	0.01227	0.13953
6(C)	0.01914	0.14556
7(O)	0.01051	0.13633
8(O)	0.01409	0.11697
9(C)	0.03996	0.02276
10(C)	0.06893	0.01431
11(N)	0.37243	0.00230
12(H)	0.01133	0.04851
13(H)	0.00617	0.06415
14(H)	0.00788	0.04681
15(H)	0.00754	0.04740
16(H)	0.00815	0.05240
17(H)	0.02655	0.02316
18(H)	0.03256	0.02658
19(H)	0.06118	0.00868
20(H)	0.07338	0.01030
21(H)	0.16503	0.00552
22(C)	0.62491	0.01164
23(H)	0.23603	0.00767
24(C)	0.06613	0.00056
25(H)	0.06378	-0.00238
26(C)	0.06606	0.00182
27(H)	0.06978	0.00201
28(H)	0.05706	-0.00007
29(H)	0.04220	0.00419
30(C)	0.01484	-0.00070
31(C)	0.05452	-0.00278
32(C)	0.03834	0.00325
33(C)	0.04605	0.00079
34(H)	0.03314	-0.00415
35(C)	0.04657	0.00487
36(H)	0.02782	0.00266
37(C)	0.07495	0.00470
38(H)	0.03110	0.00055
39(H)	0.02971	0.00344
40(H)	0.03772	0.00317

6.6 Molecular Mechanics / Poisson Boltzmann (Generalized Born) Surface Area(MMPBSA)

To calculate the binding free energy and interaction between enzyme and its substrate, molecular mechanics Poisson Boltzmann surface area (MMPBSA) was used [26] based on 1000 frames of trajectory with an interval of 1. Without considering the entropy term, the calculated value became the effective binding free energy (ΔG_{bind}), which was calculated by Eq. (1).

$$\Delta G_{\text{bind}} = \Delta G_{\text{gas}} + \Delta G_{\text{sol}} \quad (1) \tag{S1}$$

Here, ΔG_{gas} is the molecular mechanical energy in the gas phase, and ΔG_{sol} is the solvation energy. The process was covered through the thermodynamic cycle. Then, the ΔG_{gas} of protein-substrate combinations could be further calculated by Eq. (2):

$$\Delta G_{\text{gas}} = \Delta E_{\text{bonded}} + \Delta E_{\text{non-bonded}} = (\Delta E_{\text{bond}} + \Delta E_{\text{angle}} + \Delta E_{\text{dihedral}}) + (\Delta E_{\text{ele}} + \Delta E_{\text{vdW}}) \quad (2) \tag{S2}$$

In Eq. (2), ΔE_{bonded} includes the molecular internal energies: ΔE_{bond}, ΔE_{angle}, and $\Delta E_{\text{dihedral}}$. The non-bonded interaction ($\Delta E_{\text{non-bonded}}$) is composed of electrostatic (ΔE_{ele}) and van der Waals (ΔE_{vdW}) interactions. Since the dynamic process does not involve the breaking or formation of intramolecular bonds, ΔG_{gas} can also be expressed as the sum of ΔE_{ele} and ΔE_{vdW}. Then, the solvation energy was calculated by Eq. (3):

$$\Delta G_{\text{sol}} = \Delta G_{\text{polar}} + \Delta G_{\text{non-polar}} \quad (3) \tag{S3}$$

In Eq. (3), ΔG_{polar} is the electrostatic or polar component of the solvation free energy evaluated by the Poisson-Boltzmann (PB) model, and $\Delta G_{\text{non-polar}}$ is the hydrophobic or nonpolar component proportional to the molecular solvent accessible surface area (SASA).

Fig. S10. MMPBSA Energy Decomposition for Products

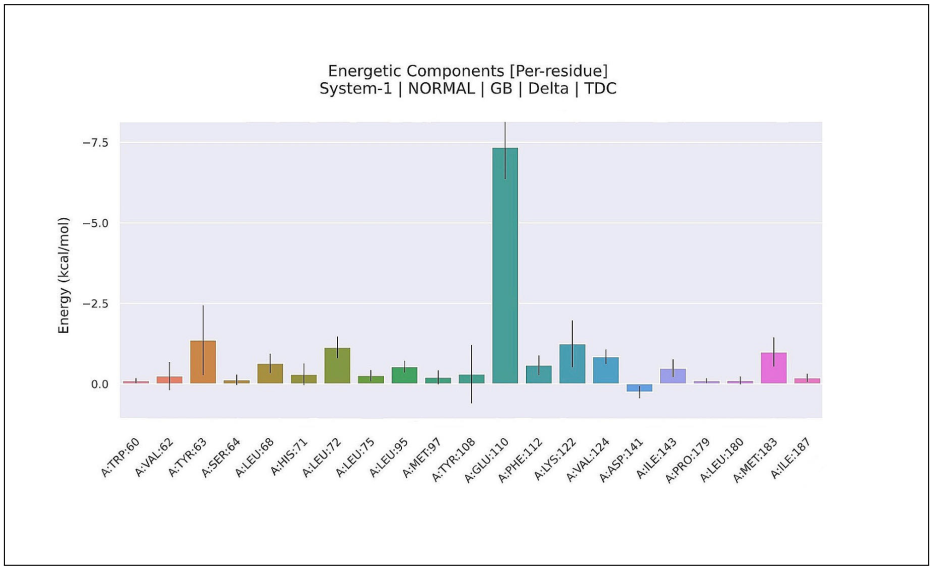


Fig. S11. MMPBSA Energy Decomposition for Dopamine

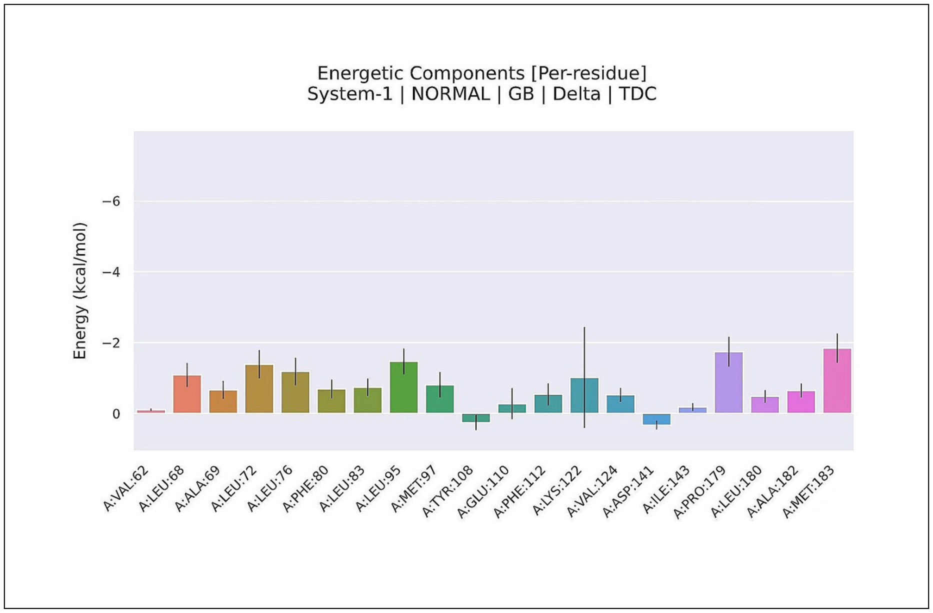
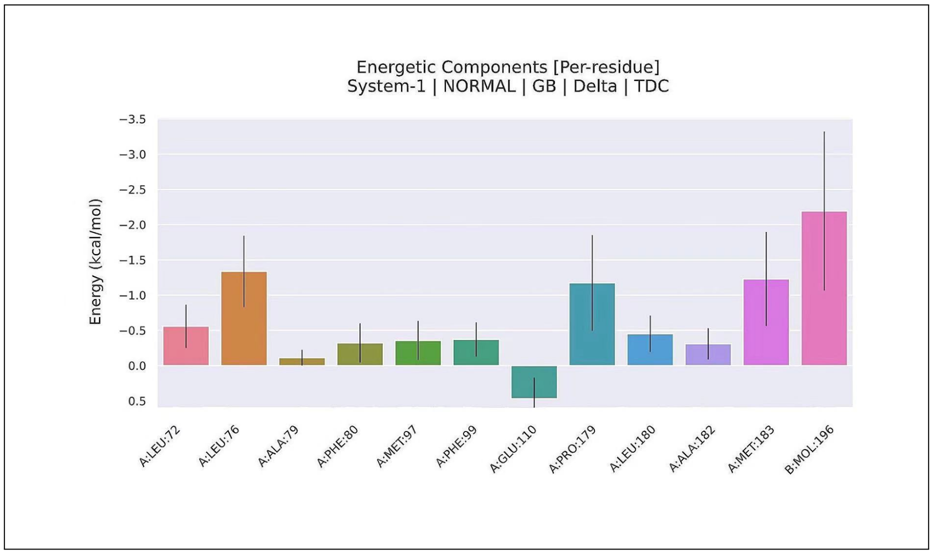
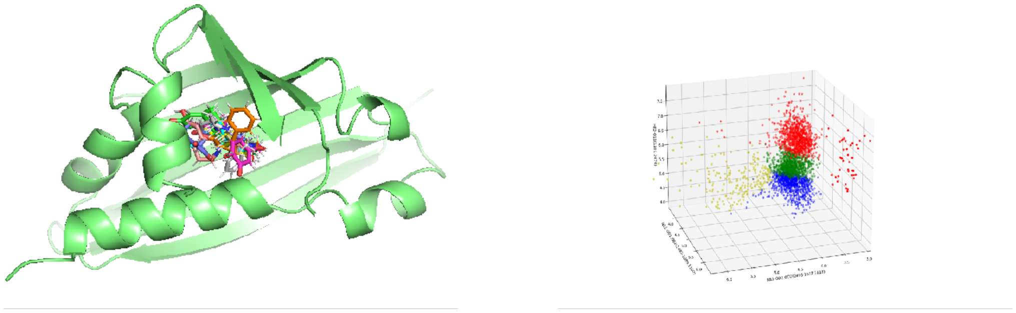


Fig. S12. MMPBSA Energy Decomposition for 2-benzenepropanal



6.7 Molecular docking and clustering analysis

To obtain more reasonable docking results, we first generated 1000 protein conformations using Rosetta optimization. Subsequently, we used Ledock to perform substrate docking on the top-scoring proteins. Finally, we employed the k-means algorithm to extract the xyz coordinates of the substrate molecules to determine the most reasonable conformations.



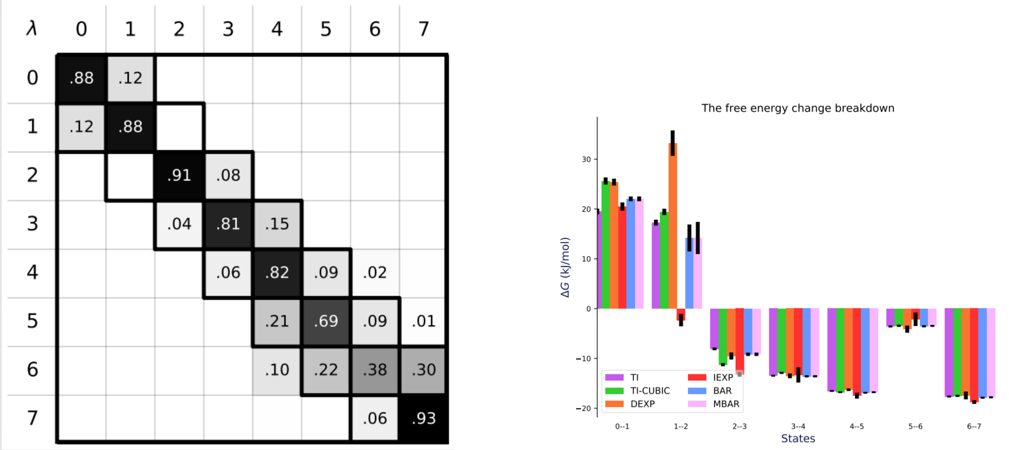
(a) Multiple potential conformations of a single molecule within a protein pocket

(b) Clustering analysis of multiple potential conformations

Fig. S13. The initial conformation of the molecular dynamics simulation is determined jointly by cluster analysis and scoring functions

6.8 Free energy perturbation(FEP) calculation

To obtain the most accurate interaction energy, we employed the Free Energy Perturbation method. Analyzing the results using the FEP module in GROMACS and the alchemist Python library, we ultimately obtained the interaction energy.



λ	0	1	2	3	4	5	6	7
0	.88	.12						
1	.12	.88						
2			.91	.08				
3			.04	.81	.15			
4				.06	.82	.09	.02	
5					.21	.69	.09	.01
6					.10	.22	.38	.30
7							.06	.93

(a) The adjacency matrix provides an intuitive reflection of the quality of the FEP calculations. Our adjacency matrix exhibits a high degree of repeatability, indicating good overlap between different windows, and therefore, the computational results are reliable.

(b) Energy breakdown reflects the contributions of various energies across different windows.

Fig. S14

7. CHEMICAL SYNTHESES

Enzymatic scale-up procedure

A solution of dopamine hydrochloride (18.9 mg, 0.10 mmol) and sodium ascorbate (10 mg, 0.05 mmol) in Hepes buffer (9 mL, 100 mM, pH 7.5) was prepared. Aldehyde (0.20 mmol) in acetonitrile (1 mL) was added followed by Δ33TfNCS (100 mL, 10 mg mL-1 in 20 mM Tris HCl, 50 mM NaCl, pH 7.5). The reaction mixture was stirred under an inert atmosphere at 37 °C for 10 h. The reaction was quenched by addition of HClaq (1 M, 1 mL) and adjust pH to about 6.5. The reaction mixture was extracted with A mixed solution of dichloromethane and methanol (in a volume ratio of 10:1, 3×15 mL). The organic phases were combined, dried and concentrated under reduced pressure to give the product.

8. LCMS DATA FOR PRODUCTS

Fig. S15. MS data for 1-(1-phenylethyl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol

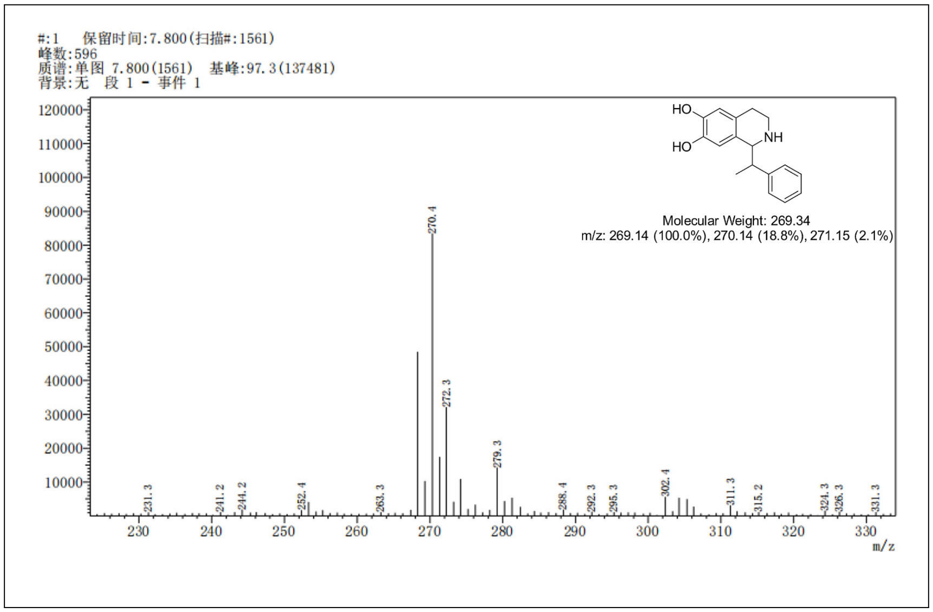


Fig. S16. MS data for 1-(2-fluorophenyl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol

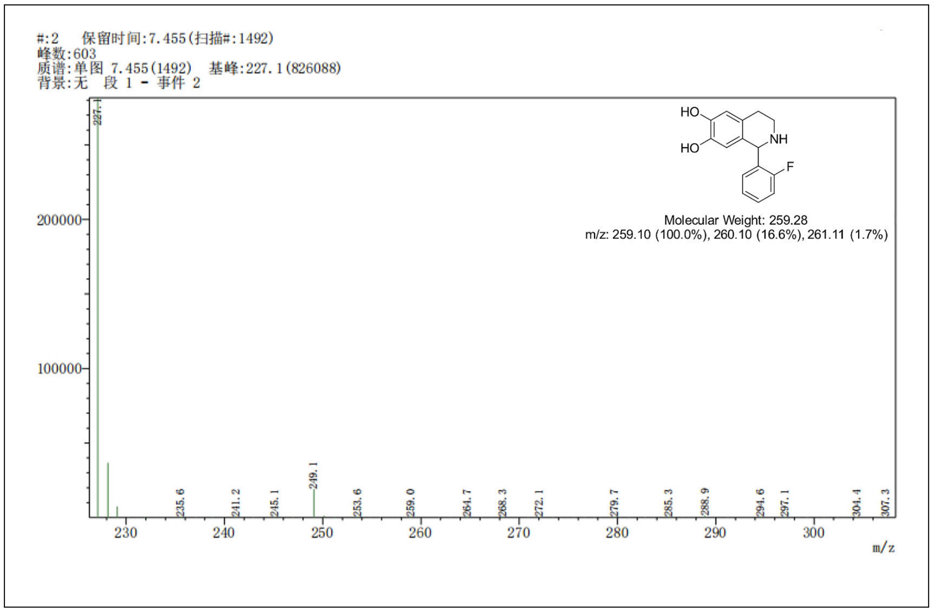
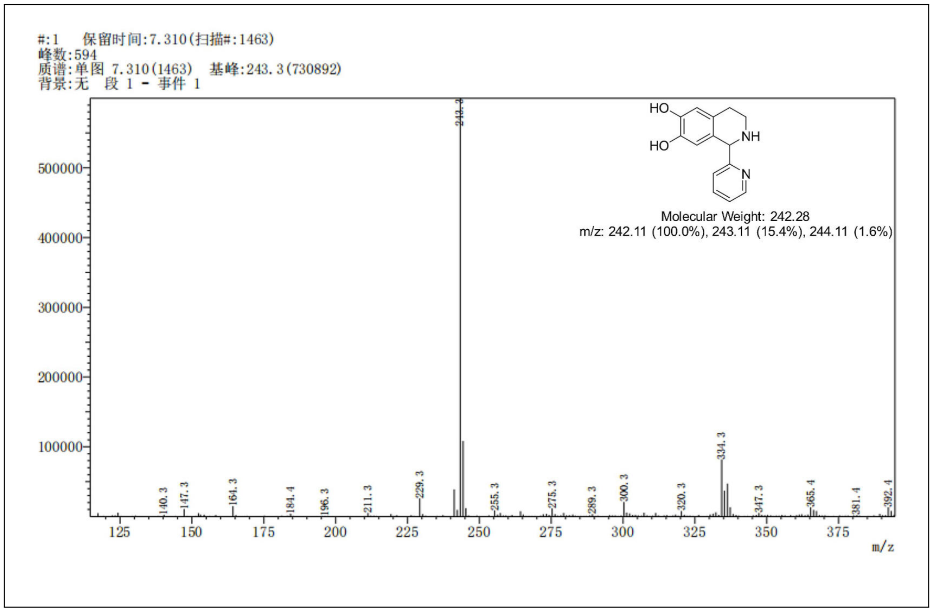


Fig. S17. MS data for 1-(pyridin-2-yl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol



9. GC DATA FOR PRODUCTS

Fig. S18. GC data for 1-(1-phenylethyl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol

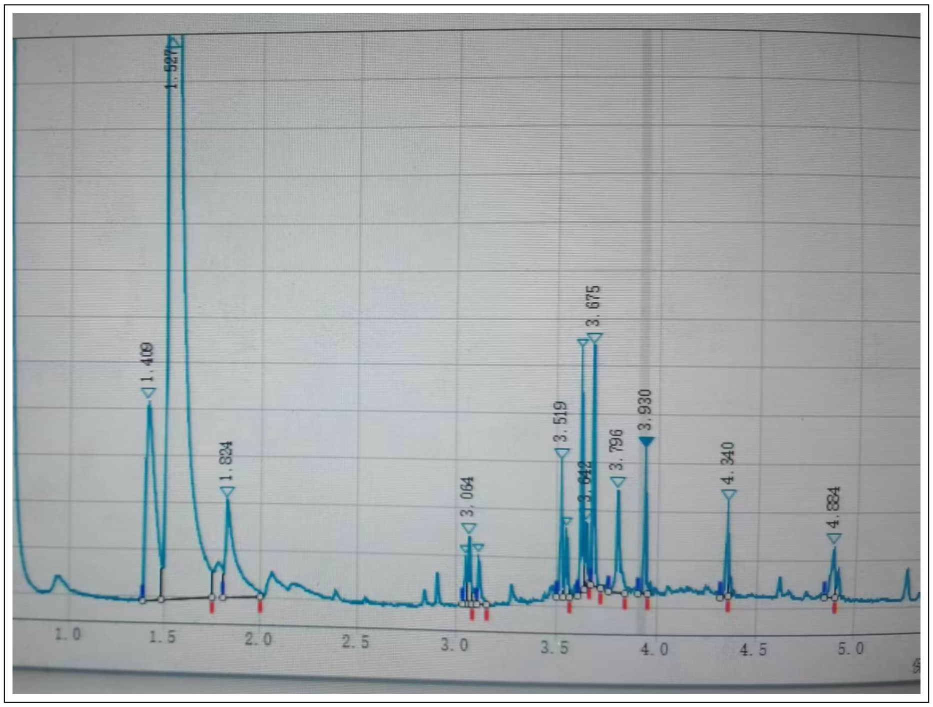


Fig. S19. GC data for 1-(pyridin-2-yl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol

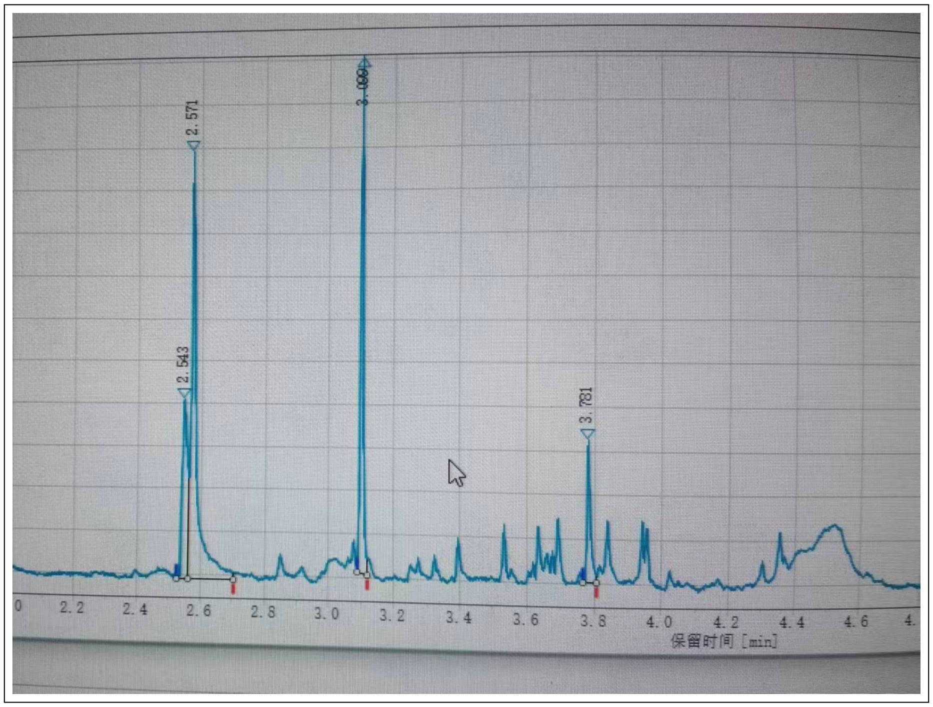


Fig. S20. GC data for 1-(pyridin-3-yl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol

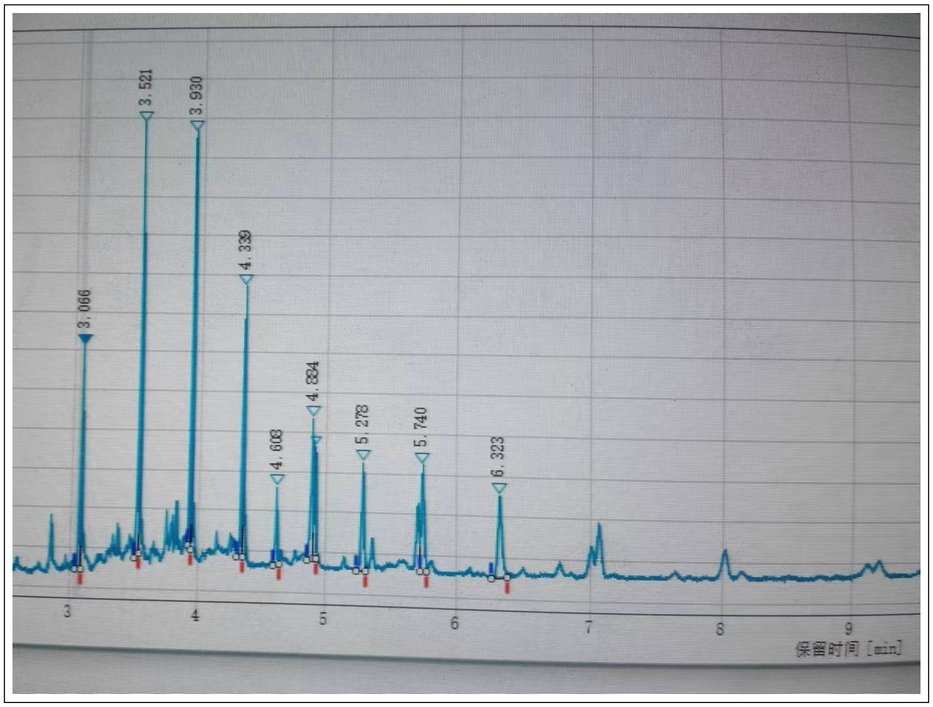


Fig. S21. GC data for 1-(1H-pyrrol-2-yl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol

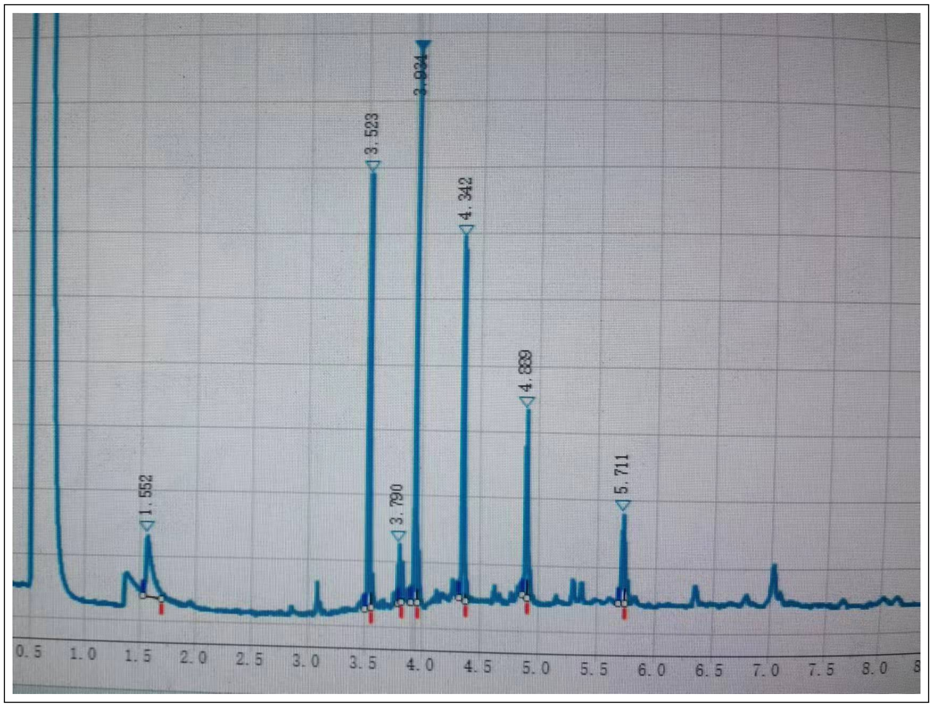
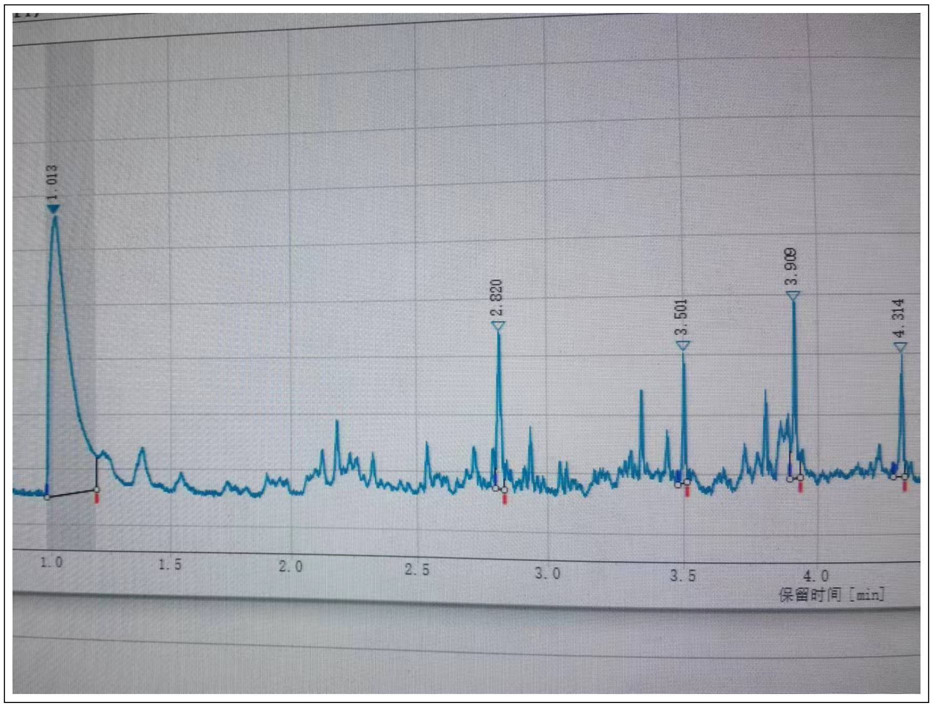


Fig. S22. GC data for 1-(2-fluorophenyl)-1,2,3,4-tetrahydroisoquinoline-6,7-diol



10. DEEP LEARNING MODEL AND BENCHMARK

10.1 Principle of RetKcat

In this work,we developed an end-to-end learning approach for in vitro Kcat value prediction by combining a GCN for substrates and a RetNet for proteins.Molecular which was atoms linked with chemical bond can be naturally converted into a graph and protein sequences can also be seen as a special format of list.

$$\begin{aligned} G_{\text{mol}} &= (V,E)_{V=\text{atoms},\, E=\text{chemical bonds}} \\ X_p &= x_1 \ldots x_{|\text{sequence length}|}, \quad x \in \{\text{Amino Acid}\} \end{aligned}$$

First,substrate SMILES information was loaded with RDKit v.2022.9.5(https://www.rdkit.org) and then each node will update each node via its neighbour around,which can be seen as divide atoms with its chemical environment[2].Moreover the adjacency of molecule was extracted ,the molecule was finally represented as adjacency and a ordered node list.Then the edge information and node information has been convoluted . The final output of the GCN is a real-valued matrix M.

$$\begin{aligned} H_{x+1} &= f(A,H_i) = \sigma(D^{-\frac{1}{2}}\widetilde{A}D^{-\frac{1}{2}}W_i H_x) \\ [A(G)]_{ij} &= \begin{cases} 1, & v_i v_j \in E \\ 0, & \text{otherwise} \end{cases} \\ D_{ij} &= \sum_{j=1}^{N} A_{Nj} \end{aligned}$$

The protein sequence is manually split into 'words' which contain N amnion acids[3].Every word is corresponding with a real number.Windows was set to limit the length of words list , every N amnion acid is transfer into number and hold by windows respectively.Then the matrix maybe embedding to appointed dimension.The protein representation and molecule representation will has same dimension and will be concentrated as the input of RetNet.

$$\begin{aligned} Q = (XW_Q) \odot \Theta, K &= (XW_K) \odot \overline{\Theta}, V = XW_V \\ \Theta n &= e^{in\theta}, \\ D_{nm} &= \begin{cases} \gamma^{n-m}, & n \geq m \\ 0, & n < m \end{cases} \\ Retention(X) &= (QK^T \odot D)V \end{aligned}$$

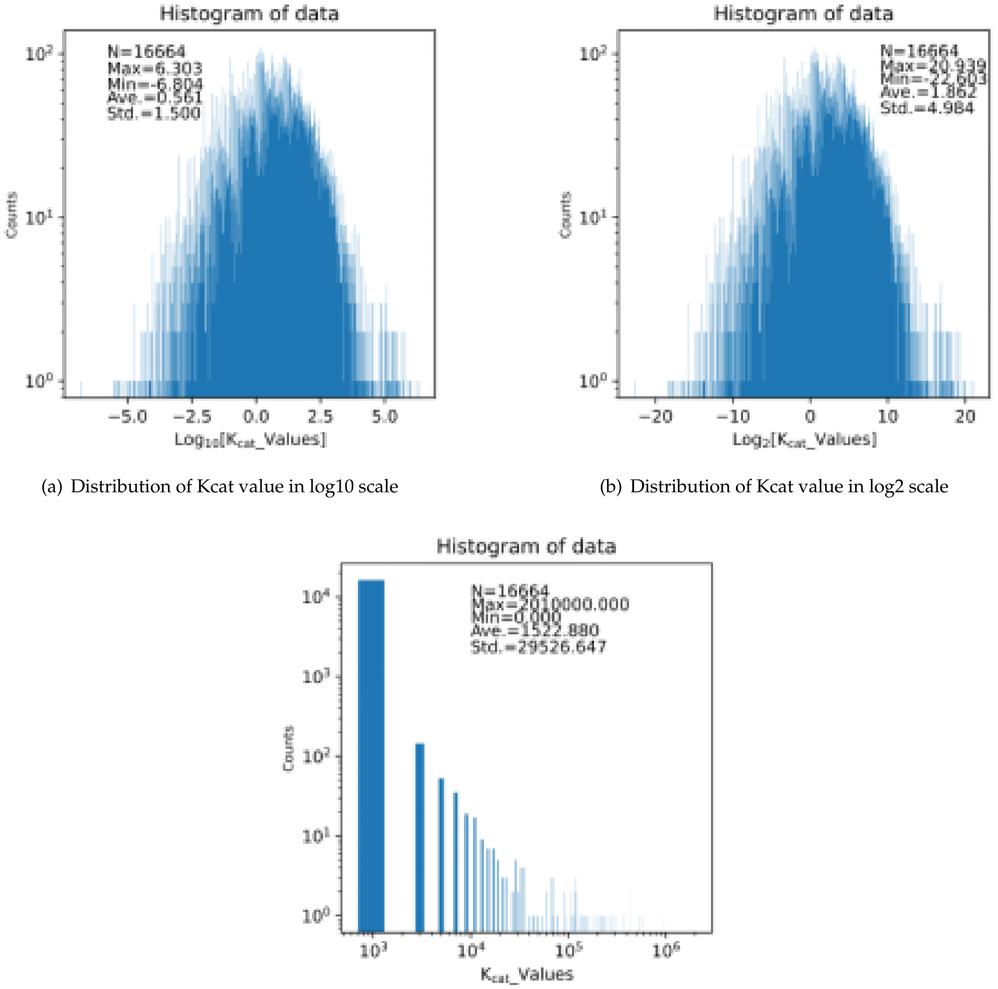
The outcome of RetNet will forward an output layers,which is consisting of several Liner ,then the vector will be turn to predict value via a single layer Liner.

$$\begin{aligned} RMSE(X,h) &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2} \\ MAE(X,h) &= \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i| \\ R\ square &= \frac{SSR}{SST} \\ SSR &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y_i})^2, \quad SST = \sum_{i=1}^{n}(y_i - \overline{y_i})^2 \end{aligned}$$

The predict will be evaluated by MAE (mean absolute error), RMSE (root mean squared error) and R square(Coefficient of determination).

10.2 Data distribution

We further integrated and curated the BRENDA[4] and SABIO-RK[5] databases to ensure that our dataset contains samples with substrates' SMILES, protein sequences, Kcat values, and that these values are unique (SI). Notably, we removed protein samples related to the NCS family (E.C.Number 4.2.1.78) from the database to ensure that our model hasn't seen these proteins in advance.Moreover the datasets was filtered by the rules below:(1)Kcat Value >=0,(2)there is no '.' in molecules' SMILES format which means no certain bond is abnormal for model to comprehend. The final datasets contained 16,416 unique entries catalysed by 7,651 unique protein sequences and converting 2,619 unique substrates. This dataset was randomly split into training, validation by 90%, 10% respectively, the test dataset is made up of the NCS samples whose Kcat were determined by experiment reported previously.



(a) Distribution of Kcat value in log10 scale

(b) Distribution of Kcat value in log2 scale

(c) Distribution of Kcat value

Fig. S23. The data analysis for Kcat value we extract for DLKcat after processing, cleaning and combination,raw data has been filter by DLKcat before.Kcat value has been convert into log scale which make the data distribution more reasonable .There was 17010 samples in DLKcat datasets 346 has been removed.

10.3 Content of testsets

Task Name	Substrate SMILES	DLKcat predict value (1/s)	Experiment	RetKcat predict value (1/s)
L68T/M97V-4-Biphenylcarboxaldehyde	Oc1c(O)cc(cc1)CC/N=C/c2ccc(cc2)-c3ccccc3	5.0320	0.002	12.1898
L68T/M97V-benzaldehyde	c1ccc(cc1)/C=N/CCc2cc(O)c(O)cc2	8.0897	0.439	11.1961
L76A-4-HPAA	Oc1ccc(cc1)C/C=N/CCc2cc(O)c(O)cc2	2.8128	13.95	8.1638
L76A-citronellal	CC(C)=CCC[C@@H](C)C/C=N/CCc1cc(O)c(O)cc1	3.2442	3.02	8.5014
L76A-Hexanal	CCCCC/C=N/CCc1cc(O)c(O)cc1	2.8628	7.3	7.0309
WT-4-Biphenylcarboxaldehyde	Oc1c(O)cc(cc1)CC/N=C/c2ccc(cc2)-c3ccccc3	4.4103	0.001	11.2398
WT-4-HPAA	Oc1ccc(cc1)C/C=N/CCc2cc(O)c(O)cc2	3.7487	21.35	7.0767
WT-benzaldehyde	c1ccc(cc1)/C=N/CCc2cc(O)c(O)cc2	7.1997	0.234	10.3177
WT-citronellal	CC(C)=CCC[C@@H](C)C/C=N/CCc1cc(O)c(O)cc1	4.4151	1.6	7.7147
WT-Hexanal	CCCCC/C=N/CCc1cc(O)c(O)cc1	3.9151	5.55	6.0782
Y108F-4-HPAA	Oc1ccc(cc1)C/C=N/CCc2cc(O)c(O)cc2	3.6528	10.2	7.4050
Y108F-Hexanal	CCCCC/C=N/CCc1cc(O)c(O)cc1	4.4288	3.55	6.3659

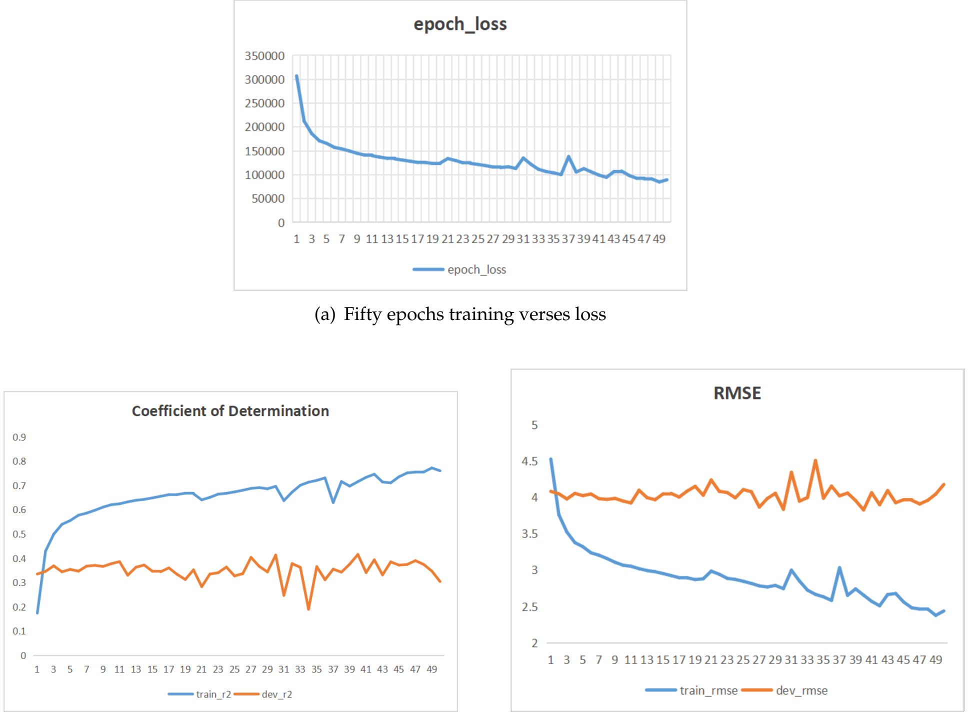
Table S5. The data analysis for Kcat value we extract for DLKcat after pre-process, cleaning and combination,raw data has been filter by DLKcat before.Kcat value has been convert into log scale which make the data distribution more reasonable .There was 17010 samples in DLKcat datasets 236 has been removed .

10.4 Sequence of enzyme used in testsets

Mutant	Sequence
L68T/M97V	MMKMEVVFVFLMLLGTINCQKLILTGRPFLHHQGIINQVSTVTKVIHHELEVAASADDIWTVYSWPGTAKHLPDLLPGAFEKLEIIGDGGVGTILDVTFVPGEFPHEYKEKFILVDNEHRLKKVQMIEGGYLDLGVTYYMDTIHVVPTGKDSCVIKSSTEYHVKPEFVKIVEPLITTGPLAAMADAISKLVLEHKSKSNSDEIEAAIITV
L76A	MMKMEVVFVFLMLLGTINCQKLILTGRPFLHHQGIINQVSTVTKVIHHELEVAASADDIWTVYSWPGLAKHLPDLAPGAFEKLEIIGDGGVGTILDMTFVPGEFPHEYKEKFILVDNEHRLKKVQMIEGGYLDLGVTYYMDTIHVVPTGKDSCVIKSSTEYHVKPEFVKIVEPLITTGPLAAMADAISKLVLEHKSKSNSDEIEAAIITV
WT	MMKMEVVFVFLMLLGTINCQKLILTGRPFLHHQGIINQVSTVTKVIHHELEVAASADDIWTVYSWPGLAKHLPDLLPGAFEKLEIIGDGGVGTILDMTFVPGEFPHEYKEKFILVDNEHRLKKVQMIEGGYLDLGVTYYMDTIHVVPTGKDSCVIKSSTEYHVKPEFVKIVEPLITTGPLAAMADAISKLVLEHKSKSNSDEIEAAIITV
Y108F	MMKMEVVFVFLMLLGTINCQKLILTGRPFLHHQGIINQVSTVTKVIHHELEVAASADDIWTVYSWPGLAKHLPDLLPGAFEKLEIIGDGGVGTILDMTFVPGEFPHEFKEKFILVDNEHRLKKVQMIEGGYLDLGVTYYMDTIHVVPTGKDSCVIKSSTEYHVKPEFVKIVEPLITTGPLAAMADAISKLVLEHKSKSNSDEIEAAIITV

Table S6.

10.5 Training convergence analysis



(a) Fifty epochs training verses loss

(b) Fifty epochs training verses R-squre

(c) Fifty epochs training verses RMSE

Fig. S24. Through plotting the curves of loss, R-squared, and RMSE, we find that 50 epochs of training lead the model to converge.

REFERENCES

1. R. Roddan, G. Gygli, A. Sula, *et al.*, “Acceptance and kinetic resolution of -methyl-substituted aldehydes by norcoclaurine synthases,” ACS Catal. p. n. pag. (2019).
2. A. Tsubaki, K. Tomii, and J. Sese, “Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences,” Bioinformatics **35**, 309–318 (2019).
3. B. Dong, Q.-W. Wang, X.-L. Wang, and L. Lin, “Application of latent semantic analysis to protein remote homology detection,” Bioinformatics **22**, 285–290 (2006).
4. I. Schomburg, L. Jeske, M. Ulbrich, *et al.*, “The brenda enzyme information system–from a database to an expert system,” J. Biotechnol. **261**, 194–206 (2017).
5. U. Wittig, M. Rey, A. Weidemann, *et al.*, “Sabio-rk: an updated resource for manually curated biochemical reaction kinetics,” Nucleic Acids Res. **46**, D656–D660 (2018).