RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2019, **9**, 13868

Using molecular dynamics simulations to evaluate active designs of cephradine hydrolase by molecular mechanics/Poisson–Boltzmann surface area and molecular mechanics/generalized Born surface area methods†

Jing Xue,[a] Xiaoqiang Huang[a] and Yushan Zhu *[ab]

The poor predictive accuracy of current computational enzyme design methods has led to low success rates of producing highly active variants that target non-natural substrates. In this report, a quantitative assessment approach based on molecular dynamics (MD) simulations was developed to eliminate false-positive enzyme designs at the computational stage. Taking cephradine hydrolase as an example, the apparent Michaelis binding constant (K_m) and catalytic efficiency (k_{cat}/K_m) of designed variants were correlated with binding free energies and activation energy barriers, respectively, as calculated by molecular mechanics/Poisson–Boltzmann surface area (MM/PBSA) and molecular mechanics/generalized Born surface area (MM/GBSA) methods with explicit water considered based on general MD simulation protocols. The correlation results showed that both the MM/GBSA and MM/PBSA methods with a protein dielectric constant ($\varepsilon_p = 4$) could rank the variants well based on the predicted binding free energies between enzyme and the substrate. Furthermore, the activation energy barriers calculated by the MM/PBSA method with an $\varepsilon_p = 24$ correlated well with k_{cat}/K_m. Thus, false-positive variants obtained by the enzyme design program PRODA were eliminated prior to experimentation. Therefore, MD simulation-based quantitative assessment of designed variants greatly enhanced the predictive accuracy of computational enzyme design tools and should facilitate the construction of artificial enzymes with high catalytic activities toward non-natural substrates.

Received 31st March 2019
Accepted 30th April 2019

DOI: 10.1039/c9ra02406a

rsc.li/rsc-advances

Introduction

As efficient biocatalysts, enzymes are used widely in industry to accelerate chemical reactions and obtain high selectivity and specificity under ambient conditions.[1,2] However, native enzymes cannot meet the increasing demands of green process developments. The use of native enzymes toward non-natural substrates often suffers from low activity and production of worthless by-products.[3] In the last two decades, directed evolution approaches have succeeded in adapting native enzymes to catalyze non-natural chemical transformations through high-throughput screening.[1] An alternative and more economical way to adapt native enzymes is the structure-based computational enzyme design approach. With the help of molecular modeling and high-performance computing, researchers can optimize or create enzyme structures that preferentially stabilize the transition state relative to the ground state, thereby reducing the major free energy barrier along the reaction coordinate. Computational enzyme design can be divided into two subclasses: the *de novo* design of new active sites and the redesign of existing active sites. The *de novo* design approach requires the construction of an active site model (theozyme)[4] for the new reaction. This model is then placed in a suitable protein scaffold[5] and the binding pocket is redesigned to stabilize the anchored active site. *De novo* design has been used successfully to generate artificial enzymes for the Kemp elimination reaction,[6] retro-aldol reaction,[7] and Diels–Alder reaction.[8] Although the catalytic activities of the *de novo* designed enzymes are always modest, this approach produces good starting points for further evolution of the enzyme by experimental approaches. In the redesign approach, the native enzyme active sites are tailored to catalyze reactions for non-natural substrates. The computational active site redesign strategy is capable of replacing the substantial workload of directed evolution to adapt native enzymes with enhanced activity and selectivity for non-natural substrates.[9–14]

Although the computational enzyme design strategy has been effective and shown potential in these cases, the success

[a]*Department of Chemical Engineering, Tsinghua University, Beijing 100084, China*
[b]*MOE Key Lab for Industrial Biocatalysis, Tsinghua University, Beijing 100084, China. E-mail: yszhu@tsinghua.edu.cn*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ra02406a

rate has not improved. Generally, assumptions such as a rigid backbone, discrete amino acid side chain rotamers and discrete ligand placement, and continuum solvent are adopted in current enzyme design algorithms. These trade-offs facilitate the search of the sequence space but sacrifice the model accuracy for a protein–ligand interaction. This results in the algorithms producing a large number of false positive variants besides the few beneficial variants. To filter out the false positive variants at the computational stage, various methods such as mixed quantum and molecular mechanics (QM/MM),[15,16] semi-empirical valence bond[17] and molecular dynamics (MD) simulations[18–22] have been applied. The results of these studies show that MD simulations can be used to evaluate enzyme designs effectively.[16] However, evaluation criteria applied in these studies such as active site root-mean-square deviations (RMSDs), water coordination numbers, hydrogen-bond distances and angles along the simulation trajectories can only be used to rank the designed variants qualitatively. In this report, we develop a direct way to correlate the free energy changes calculated based on the MD trajectories with kinetic parameters, *i.e.*, the turnover number k_{cat} and the apparent Michaelis constant K_m, of the designed variants to rank these variants quantitatively. Several MD simulation based methods are developed to calculate ensemble averaged thermodynamic quantities, such as the linear interaction energy (LIE),[23] thermodynamic integration (TI),[24] free energy perturbation (FEP),[25] and Molecular Mechanics/Generalized Born or Poisson–Boltzmann Surface Area (MM/GBSA and MM/PBSA).[26,27] Although LIE, TI and FEP provide relatively accurate free energy calculations, the computational cost of these methods is very unfavorable. In contrast, the MM/GBSA and MM/PBSA methods are computationally economical and have been used successfully to predict ligand binding free energies[28,29] to screen small molecule inhibitors in drug design. Here, we use MM/GBSA and MM/PBSA methods to calculate the binding free energies between enzyme active sites and the substrates at various states based on MD simulation trajectories to evaluate the designed variants quantitatively.

In our previous study,[30] a new scaffold (PDB ID: 1JU3), which is a *Rhodococcus* sp. cocaine esterase (EC 3.1.1.84, Uniprot ID: Q9L9D7), was identified from structural databases to catalyze the hydrolysis and synthesis of cephradine, an important semi-synthetic β-lactam antibiotic. The hydrolytic reaction of cephradine catalyzed by the enzyme is shown in Fig. 1A. The catalytic hydrolysis of cephradine by the wild-type enzyme is low. Several hundred variants were designed by our computational enzyme design program PRODA to increase the activity[31–34] and the kinetic parameters of the top-ranked eleven variants were experimentally measured. Among the eleven designed variants, only one variant showed higher catalytic efficiency (k_{cat}/K_m) than the wild-type enzyme, indicating a success rate of less than 10%. To eliminate all or at least part of the false positive predictions by PRODA, an MD simulation was used here to evaluate the active designs of cephradine synthase by MM/GBSA and MM/PBSA methods quantitatively.

Materials and methods

Structure modeling

The crystal structure of wild-type cocaine esterase from *Rhodococcus* sp. (PDB ID: 1JU3) was taken directly from the Protein Data Bank (PDB) without further minimization. The water molecules were removed, and hydrogen atoms were added using PRODA by virtue of the topology parameters of the all-atom CHARMM 22 force field. The crystal structure of cephradine was taken directly from the Cambridge Structural Database and the heavy atom names of cephradine are shown in Fig. 1B. The geometries of cephradine in the transition and Michaelis binding states (Fig. 2A) were calculated by PRODA.

The geometry of cephradine at the transition state (TS) is shown in Fig. 2A, where the central atom (C15) adopts a tetrahedral intermediate form. The catalytic geometrical relationships between the TS and the catalytic residues of the protein scaffold are shown in Fig. 2A. The catalytic triad (Ser117–His287–Asp259) forms the nucleophilic attacking group while the backbone amido group of Try118 and the hydroxyl group of Tyr44 constitute the oxyanion hole that stabilizes the negatively charged O16 atom. To represent the translational, rotational and conformational freedoms of the TS in the active site, a rotamer library of the TS with 5470 conformers was generated using the targeted small molecule placement approach developed in our earlier work,[33] which is based on the catalytic geometrical constraints (Table S1†) and placing rules (Table S2†). The atomic parameters for cephradine were obtained from the model compounds of the CHARMM 22 force field and the atomic partial charges were assigned based on the PARSE model. In the active site of the protein scaffold, a total of 36 residues (N42, Y44, W52, T54, Q55, S56, H87, V116, S117, Y118, L119, V121, S140, M141, S143, L146, A149, P150, W151, A162, W166, L169, I170, W220, W235, D259, F261, E264, S265, W285, S286, H287, S288, L290, L407, F408) were subjected to side chain conformational optimization. A backbone-independent rotamer library compiled by Xiang and Honig,[35] which contains 11 810 original rotamers, was used to model the side chain conformations of the design sites. The atomic and internal coordination parameters for amino acids were taken from the all-atom force field CHARMM 22. The protein–ligand interaction for the enzyme–TS complex in PRODA was calculated using the MM/GBSA form of the free energy function. The enzyme–TS complex was repacked and global minimum energy conformations (GMEC) were identified using the combinatorial optimization algorithm developed in our previous work.[33] The optimized conformation was used as the starting geometry of cephradine for the subsequent MD simulations at the TS. The geometry of cephradine at the Michaelis binding state (Fig. 2B) was obtained in a similar way to that described above. However, the amide bond in the geometry of cephradine adopted the planar form and the distance between the C15 atom of cephradine and the OG atom of Ser117 was extended to van der Waals contact ranges. The catalytic geometrical constraints and placing rules for small molecular placement at the Michaelis binding state are presented in Tables S3 and S4,† respectively.

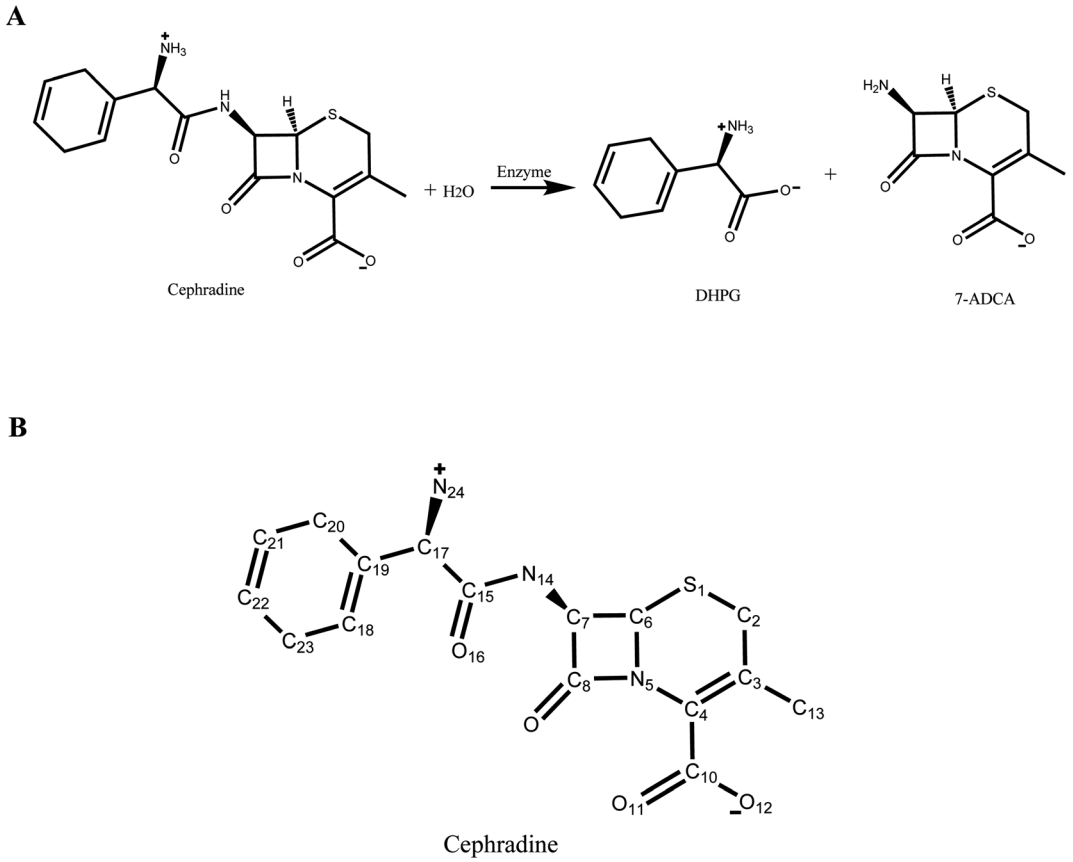

Fig. 1 Hydrolysis of cephradine catalyzed by enzyme. (A) Reaction scheme. 7-ADCA: 7-aminodesacetoxycephalosporanic acid, DHPG: D-dihydrophenyl-glycine. (B) Structural formula of cephradine marked with heavy atom names.

Molecular dynamics simulation

The MD simulations were carried out with the Desmond package.[36] In each case, an orthorhombic solvent box was constructed with a distance of 10 Å from the solute, and up to 15 000 water molecules represented by TIP3P[37] were added to the box. The charge of the system was neutralized by addition of an appropriate number of Na^+ or Cl^- counter ions to the solvent box. The atomic partial charges of cephradine were fitted using the RESP techniques[38] by the AmberTools module *antechamber*, where the electrostatic potential of cephradine was calculated using the Hartree–Fock (H F)/6-31G* basis set by Gaussian 09.[39] In the molecular mechanics energy calculation, the Amber force field *ff99SB-ILDN*[40] was used for enzymes and the general Amber force filed *gaff*[41] was used for cephradine. The system was relaxed at 10 K by 100 ps Brownian dynamics NVT simulation with restraints on the solute heavy atoms (50 kcal mol^{-1}). The solvent box was then equilibrated at 10 K by 12 ps of NVT simulation and 12 ps of NPT simulation with restraints on the solute heavy atoms (50 kcal mol^{-1}). The system was then heated to 300 K and full equilibration was performed in the NPT ensemble for 12 ps with restraints on the solute heavy atoms (50 kcal mol^{-1}). The last relaxation procedure was a 24 ps NPT dynamics run at 300 K without restraints. Finally, unrestrained production runs were performed at 300 K for 20 ns, which is a length that is considered adequate for similar calculations.[16] The M-SHAKE[42] algorithm was used to restrain bonds involving hydrogen atoms to their equilibrium lengths. Non-bonded interactions were truncated at 9 Å and the Gaussian split Ewald method[43] with a 60 × 40 × 60 mesh was used for electrostatic interactions. The constant temperature and pressure were controlled using a Berendsen thermostat or manostat.[44] Approximately 20 000 snapshots were deposited with a time space of 1 ps. The Desmond built-in post-analysis utility *Simulation Event Analysis* was used to extract statistics information, including distances, angles and RMSDs from the MD records. The MD simulation was run on a computer cluster with 64 cores and usually took 35 h to finish.

Binding free energy calculations by MM/PBSA and MM/GBSA

In the MM/PBSA or MM/GBSA methods, the binding free energy (ΔG_{bind}) between the substrate (S) and the enzyme (E) to form a complex ES is calculated as:

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{pol} + \Delta G_{np} - T\Delta S \quad (1)$$

where ΔE_{MM} represents the gas-phase molecular mechanics energy, including bond stretching, angle bending, torsion rotation, van der Waals and electrostatic contributions, as:

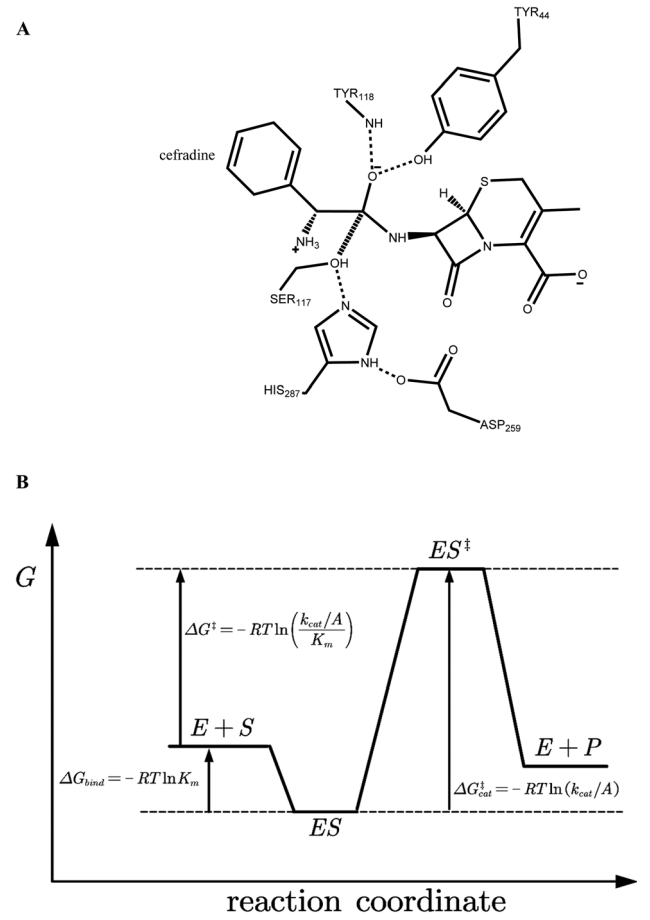

Fig. 2 (A) Catalytic geometrical constraints of cephradine hydrolase. The hypothesized bond between atoms C15 and OG is shown in hashed line. Hydrogen bonds are shown in dashed lines. (B) Reaction coordinate of enzymatic reaction. E + S represents the unbound form of enzyme and substrate, ES for the Michaelis binding state, and ES‡ for the transition state. R indicates the gas constant, K_m is the apparent Michaelis parameter, k_{cat} is the turnover number, and A is the pre-exponential factor.

$$\Delta E_{MM} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{tors} + \Delta E_{vdw} + \Delta E_{elec} \quad (2)$$

The ΔG_{pol} term represents the polar contribution to the solvation free energy, while the ΔG_{np} term stands for the non-polar contribution. The ΔS term represents the conformational entropy, which can be estimated by normal-mode analyses. The conformational entropy contribution to the binding free energy is controversial[45–47] and this term was neglected to reduce computational cost. In this work, the binding free energies were calculated using the single trajectory approach based on the MD simulation of the protein–ligand complex, as this computational strategy can reduce noise and errors in simulations.[28] Therefore, the internal energy terms (ΔE_{bond}, ΔE_{angle}, ΔE_{tors}) in eqn (2) cancel and the gas-phase interaction energy (ΔE_{MM}) between the enzyme and the substrate is the sum of van der Waals (ΔE_{vdw}) and electrostatic (ΔE_{elec}) interaction energies. All analyses by MM/PBSA and MM/GBSA were implemented using the MMPBSA.py[48] python script in the AmberTools17 package. The analyses were carried out on 100 evenly spaced snapshots extracted from the production MD run trajectory between 10 and 20 ns, where the trajectory files from Desmond were transformed into Amber coordinate files *via* VMD software.[49] In MM/PBSA, the ΔG_{pol} term was calculated by solving the linearized Possion–Boltzmann equation; the default finite-difference PB solver with default parameters was adopted. In MM/GBSA, the ΔG_{pol} term was calculated by the modified GB models developed by Onufriev[50] and his colleagues, where the atomic radii (igb = 5) were chosen. In both PB and GB calculations, the salt molar concentration was set at 0.1 M in solution. The ΔG_{np} term was determined based on the solvent accessible surface area (SASA): ΔG_{np} = surften × ΔSASA, where the default parameter (surften = 0.0072 kcal mol^{-1} Å^{-2}) was adopted. In this work, explicit waters around the substrate were considered in the MM/PBSA and MM/GBSA analyses, and the corresponding MD trajectories were obtained using the Amber module *cpptraj*[51] with the keyword "*closest*", which retains the requested number (N_{wat}) of water molecules that are closest to the substrate.

Results and discussion

Validation of the MD simulation

The rapid MD simulation protocols introduced in the preceding section were verified by testing the stability of the wild-type (WT) cocaine esterase (cocE). The ligand and the water molecules in the crystal structure (PDB ID: 1JU3) of cocE were removed. The RMSD of all backbone atoms and side chain atoms of cocE relative to their initial crystal structural positions were monitored over the course of 20 ns MD simulations, and the resulting plots of RMSDs are shown in Fig. 3A. After 10 ns, the observed RMSDs for backbone and side chain atoms lie within a low and narrow range of approximately 1.25–1.75 Å and 2.0–2.5 Å, respectively. The time averaged RMSD for each residue of cocE is shown in Fig. 3B, where the five catalytic residues are indicated by red lines. Ser117 is the most flexible residue in the active site with a RMSD value of 1.82 Å. Ser117 plays the nucleophilic attacking role in the catalytic machinery and its structural flexibility is consistent with this function. All other catalytic residues, *i.e.*, His287 (1.00 Å), Asp259 (1.37 Å), Tyr44 (1.80 Å) and Tyr118 (0.96 Å), displayed very low RMSDs as they contribute hydrogen bonds to stabilize either the catalytic triad or the oxyanion. The whole structure and the active site geometries were well maintained during the simulation. Moreover, the system volume kept stable during the 20 ns MD simulations (Fig. S2†). We concluded that the MD simulation protocols recapitulate the crystal structure of cocE and this protocol should behave equally well in the presence of substrates.

The explicit waters around the substrate were considered in the binding free energy calculations by MM/PBSA or MM/GBSA. In addition, explicit water molecules that surround the substrate were considered in this study because these water molecules may influence energetic parameters. Computations were performed on the WT cocE–cephradine complex to probe the influence of the number of explicit water molecules on the

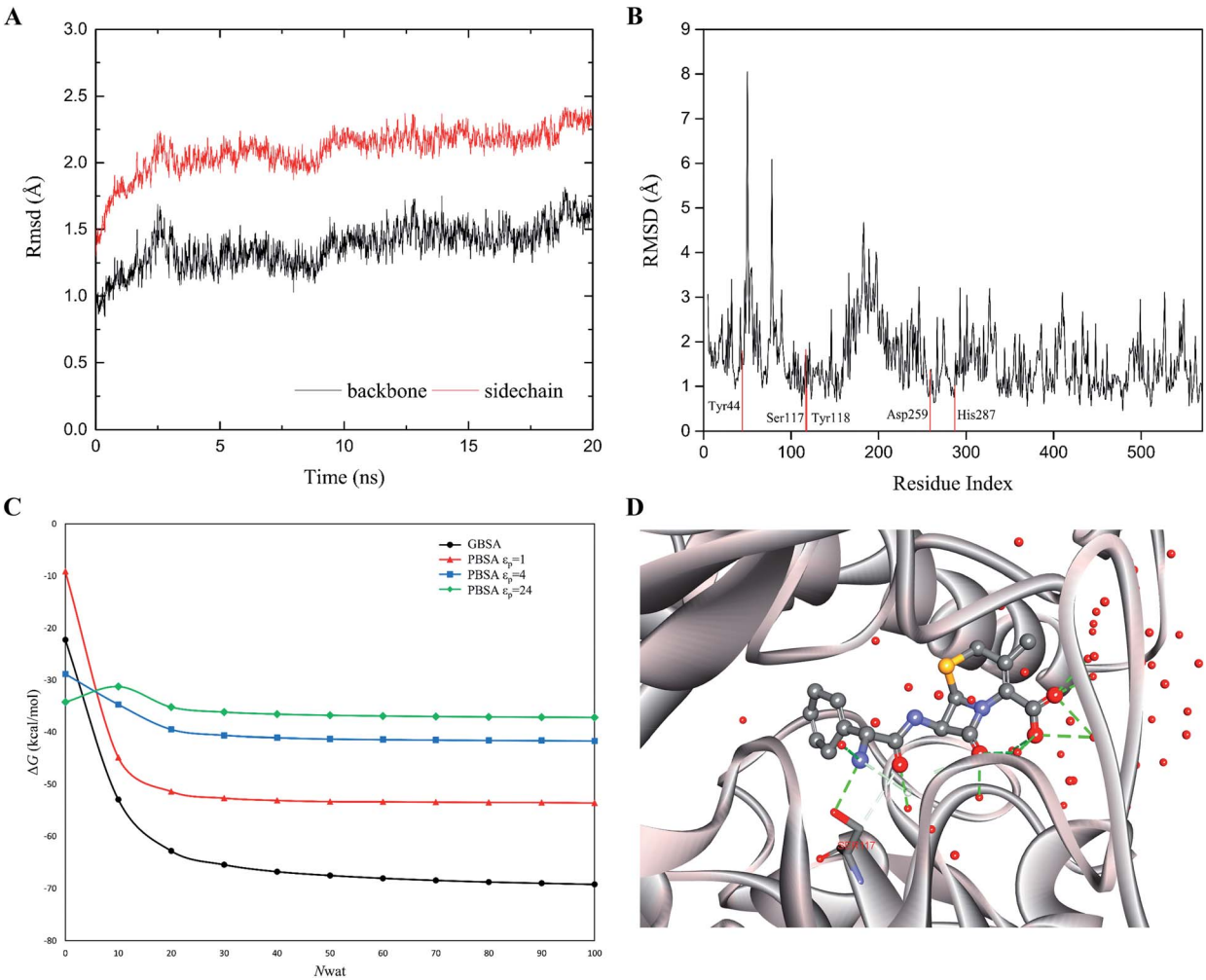

Fig. 3 (A) The RMS deviations between the crystal structure and the MD snapshots for wild type cocE during 20 ns. The RMSDs of backbone atoms are represented by black line, while the RMSDs of sidechain atoms are shown in red line. (B) The 20 ns averaged RMSDs of heavy atoms are shown in per-residue form. Five catalytic residues Tyr44, Ser117, Tyr118, Asp259, and His287 are labeled in red. (C) The effect of number of explicit waters (N_{wat}) considered in MM/GBSA and MM/PBSA methods on the predicted binding free energies of WT. The black line indicates the results from the MM/GBSA method, while the red, blue, and green lines for results from the MM/PBSA methods based on $\varepsilon_p = 1$, 4, or 24. (D) The active site geometry with 50 explicit water molecules. The protein secondary structure is shown in silver ribbon. The cefradine is shown in ball-and-stick model. The hydrogen bonds are represented by green dashed lines. Oxygen atoms are colored in red, nitrogen atoms in cyan, and carbon atoms in grey.

calculated binding free energies. The calculated results of the binding free energies by MM/PBSA and MM/GBSA against the number of explicit waters from 0 to 100 with increments of 10 are presented in Fig. 3C. The calculated binding free energies decreased drastically when considering only a few explicit water molecules in the MM/PBSA and MM/GBSA methods. This result may be due to the closest water molecules forming hydrogen bonds with the substrate (Fig. 3D). The calculated binding free energies were essentially constant once the number of explicit water molecules (N_{wat}) increased above 30. Since changes in the electrostatic environment of the active site of cocE variants may change, 50 explicit waters were considered in the following binding free energy calculations by both MM/PBSA and MM/GBSA methods.

Ranking variants by calculated binding free energy

Wild-type cocE and its variants belong to the family of serine proteases. For the amide hydrolysis reaction that is catalyzed by serine proteases,[52] the acylation process is always the rate-limiting step when compared with that of the deacylation process. Moreover, WT cocE and its variants are poor enzyme catalysts towards the hydrolysis of cephradine,[30] indicating that the kinetics of cephradine hydrolysis is reaction-controlled rather than diffusion-controlled. This can be explained that the scaffold of the esterase cocE and its variants lacks of suitable hydrogen bonding acceptor to stabilize the leaving amino group of cephradine in the transition state for proton shuttling with His287 and results in low amidase activity.[54,55] Therefore, the Michaelis–Menten equation for cephradine hydrolysis can be greatly simplified and the Michaelis constant, K_m, can be

approximated by the dissociation constant K_s of the enzyme–substrate complex. The binding free energy ΔG_{bind} corresponds to the Gibbs free energy change between the enzyme–substrate bound form (ES state in Fig. 2B) and the unbound form (E + S state in Fig. 2B). Based on the MD simulations, the binding free energies (ΔG_{bind}) of WT cocE and variants were calculated by the MM/PBSA and MM/GBSA methods. Importantly, the MM/PBSA and MM/GBSA methods are used here to rank the binding free energies of the variants rather than give accurate predictions of the absolute binding free energies.

The calculated binding free energies (ΔG_{bind}) were semi-log plotted against the experimentally measured K_m values for the cocE WT and variants in Fig. 4A for the MM/GBSA method and Fig. 4B–D for the MM/PBSA method based on different protein dielectric constants ε_p. The results of the statistical significance test for these correlations were shown in Table S5,† and all these fits passed the *t*-test. In Fig. 4A, WT cocE and the majority of the variants concentrated around the regression line, resulting in a correlation coefficient of $R^2 = 0.7053$. A hypothetical coordinate system with the WT protein as the origin was added to Fig. 4A, and the results of all variants, except Q6 and Q10, resided in the first and third quadrants, indicating that the MM/GBSA method can differentiate variants with that of the WT protein according to the predicted binding free energies between enzyme and the substrate. However, results for the variants lying in the first and third quadrants were widely distributed around the regression line, indicating that the MM/GBSA method cannot rank the variants well. Fig. 4B shows the predicted binding free energies calculated by the MM/PBSA method based on $\varepsilon_p = 1$, and the correlation coefficient improved (*i.e.*, $R^2 = 0.7531$) when compared with that of the MM/GBSA method. The results of all variants except Q10 in Fig. 4B resided in the first and third quadrants, and more importantly the results for these variants concentrated on the regression line, indicating that the MM/PBSA method not only can differentiate variants from that of the WT protein, but can also rank the variants. This can be attributed to the finite-difference Poisson–Boltzmann method embedded in the MM/PBSA method calculating the electrostatic component of the solvation energy more accurately than the generalized Born model. The active site of the enzyme is a typical anisotropic electrostatic environment and the dielectric constant in the active pocket should be higher than that in the hydrophobic core of the protein since it is always directly in contact with the polar solvent. In Fig. 4C and D, the binding free energies between enzymes and substrate were calculated by the MM/PBSA method based on $\varepsilon_p = 4$ and $\varepsilon_p = 24$, respectively. In Fig. 4C, the correlation coefficient $R^2 = 0.7728$ was further

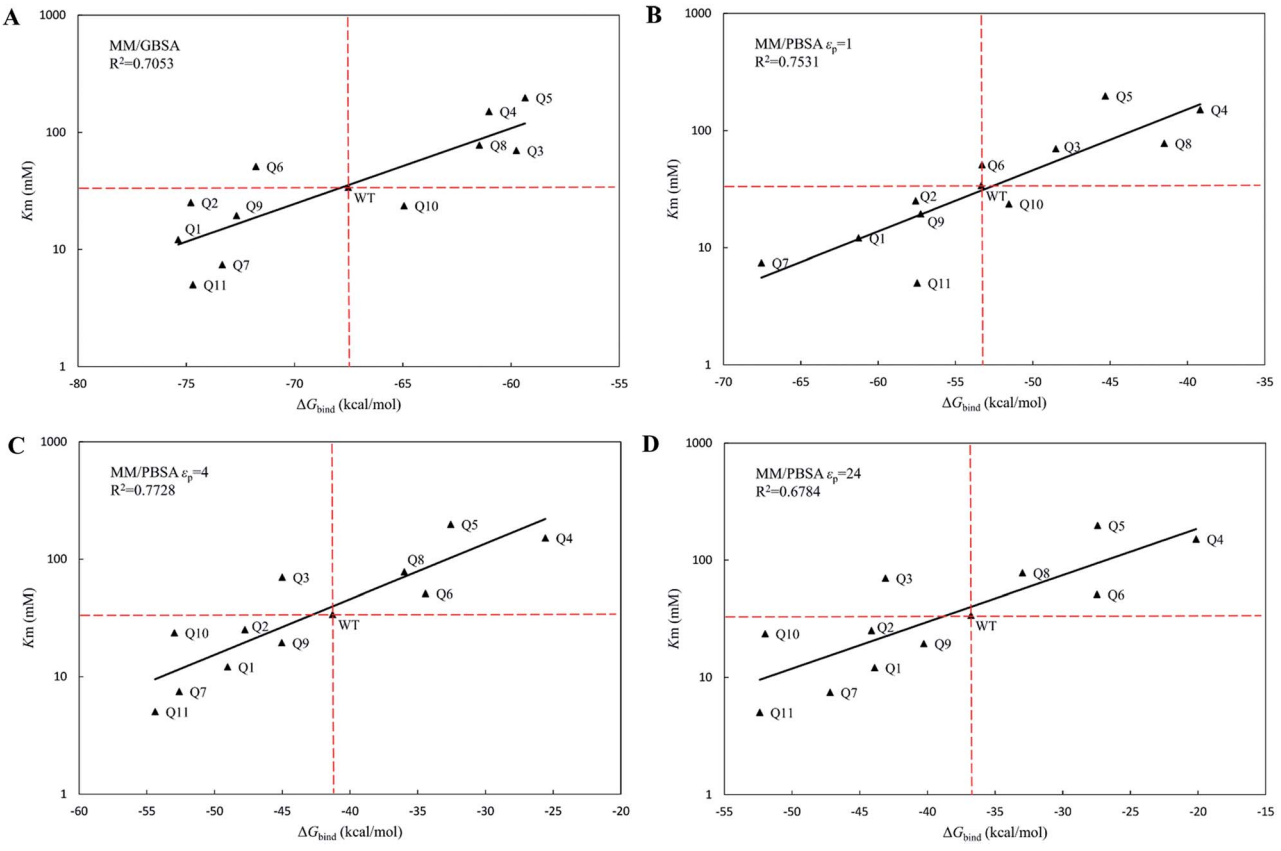

Fig. 4 The regression between the experimental Michaelis binding constants and the predicted binding free energies of WT and its eleven variants by MM/GBSA and MM/PBSA methods. The ordinate is scaled logarithmically, and the hypothetical coordination systems are shown in dashed lines with the WT as the origin. (A) MM/GBSA method; (B) MM/PBSA method based on $\varepsilon_p = 1$; (C) MM/PBSA method based on $\varepsilon_p = 4$; (D) MM/PBSA method based on $\varepsilon_p = 24$.

improved. The results for all variants except Q3 reside in the first and third quadrants. Note that variant Q6, which was a false positive prediction in Fig. 4A, moved completely into the first quadrant, and variant Q10, which was a false negative prediction in Fig. 4B, moved into the third quadrant, indicating that these two variants were predicted correctly by the MM/PBSA method based on $\varepsilon_p = 4$. Moreover, variant Q11, whose apparent binding constant was the smallest among all variants, ranked best in Fig. 4C according to the predicted binding free energies. The predicted binding free energies by the MM/PBSA method deteriorated when the ε_p was increased to 24. The correlation coefficient decreased noticeably in Fig. 4D ($R^2 = 0.6784$), indicating that selection of the appropriate protein dielectric constant is critical for the MM/PBSA method to correctly predict the binding free energies. In summary, a relatively high protein dielectric constant ($\varepsilon_p = 4$) is beneficial for ranking purposes using the MM/PBSA method, which is consistent with the suggestion from Hou and his colleagues.[53]

Ranking variants by the calculated activation energy barrier

On the reaction coordinate for enzymatic catalysis shown in Fig. 2B, the Gibbs free energy change (activation energy, $\Delta G^{\ddagger}_{cal}$) from the Michaelis binding state (ES) to the transition state ($ES^{\ddagger}$) corresponds to the turnover number (k_{cat}). The activation process involves formation of a bond between the OG atom of Ser117 and the central C15 atom of cephradine during the acylation step for the hydrolytic reaction of cephradine, and that the molecular dynamics simulation using the *ff99SB-ILDN* force field cannot handle bond breaking and formation. Therefore, we cannot use the MM/GBSA method or the MM/PBSA method to calculate the activation energies of the variants. In fact, the experimental turnover numbers (k_{cat}) and the calculated activation energies ($\Delta G^{\ddagger}_{cat}$) of WT and its eleven variants by MM/GBSA and MM/PBSA methods were poorly correlated (Fig. S1(A–D)†), and all fits did not pass the *t*-test but for MM/PBSA method based on $\varepsilon_p = 24$ (Table S5†). However, the acylated state is relatively stable when compared with that of the transition state and the bond between the OG atom of Ser117 and the central C15 atom of cephradine was formed. Therefore, the MD simulation can be run for this state. The MD simulation time (20 ns) is much longer than the actual lifetime of the intermediate state (~fs). In the reference state (E + S), the solvation processes of the isolated enzyme (E) and substrate (S) can be performed by two individual MD simulations. Using the trajectories of the three individual MD simulations for E, S and $ES^{\ddagger}$, the free energy change ($\Delta G^{\ddagger}$) between the transition state

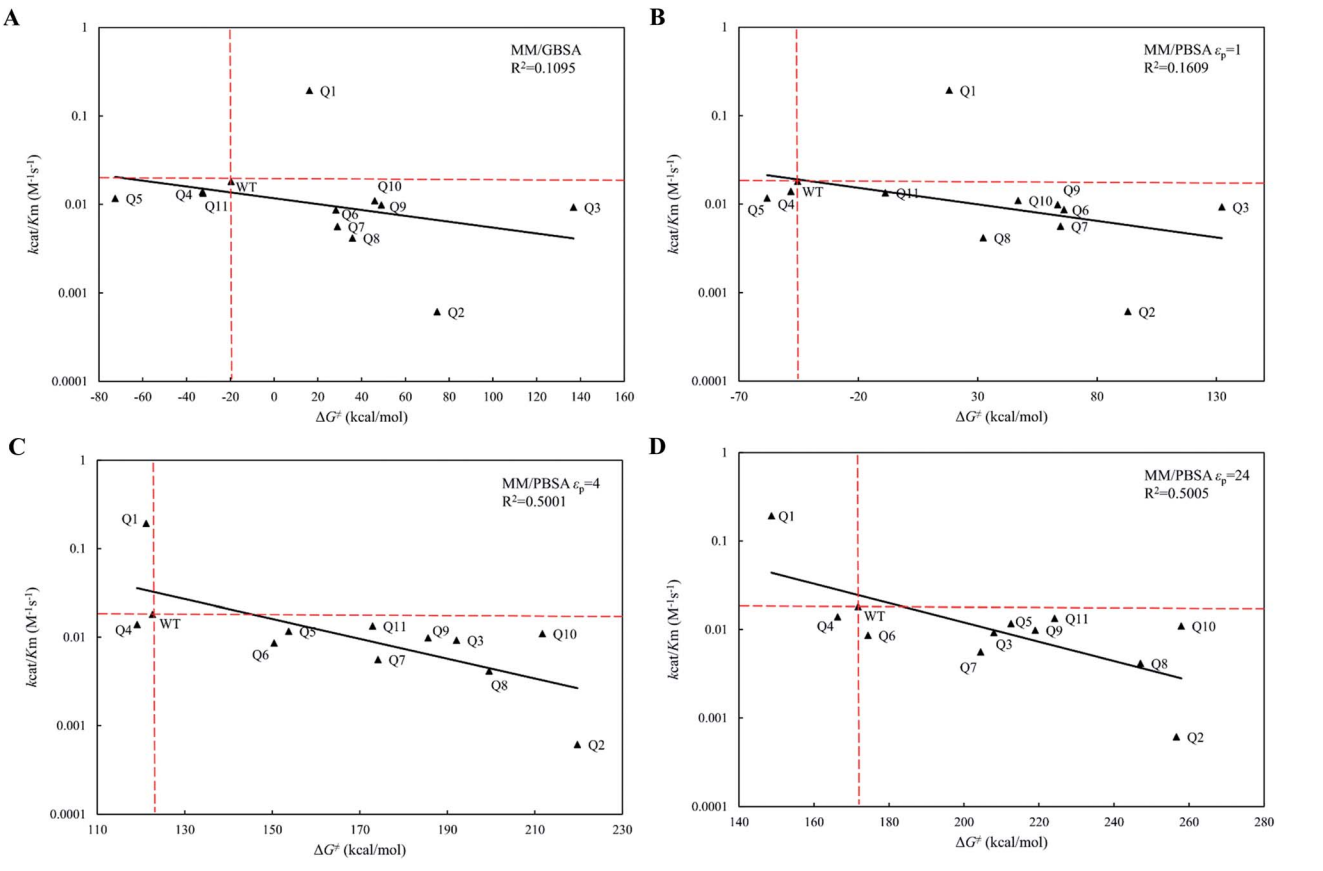

Fig. 5 The regression between the experimental catalytic efficiencies and the predicted activation energy barries of WT and its eleven variants by MM/GBSA and MM/PBSA methods. The ordinate is scaled logarithmically, and the hypothetical coordination systems are shown in dashed lines with the WT as the origin. (A) MM/GBSA method; (B) MM/PBSA method based on $\varepsilon_p = 1$; (C) MM/PBSA method based on $\varepsilon_p = 4$; (D) MM/PBSA method based on $\varepsilon_p = 24$.

Table 1 The kinetic parameters of the cephradine hydrolase and its variants predicted by PRODA. The data is taken from the ESI of the ref. 30. Reaction conditions are: pH 7.0 at 22 °C

Variants	Mutations	k_{cat} ($\times 10^{-3}$ s^{-1})	K_m (mM)	k_{cat}/K_m ($\times 10^{-3}$ M^{-1} s^{-1})
WT	—	0.61	33.90	18.11
Q1	F261T	2.33	12.10	192.72
Q2	Q55R	0.02	25.16	0.61
Q3	F408L	0.65	70.03	9.26
Q4	P150E	2.11	151.34	13.93
Q5	F261T/F408L	2.32	198.12	11.68
Q6	F261T/P150E	0.44	51.16	8.63
Q7	F261T/L407Q	0.04	7.46	5.61
Q8	F261T/L407H	0.32	77.98	4.16
Q9	F261T/F408Y	0.19	19.47	9.84
Q10	F261T/L407Q/F408Y	0.26	23.68	10.99
Q11	F261T/L407H/F408Y	0.07	5.02	13.41

($ES^{\ddagger}$) and the reference state (E + S) can be calculated by the MM/GBSA method or the MM/PBSA method. As this free energy change ($\Delta G^{\ddagger}$) corresponds to the catalytic efficiency (k_{cat}/K_m) of the enzyme, we can use the calculated $\Delta G^{\ddagger}$ to evaluate the variants. Note that the covalent bond energy between the enzyme and substrate was not included in the calculated free energy change ($\Delta G^{\ddagger}$) and was cancelled during the evaluation of the variants. To obtain convergent results in the triple trajectories, water molecules within a 10 Å range to the OG atom of Ser117 were explicitly considered for the states (E) and ($ES^{\ddagger}$) during the calculation of the solvation energy, whereas the implicit method was used for the S state.

The calculated free energy changes ($\Delta G^{\ddagger}$) are semi-log plotted against the experimentally measured catalytic efficiencies (k_{cat}/K_m) for WT cocE and variants in Fig. 5A for the MM/GBSA method and for the MM/PBSA method (Fig. 5B–D) based on different ε_p. In Fig. 5A, the MM/GBSA method showed a very poor correlation coefficient ($R^2 = 0.1095$). The variants Q4, Q5 and Q11 appeared to be false positives because they showed even lower energy barriers than the WT enzyme even though their catalytic efficiencies were lower. Moreover, the best variant Q1 showed a ten-fold higher catalytic efficiency (k_{cat}/K_m) than the WT protein (Table 1), but this was a false negative because the predicted energy barrier of the variant was higher than that of the WT enzyme. In Fig. 5B, the MM/PBSA method based on $\varepsilon_p = 1$ yielded similar performances with only an improvement observed for variant Q10, *i.e.*, the result was no longer in the false positive quadrant. These results indicate that the protein dielectric constant might be set too low ($\varepsilon_p = 1$) to correctly reflect the electrostatic environment in the active site of the enzyme. Moreover, the correlation between k_{cat}/K_m and $\Delta G^{\ddagger}$ for MM/GBSA method and MM/PBSA method based on $\varepsilon_p = 1$ did not pass the *t*-test (Table S5†). In Fig. 5C and D, the ε_p was increased to 4 and 24, respectively. In Fig. 5C, the MM/PBSA method based on $\varepsilon_p = 4$ showed a better correlation coefficient ($R^2 = 0.5001$) and most variants were distributed along the regression line. The best variant, Q1, ranked second among all variants and its relative position to that of the WT protein was predicted correctly, although the free energy gap was small. The worst variant Q2 was predicted to have the highest energy barrier, which is also consistent with its experimental catalytic efficiency. The variant Q4 was the only false positive prediction and no false negative prediction was observed. When the ε_p was increased to 24 in Fig. 5D, the best variant Q1 ranked top and its free energy gap relative to the WT enzyme increased when compared with the results in Fig. 5C. Although the correlation coefficient ($R^2 = 0.5005$) did not improve, the better prediction results showed that the evaluation performance of the MM/PBSA method could be greatly improved if higher protein dielectric constants are selected in the active site. All eleven variants were predicted by PRODA to have higher catalytic efficiencies when compared with that of the WT enzyme, but the results shown in Table 1 indicate that all variants except Q1 were false positive predictions. Based on MD simulations, the energy barriers shown in Fig. 5D, which were calculated by the MM/PBSA method under a suitable ε_p (*i.e.*, 24), can help to eliminate all false positive variants except Q4 before experiments were recorded. Therefore, MD based simulations can greatly improve the accuracy of energy-based sequence prediction tools, such as PRODA, for computational enzyme design.

Conclusions

In this paper, the computationally designed variants of a *Rhodococcus* sp. cocaine esterase for the hydrolysis of cephradine were assessed quantitatively by MM/GBSA and MM/PBSA methods based on MD simulations. The apparent K_m of the variants correlated well with the binding free energies calculated by the MM/GBSA and MM/PBSA methods with explicit waters around the substrate considered. The catalytic efficiencies k_{cat}/K_m of the variants correlated poorly with the activation energy barriers calculated by the MM/GBSA and MM/PBSA methods when a low protein dielectric constant ($\varepsilon_p = 1$) was used. In contrast, when $\varepsilon_p = 4$ or 24 for the enzyme, the activation energy barriers calculated by the MM/PBSA method correlated well with the catalytic efficiencies (k_{cat}/K_m) of the variants, and most false positive predictions obtained by the computational enzyme design program PRODA could be eliminated prior to experimental analysis. Although the methods developed here need to be tested on more systems, we showed

that MD-based quantitative assessment of cocE variants provided an effective and efficient screening tool for energy-based computational enzyme design.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

We acknowledge the financial support from the National Natural Science Foundation of China (Grant Numbers: 21476123 and 21878170) and the Ministry of Science and Technology of China (Grant Number: 2012AA021204).

References

1 U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore and K. Robins, *Nature*, 2012, **485**, 185–194.
2 A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts and B. Witholt, *Nature*, 2001, **409**, 258–268.
3 F. H. Arnold, *Nature*, 2001, **409**, 253–257.
4 D. J. Tantillo, C. Jiangang and K. N. Houk, *Curr. Opin. Chem. Biol.*, 1998, **2**, 743–750.
5 A. Zanghellini, L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Röthlisberger and D. Baker, *Protein Sci.*, 2006, **15**, 2785–2794.
6 D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik and D. Baker, *Nature*, 2008, **453**, 190–195.
7 L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard and D. Baker, *Science*, 2008, **319**, 1387–1391.
8 J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael and D. Baker, *Science*, 2010, **329**, 309–313.
9 S. R. Gordon, E. J. Stanley, S. Wolf, A. Toland, S. J. Wu, D. Hadidi, J. H. Mills, D. Baker, I. S. Pultz and J. B. Siegel, *J. Am. Chem. Soc.*, 2012, **134**, 20513–20520.
10 S. D. Khare, Y. Kipnis, P. J. Greisen, R. Takeuchi, Y. Ashani, M. Goldsmith, Y. Song, J. L. Gallaher, I. Silman, H. Leader, J. L. Sussman, B. L. Stoddard, D. S. Tawfik and D. Baker, *Nat. Chem. Biol.*, 2012, **8**, 294–300.
11 J. B. Siegel, A. L. Smith, S. Poust, A. J. Wargacki, A. Bar-Even, C. Louw, B. W. Shen, C. B. Eiben, H. M. Tran, E. Noor, J. L. Gallaher, J. Bale, Y. Yoshikuni, M. H. Gelb, J. D. Keasling, B. L. Stoddard, M. E. Lidstrom and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 3704–3709.
12 H. J. Wijma, R. J. Floor, S. Bjelic, S. J. Marrink, D. Baker and D. B. Janssen, *Angew. Chem., Int. Ed.*, 2015, **54**, 3726–3730.
13 M. J. Grisewood, N. J. Hernández-Lozada, J. B. Thoden, N. P. Gifford, D. Mendez-Perez, H. A. Schoenberger, M. F. Allan, M. E. Floy, R.-Y. Lai, H. M. Holden, B. F. Pfleger and C. D. Maranas, *ACS Catal.*, 2017, **7**, 3837–3849.
14 R. Li, H. J. Wijma, L. Song, Y. Cui, M. Otzen, Y. Tian, J. Du, T. Li, D. Niu, Y. Chen, J. Feng, J. Han, H. Chen, Y. Tao, D. B. Janssen and B. Wu, *Nat. Chem. Biol.*, 2018, **14**, 664–670.
15 A. N. Alexandrova, D. Röthlisberger, D. Baker and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2008, **130**, 15907–15915.
16 G. Kiss, D. Röthlisberger, D. Baker and K. N. Houk, *Protein Sci.*, 2010, **19**, 1760–1773.
17 M. P. Frushicheva, J. Cao, Z. T. Chu and A. Warshel, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 16869–16874.
18 S. Osuna, G. Jiménez-Osés, E. L. Noey and K. N. Houk, *Acc. Chem. Res.*, 2015, **48**, 1080–1089.
19 H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk and S. L. Mayo, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 3790–3795.
20 L. Mollica, G. Conti, L. Pollegioni, A. Cavalli and E. Rosini, *J. Chem. Inf. Model.*, 2015, **55**, 2227–2241.
21 G. Jiménez-Osés, S. Osuna, X. Gao, M. R. Sawaya, L. Gilson, S. J. Collier, G. W. Huisman, T. O. Yeates, Y. Tang and K. N. Houk, *Nat. Chem. Biol.*, 2014, **10**, 431–436.
22 Q. Li, X. Huang and Y. Zhu, *J. Mol. Model.*, 2014, **20**, 2314.
23 J. Åqvist, V. B. Luzhkov and B. O. Brandsdal, *Acc. Chem. Res.*, 2002, **35**, 358–365.
24 D. A. Pearlman and P. S. Charifson, *J. Med. Chem.*, 2001, **44**, 3417–3423.
25 D. J. Price and W. L. Jorgensen, *J. Comput.-Aided Mol. Des.*, 2001, **15**, 681–695.
26 I. Massova and P. A. Kollman, *Perspect. Drug Discovery Des.*, 2000, **18**, 113–135.
27 P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham, *Acc. Chem. Res.*, 2000, **33**, 889–897.
28 T. Hou, J. Wang, Y. Li and W. Wang, *J. Chem. Inf. Model.*, 2011, **51**, 69–82.
29 S. Genheden and U. Ryde, *Expert Opin. Drug Discovery*, 2015, **10**, 449–461.
30 X. Huang, J. Xue and Y. Zhu, *Chem. Commun.*, 2017, **53**, 7604–7607.
31 Y. Zhu, *Ind. Eng. Chem. Res.*, 2007, **46**, 839–845.
32 Y. Lei, W. Luo and Y. Zhu, *Protein Sci.*, 2011, **20**, 1566–1575.
33 X. Huang, K. Han and Y. Zhu, *Protein Sci.*, 2013, **22**, 929–941.
34 J. He, X. Huang, J. Xue and Y. Zhu, *Green Chem.*, 2018, **20**, 5484–5490.
35 Z. Xiang and B. Honig, *J. Mol. Biol.*, 2001, **311**, 421–430.
36 K. J. Bowers, D. E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan and D. E. Shaw, in *Proceedings of the ACM/IEEE SC 2006 Conference*, 2006, pp. 43–43.
37 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.
38 C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280.

39 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision A.2*, 2009.

40 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950–1958.

41 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

42 V. Kräutler, W. F. van Gunsteren and P. H. Hünenberger, *J. Comput. Chem.*, 2001, **22**, 501–508.

43 Y. Shan, J. L. Klepeis, M. P. Eastwood, R. O. Dror and D. E. Shaw, *J. Chem. Phys.*, 2005, **122**, 54101.

44 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

45 T. Yang, J. C. Wu, C. Yan, Y. Wang, R. Luo, M. B. Gonzales, K. N. Dalby and P. Ren, *Proteins*, 2011, **79**, 1940–1951.

46 H. G. Wallnoefer, K. R. Liedl and T. Fox, *J. Comput. Chem.*, 2011, **32**, 1743–1752.

47 A. Weis, K. Katebzadeh, P. Söderhjelm, I. Nilsson and U. Ryde, *J. Med. Chem.*, 2006, **49**, 6596–6606.

48 B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, *J. Chem. Theory Comput.*, 2012, **8**, 3314–3321.

49 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38.

50 A. Onufriev, D. Bashford and D. A. Case, *Proteins*, 2004, **55**, 383–394.

51 D. R. Roe and T. E. Cheatham, *J. Chem. Theory Comput.*, 2013, **9**, 3084–3095.

52 L. Hedstrom, *Chem. Rev.*, 2002, **102**, 4501–4524.

53 H. Sun, Y. Li, M. Shen, S. Tian, L. Xu, P. Pan, Y. Guan and T. Hou, *Phys. Chem. Chem. Phys.*, 2014, **16**, 22035–22045.

54 P.-O. Syrén and K. Hult, *ChemCatChem*, 2011, **3**, 853–860.

55 P.-O. Syrén, *FEBS J.*, 2013, **280**, 3069–3083.