NJC

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *New J. Chem.*, 2024, **48**, 591

Enhanced thermostability of *Streptomyces mobaraensis* transglutaminase *via* computation-aided site-directed mutations and structural analysis†

Yongzhen Li,[a] Banghao Wu,[a] Yumeng Zhang,[a] Lanxuan Liu,[b] Linquan Bai [a] and Ting Shi *[a]

Streptomyces mobaraensis transglutaminase (smTG) has been widely used in the food processing industry for protein crosslinking. However, its poor thermostability becomes a major obstacle to further applications. It is significant to develop a feasible strategy to improve the thermostability of smTG. Here, we developed a computational model based on the Siamese graph neural network framework to identify residues critical to the thermostability of smTG by predicting changes in folding free energy ($\Delta\Delta G$) between the wild type and mutants. Four candidate residues were selected for mutation experiments, and the single mutant H44C exhibited 2.7 U mg^{-1} residual enzyme activity after 10 min incubation at 60 °C, twice as much as the wild type. Mutants Q74F and E87W also exhibited 3- and 2.3-times greater activity after 20 min. Molecular dynamics (MD) simulations revealed that smTG mutants (H44C, Q74F, and E87W) improved thermostability by enhancing hydrogen bond interactions, increasing additional residue interactions, and reducing loop flexibility. The MM-GBSA calculation demonstrated that mutants H44C and Q74F enhanced binding affinity with the substrate, and six residues crucial for substrate binding were identified. This study combines computational analysis with mutation assays for the rational design of smTG and offers a facile and efficient strategy to understand and improve the thermostability of proteins for industrial applications.

Received 10th October 2023,
Accepted 12th November 2023

DOI: 10.1039/d3nj04708c

rsc.li/njc

1 Introduction

Transglutaminases (protein-glutamine γ-glutamyltransferase, TGase, EC 2.3.2.13) are widespread in nature and have been discovered in vertebrates, invertebrates, microorganisms, and plants.[1] TGases generally catalyze an acyl-transfer reaction between amino groups and glutamine residues. As cross-linking and protein-modifying enzymes, TGases have garnered significant interest due to their versatility in industrial applications by the formation of ε-(γ-glutamyl) isopeptide bonds from glutamine and lysine residues in proteins.[2]

Due to the limitations of high costs and low yields derived from animal and plant TGases, industries have gradually turned their attention to microbial TGases.[3] Compared to mammalian TGases, microbial TGases show promising and valuable properties such as a higher catalytic rate, smaller size, broader substrate specificity, and lower deamidation activity.[4] Among them, TGase generated from *Streptomyces mobaraensis* (smTG) has emerged as the most important source of commercial TGase, which is the most widely utilized biocatalyst in food products such as milk, beef, noodles, and bread.[5–7] However, food processing requires smTG to maintain high reactivity at temperatures of 50–60 °C. The strict requirement for smTG impels further exploration of its thermo stability. Therefore, how to improve the thermostability of smTG has become a critical and urgent problem.

Engineering modifications to improve the thermostability of smTG were attempted after its successful heterologous expression and activation.[8] Marx *et al.* successfully generated a library of mutants through random mutagenesis, coupled with targeted single amino acid substitutions. Among the vast pool of 5500 mutants, the researchers identified six variants that exhibited desired characteristics.[9] By using DNA shuffling and saturating mutagenesis, the half-life of the optimal smTG at 60 °C was extended.[10] Subsequently, rational design has also been applied in smTG since the crystal structures of TGs were resolved.[11] Yokoyama *et al.* improved the T_m value of smTG by

[a] *State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: tshi@sjtu.edu.cn*
[b] *Digital Innovation of AI, WuXi Biologics, Shanghai 201400, China*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3nj04708c

artificially constructing disulfide bridges.[12] Besides, more hot spots were discovered by combined mutation, which were found to improve the thermostability of smTG after testing for residual enzyme activity with the standard substrate CBZ-Gln-Gly.[13]

Compared to the time-consuming traditional experimental methods, rational design is a suitable alternative to change the function or improve the activity of proteins after mutation. Especially, the change in thermostability upon site-directed mutations turns out to be one of the foundational knowledges for enzyme or drug design.[14] On the other hand, the Gibbs free energy ΔG from the unfolding state to the folding state can quantify the thermodynamic stability of the protein.[15] The difference in Gibbs free energy ($\Delta\Delta G$) between the wild-type and mutant structures becomes a critical indicator for the impact of site-directed mutations on protein stability and function. Therefore, a robust computational strategy leveraging structural information is essential for predicting protein thermostability accurately and thus providing reliable criteria for rational design. Researchers have developed several bioinformatics approaches to predict $\Delta\Delta G$ upon mutations, including FoldX,[16] Rosetta_ddg,[17] and consensus analysis algorithms. However, the neglect of the anti-symmetry of values between mutants and native proteins usually causes prediction biases,[18,19] while the sequence similarity between training and test data possibly results in over-optimistic prediction performance. Since geometric deep learning has achieved great success in predicting protein binding affinity and protein–protein interactions,[20,21] it is necessary to develop a new effective model for $\Delta\Delta G$ prediction by integrating the network topology of residue interaction from structures with an artificial intelligence model.

In this study, we combined computation-aided design with mutation experiment to discover engineering smTG with improved thermostability. First, a Siamese graph neural network (GNN) model[22] was developed to detect the key residues that might affect the thermostability by calculating the $\Delta\Delta G$ values between wild-type and mutant smTG. Through the screening of the prediction results, four candidate residues (H44, E54, Q74, and E87) were mined. Then, the residual enzyme activity experiments demonstrated that three mutants (H44C, Q74F, and E87W) significantly improved the thermostability of smTG at 60 °C. Finally, molecular dynamics simulations were carried out to explain the structural basis of the mutants. With the improved thermostability, the engineered smTG has the potential to extend its applications in the modern food industry. What's more, this work provides a feasible strategy to understand and improve the thermostability of proteins for extensive applications.

2 Materials and methods

2.1 Data collection and pre-processing

The training dataset was derived from the curated records with experimental $\Delta\Delta G$ values in FireprotDB,[23] while the testing data consisted of S^{sym} (342 mutations and their reverse of 13 proteins),[18] p53 (42 variants from P53 structure 2OCJ),[24] and myoglobin (134 variants from myoglobin structure 1BZ6) datasets.[25] After processing replicated mutations (taking average values) and removing homologous proteins against the test data (BLASTp e-value ≤ 0.001), there were 2640 non-redundant records (from 140 unique proteins) in the training data. Ten-fold cross-validation was performed on the training set, namely using 2376 thermostability records for training, and the remaining 264 records in the validation set were evaluated. PDBrenum[26] was used to align mutation positions in the database to those in the PDB structures.

2.2 Refinement on protein structures

The model required a pair of protein structures (wild type and mutant) as inputs. The wild-type protein structures were obtained from the Protein Data Bank (PDB)[27] and relaxed using the Rosetta FastRelax protocol[28] with the Rosetta all-atom energy function REF2015.[29] The mutant structure was generated from the refined wild-type structure using the same protocol by Rosetta. Atom positions were constrained to avoid significant conformational changes.

2.3 Construction of the residue interaction graph

The residue interaction graph depicted the interaction network between residues around the mutation site. The Euclidean distances from the C_α atom of all residues to the C_α atom at the mutation site were measured. Residues within a neighborhood radius of 12.0 Å were considered as nodes in the mutation neighborhood graph. If the distance between two nodes was below a contact threshold of 5.0 Å, they were connected by an edge. In more than half of the cases, the network topology of the mutant protein's residue interaction graph differed from that of the wild-type protein.

2.4 Node features in the residue interaction graph

Node features represented prior knowledge about the corresponding amino acid residues before training. Descriptors for residues were divided into three categories: (a) amino acid encoding, including skip-gram model representation, one-hot vector encoding dipole scale and volume scale, and an 8-dimensional vector summarizing hydrophobicity, charge, secondary structure, and solvent accessibility.[30] (b) Energy encoding, a 20-dimensional representation obtained from Rosetta scoring functions, incorporating both physics-based and knowledge-based energy terms. (c) Evolutionary encoding, a 20-dimensional representation derived from multiple sequence alignment against the Uniclust_30 database using *hhblits*.[31,32] These descriptors were concatenated to form initial node embeddings in the graph.

2.5 Deep learning model architecture

The deep learning model consisted of a Siamese neural network based on a graph attention network (GAT)[22] and a multi-layer perceptron (MLP) for regression. It utilized two graph attention layers to aggregate biological information from adjacent nodes and update the node embeddings (300-dimensions). The graph

attention operator aggregated the information from neighbouring nodes using attention coefficients. A dropout layer and global mean pooling layer were employed to avoid overfitting and produce the graphical representations for the wild-type and mutant structure, respectively. The final representation vector was obtained by combining the node embedding vector for the residue at the mutation site with the pooled output. Both wild-type and mutant graph representations were then fed into the MLP regressor to predict $\Delta\Delta G$.

2.6 Model training and performance assessment

The model was trained using ten-fold cross-validation. Nine subsets of the training dataset were used for training, while the remaining subset was used for monitoring model generalizability. This process was repeated ten times, and the output of the ten models was combined to predict $\Delta\Delta G$ for the test data. Log-cosh loss was used as the objective function of the regression task. The Adam optimizer with a learning rate of 0.001 and weight decay of 0.0005 was employed to train the model for 100 epochs. A dropout ratio of 0.2 was applied after the graph attention layers and fully-connected layer to address overfitting. A batch size of 256 was chosen for efficiency and convergence. Warm-up learning rate (10 epochs) and a cosine annealing schedule were used for stability and convergence. Early stopping with a patience of 10 epochs was employed based on the mean square error (MSE) on the validation subset. The model was implemented using PyTorch 1.9.0 and PyTorch Geometric 2.0.2.[33] Pearson correlation coefficient and root mean square error (RMSE) were estimated between the experimental and predicted $\Delta\Delta G$ values to assess model performance. Performance metrics were recorded for both direct and reverse mutations. The average of $\delta = \Delta\Delta G_{dir} + \Delta\Delta G_{rev}$ was also used to quantify the prediction bias.[34]

2.7 Construction of plasmids expressing TG wild-type (smTG) and its mutants

The plasmid pET-28a (+) and *Escherichia coli* BL21(DE3) pGro7 were used for expressing smTG and its mutants. The sequence of smTG was the mature region of *Streptomyces mobaraensis* TGase, which removed the nucleotide sequence of signal peptide and FRAP (Fig. S1, ESI†).[35] The gene fragment smTG was cloned from the genome of *Streptomyces mobaraensis* (NCBI reference sequence: NZ_CP083590.1) using the primer MTGase-F/R (Table S1, ESI†) and cloned into *Nde*I-*Eco*RI sites of pET-28a (+), yielding the smTG expression plasmid pET-28a/smTG. Meanwhile, a His-tag has been placed at the N-terminus of smTG for facilitating the affinity purification on a nickel column (Fig. S2, ESI†). The plasmids encoding the smTG with one amino acid substitution (E54K, Q74Y, H44C, H44I, E87F, E87W, E54R, or Q74F) were constructed by the PCR-based method. PCR procedures were performed using pET-28a/smTG as a template, and the primer pair for each mutation is listed in Table S1 (ESI†). The purified PCR products were incubated with *Dpn*I and T4 DNA ligase, and then introduced into BL21(DE3) pGro7. Each mutation was verified by DNA sequencing.

2.8 Protein expression

The plasmid pET-28a/smTG and its derivatives carrying different mutations were transformed to *E. coli* BL21 (DE3) pGro7. A single colony of each recombinant *E. coli* strain was inoculated into 50 mL LB medium with 50 μg mL^{-1} kanamycin and 34 μg mL^{-1} chloramphenicol for seed culture at 37 °C for 12–16 h. The seed culture was transferred into 1 L LB medium with 50 μg mL^{-1} kanamycin, 34 μg mL^{-1} chloramphenicol and 0.5 mg mL^{-1} L-arabinose at 37 °C 220 rpm until a OD_{600} value of 0.6–0.8 was obtained. The culture was cooled to 16 °C and induced with 0.4 mM IPTG. The culture was further cultivated at 16 °C and 220 r.p.m. for 18 h.

2.9 Protein purification

Cells were harvested from the fermentation culture by centrifugation and resuspended in Buffer A (25 mM Tris–HCl, 300 mM NaCl, pH 7.5). After ultra-sonification and centrifugation, the supernatants were subjected to affinity purification using gravity chromatography columns with Ni Sepharose 6 Fast Flow beads (Cytiva) at 4 °C. The recombinant smTG and its mutants were eluted with elution buffer (25 mM Tris–HCl, 300 mM NaCl, and 250 mM imidazole) and desalted with Buffer A (25 mM Tris–HCl, 300 mM NaCl, pH 7.5) (Fig. S3, ESI†).

2.10 Determination of the enzyme activity of mutants

The purified smTG and its mutants were diluted to 0.5 mg mL^{-1} and transferred into an Eppendorf tube for enzyme activity analysis. Proteins were incubated at 60 °C for 10, 20 and 30 min. After heat treatment, samples were subsequently chilled on ice to prevent further inactivation. The activity was measured at 37 °C using the standard assay with 200 μL preincubated solution A (214.3 mM TRIS/acetate, 107.14 mM hydroxylamine, 10.7 mM reduced glutathione, 32.1 mM CBZ-Gln-Gly-OH), starting with 10 μL enzyme solution. After 10 min at 37 °C, the reaction was stopped with 200 μL solution B (1 vol. 3 M HCl, 1 vol. 12% trichloroacetic acid, 1 vol. 5% $FeCl_3 \cdot 6H_2O$ (in 0.1 M HCl)). Then the samples were centrifuged and the supernatant was transferred to a 96-well plate. Finally, the absorption was measured at 525 nm with microplate reader.

2.11 System preparation

The high-resolution crystal structure of smTG (PDB ID: 6GMG)[36] served as the basis to model mutant H44C, Q74F and E87W. Molecular docking of protein and CBZ-Gln-Gly was conducted with AutoDock Tools (version 1.5.6). Protein protonation states were determined using the PDB2PQR web server,[37] and the protonation state of CBZ-Gln-Gly was predicted with the MolGpKa web server.[38] The preparation of protein and ligand structures involved adding appropriate charges as well as hydrogen atoms, leading to the generation of pdbqt files. Docking grid space was determined based on the binding pocket in the crystal structures. The atom-specific affinity maps, electrostatic and desolvation potentials were generated using AutoGrid. Rigid proteins and flexible ligands were combined for docking, with each protein–ligand pair

undergoing 200 iterations of docking simulations using genetic algorithm methods.[39] The docking pose selected for further analysis was determined by its best energy score and the distance between Cys64 and Gln of CBZ-Gln-Gly being less than 4.0 Å. To derive force field parameters for the substrate CBZ-Gln-Gly, geometry optimization was performed at the M062X/6-31G* level. The electrostatic surface potential (ESP) charge calculations ensued at HF/6-31G* level, facilitated by the Gaussian 09 program.[40–43] Subsequently, we employed a two-step restrained electrostatic potential (RESP) calculation method, which generated charge distribution information and crucial parameters, like bond and angle parameters. This computational procedure was executed using the Multiwfn program and the antechamber package implemented in AMBER 18.[44–47] These complexes were immersed within an octahedral box of TIP3P water molecules to mimic a solvent environment. Electroneutrality was achieved by adding suitable sodium ions *via* tleap module of AMBER 18.

2.12 Classical molecular dynamic simulations and analysis

Classical molecular dynamic (MD) simulations were conducted for these computational systems with an ff14SB force field using for the enzyme through the AMBER 18 program suite.[48] To prepare each system, a two-step energy minimization was performed for water molecules and the rest of the entire system. Subsequently, a gradual heating process was applied, raising the temperature of the systems from 0 to 333 K in 50 ps. Next, equilibration was carried out for 50 ps under the isothermal–isobaric ensemble (at 333 K and 1 bar), establishing an appropriate initial conformation for the subsequent three sets of 100 ns MD simulations per system. During MD simulations, the Particle Mesh Ewald (PME) method along with the SHAKE algorithm was employed to describe long-range electrostatic interactions and correct bond lengths.[49,50] The non-bonded interactions were considered within a cut-off distance of 10.0 Å. To analyze the MD trajectories, various parameters were computed using the cpptraj module of AMBER 18, including root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (R_g), temperature factor (*B*-factor), and hydrogen bonds.[51] For the evaluation of binding free energy between substrates CBZ-Gln-Gly and various mutants of smTG, the molecular mechanics Poisson–Boltzmann surface area (MM-GBSA) method was employed. The Python program MMPBSA.py was utilized for this purpose, enabling the decomposition of calculated free energy into specific residue contributions.[52]

3 Results

3.1 Construction of graph neural network model

Protein thermostability can be dramatically influenced by structural variation introduced by the substitution of critical amino acids. Although protein structure analysis is always beneficial in understanding target mutations, it is hard to predict unknown key residues. Herein, a Siamese graph neural network (GNN) model, as an effective approach to capture the topological information inside the protein structure, was constructed to identify single site mutation that could improve the thermostability of smTG.

The model was designed as follows: (1) given the three-dimensional (3D) structures of wild-type protein and its mutants as the input, protein 3D structure was projected onto a 2D graph maintaining the relationship between residues based on their distances in the 3D Euclidean space. (2) Each node in the graph represents a protein residue with its biological and physicochemical properties. (3) Residue interactions depended on not only spatial locations but also the side chains of the amino acids which determine the hydrogen bond

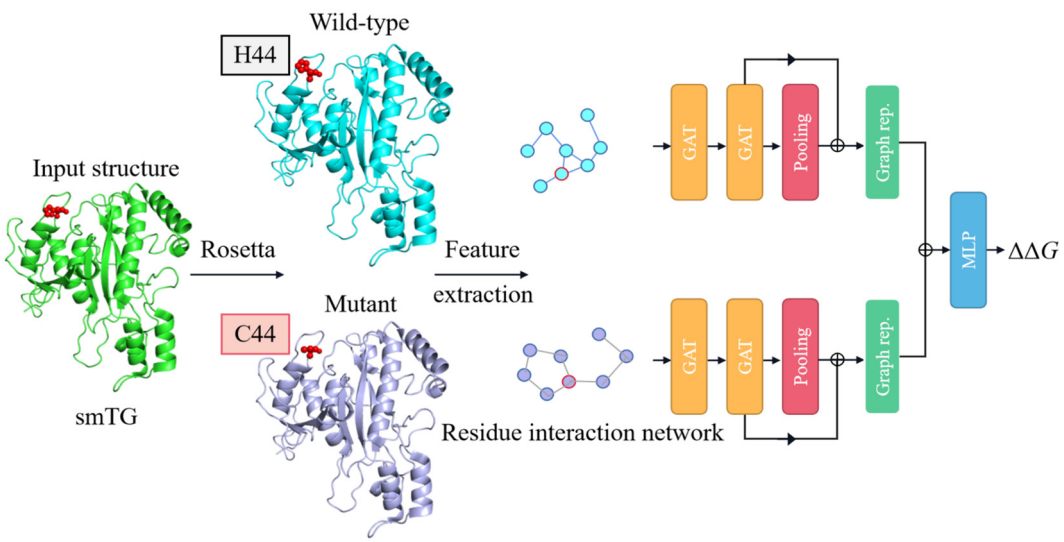

Fig. 1 The workflow of the GNN model to predict the protein thermostability. Firstly, inputting wild-type and mutant protein (*e.g.*, smTG) to Rosetta. Feature extraction was then performed by generating residue interaction networks, and the mutation site was represented by nodes with a red border. Finally, the effects of the mutation on protein thermostability were quantified using a Siamese graph attention network.

(H-bond), hydrophobic interaction and van der Waals force, so the Rosetta energy terms were included. (4) The attention mechanism was utilized to learn the intensity of residue interaction, then the node embedding was aggregated from adjacent nodes based on the attention scores. (5) A multi-layer perceptron was utilized to quantify the changes in free energy $\Delta\Delta G$ from the graph embeddings representing the wild-type and mutant structures. The workflow of the GNN model was illustrated in Fig. 1.

3.2 Quality assessment of graph neural network model

S^{sym} is a widely used dataset for assessing the predictive biases of $\Delta\Delta G$ predictors. It consists of experimentally measured $\Delta\Delta G$ values or 342 direct and corresponding reverse mutations from 15 protein chains. In the dataset, the structures of both wild-type proteins and mutants have been resolved by X-ray crystallography with a resolution of 2.5 Å or better. An ablation study was conducted on the selection of the GNN operator, radius of mutation neighborhood, and node features (Table S2, ESI†). We evaluated the predictive performance of our GNN-based model and compared it with various existing methods (Table S3, ESI†).

The performance of our GNN-based model surpassed the state-of-the-art models for protein thermostability prediction, especially in RMSE, it was 1.45 kcal mol^{-1} using ACDC-NN[53] while it was 1.27 kcal mol^{-1} in ours. Besides, our model achieved strong quantification accuracy for both direct and reverse mutations ($\sigma_{dir} = \sigma_{rev} = 1.27$ kcal mol^{-1}, $r_{dir} = r_{rev} = 0.60$), indicating a low prediction bias (Fig. S4, ESI†).

To further evaluate the generalizability of the GNN model, the same data-processing pipeline was performed to examine the model capacity of estimating the thermodynamic effects of the mutations from the tumor suppressor proteins p53 (including 42 mutants) and myoglobin (including 134 mutants) datasets (Tables S4, S5 and Fig. S5, ESI†). It turned out that the predicted $\Delta\Delta G$ values by our GNN model were highly correlated to experimental ones for mutations in both p53 protein and myoglobin.

3.3 Key residues of smTG predicted by GNN model

Since the GNN model fully conformed to the expectation, it was feasible to employ it for site prediction on our target smTG. Previous studies have indicated that flexible conformations are unfavorable for protein thermostability. Considering the length of the loop region and the distance from the catalytic center, we selected three segments of the loops for site-saturation mutagenesis using the GNN model to find out key residues improving the thermostability of smTG (Fig. S6, ESI†).

The input mutant structures were optimized by Rosetta, and the structural features were compared with the wild-type smTG to obtain predictable $\Delta\Delta G$ values. The prediction of saturation mutagenesis at all 45 sites including 855 mutations was performed. Among them, 56 mutations with $\Delta\Delta G$ value less than -0.50 kcal mol^{-1} were observed (Table S6, ESI†), suggesting these mutations might improve the thermostability of smTG. These mutations were mapped to their corresponding sites, and the number at each site was counted. The top four mutation sites were noted. Residues E54 and Q74 appeared nine times in all favorable mutations, residue H44 emerged eight times, and residue E87 emerged seven times. A total of 8 mutations with the largest $\Delta\Delta G$ changes were obtained. They were E87F (-1.79 kcal mol^{-1}), E87W (-1.18 kcal mol^{-1}), Q74Y (-1.25 kcal mol^{-1}), Q74F (-1.13 kcal mol^{-1}), E54K (-0.87 kcal mol^{-1}), H44C (-0.75 kcal mol^{-1}), H44I (-0.73 kcal mol^{-1}), and E54R (-0.62 kcal mol^{-1}), which were selected for next experimental verification.

3.4 Detection of residual enzyme activity of smTG mutants

To characterize the thermal stability, smTG and its mutants were heat treated at 60 °C. After that, the residual enzyme activity with CBZ-Gln-Gly as substrate, which was usually employed to determine the transglutaminase activity in previous studies,[20] was detected at 37 °C and shown in Fig. 2. Compared with the wild type, five of the eight mutants mentioned above exhibited better residual enzyme activity, suggesting the accuracy of our predictive model. Besides, four of them had higher activity than the wild-type smTG (Table S7, ESI†), although the residual enzyme activities of all smTGs were sharply reduced after incubation at 60 °C for 10 min. It should be noted that the mutant H44C displayed a remarkable enhancement in activity, twice as much as the wild type. What's more, extending the incubation time to 20 min, the mutants Q74F and E87W even displayed 3-fold and 2.3-fold residual enzyme activity compared to the wild-type smTG, respectively. The mutant E87W performed well after treatment for both 20 and 30 min and the mutant Q74F had the best residual enzyme activity after 30 min incubation at 60 °C. Based on the results, it was uncovered that compared to the wild type, the mutants H44C, Q74F, and E87W could exhibit better thermostability and reactivity at 60 °C for a period of time, which was a small step forward for potential applications in industry.

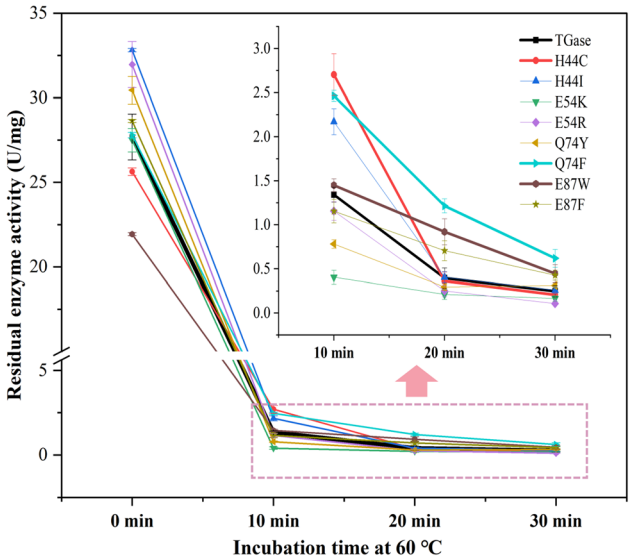

Fig. 2 Residual enzyme activities of smTG in wild type and its mutants.

3.5 Mutation-induced structural changes

Experimental evidence demonstrated that the thermostability of smTG was efficiently improved by mutants H44C, Q74F, and E87W. Three replicas of MD simulations of the wild-type and these three mutants were performed at 333 K for 100 ns to understand the structural rationale. Firstly, in order to investigate the changes in the overall structure of smTG, the temperature factor (*B*-factor) was calculated for the four systems, which is usually used as an important parameter to evaluate protein stability. The larger the *B*-factor value of the amino acid constituting the protein, the smaller the effect of the amino acid on stabilizing the protein. According to the calculated *B*-factor values, the mutants H44C, Q74F, and E87W appeared lower *B*-factor values compared to the wild-type smTG, indicating these mutations were beneficial to stabilize smTG. According to *B*-factor, the flexibility order of these three residues was identified. The mutant H44C remained that with the lowest flexibility and mutant E87W the highest. Staining was performed using Pymol in Fig. 3, and the color changing from red to blue represented an increase in structural stability.

Secondly, the radius of gyrification (R_g) and root mean square deviation (RMSD) were employed to evaluate protein flexibility of smTG. Considered as a metric, R_g determines the expansibility of protein during MD simulation and RMSD reflects the stability of protein main chain structure. They were collected to describe the structural diversity of four systems. In Fig. 4, for all systems, the fluctuation of RMSD was stable within 2.0 Å, indicating the mutations did not have an obvious impact on the overall structure. In addition, the value of R_g in each mutant was observed to be slightly lower than that of the wild type, mainly distributed below 20.8 Å. These results

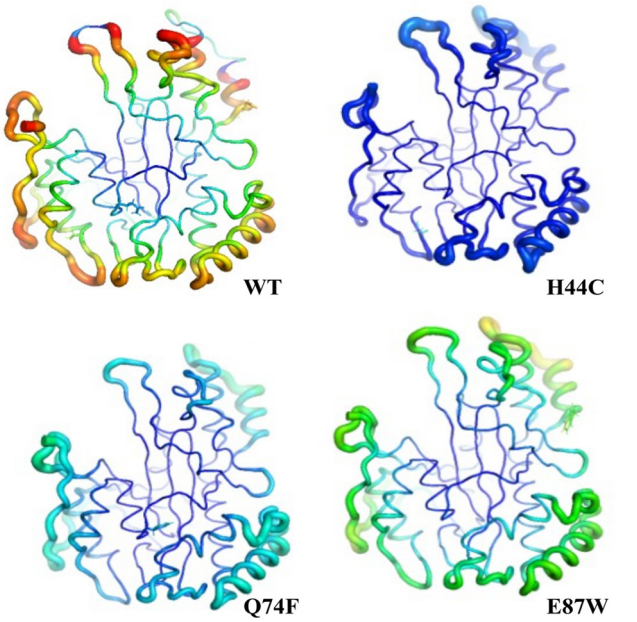

Fig. 3 Changes of *B*-factor in four systems. Average B-factor values (from small to large) are mapped to the color (blue-green-red color scale) and the tube width of the cartoon (from small to large).

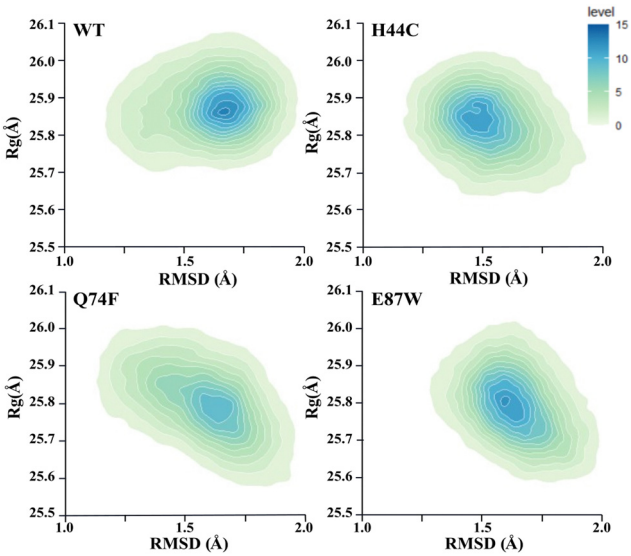

Fig. 4 Changes of RMSD and Rg in four systems. The *X*-axis represents the value of RMSD and the Y-axis represents the value of Rg.

demonstrated that all smTG systems reached their respective equilibriums after simulations, and smTG mutants exhibited more stability in the overall structure compared to the wild type.

Thirdly, we focused on some flexible regions with large variations. According to root mean square fluctuations (RMSF) values, there were three highly flexible regions (HFRs) in smTG (Fig. 5(A)). HFR I and HFR II were shown as loops and composed of residues 245–251 and 276–289, respectively, which were located near the catalytic center. HFR III encompassed residues 206–220 where a short α-helix was mixed with a long loop. As depicted in Fig. 5(B), the RMSF values of HFR I and HFR III mutants were basically less than half of those of the wild type's, decreasing from 4.5 Å to about 2.0 Å. In HFR II, it also reduced from 3.0 Å to less than 1.5 Å. This meant that the flexible regions in mutants became more stable than the wild type, particularly at HFR I and HFR II, showcasing that mutation either in close proximity to or far away from the catalytic site could impact the protein stability to some extent.

3.6 Additional interactions in mutants

The aforementioned analytical results suggested that the mutations (H44C, Q74F, E87W) would increase the stability of smTG, however, its underlying structural rationale remained unclear. As we know, the H-bond played a crucial role in maintaining the stability of the protein structure. Therefore, further analysis was focused on the H-bond interactions.

We counted the H-bonds by detecting the distance between donor and acceptor heavy atoms less than 3.0 Å and the H-bond angle larger than 135°. The frequency of eligible H-bonds occuring more than 80.0% in all MD trajectories were recorded (Table S8, ESI†). As expected, the mutants had more H-bonds than the wild type. Specifically, the mutant H44C exhibited 19 H-bonds, whereas the wild type had 15 H-bonds. Besides, both mutants Q74F and E87W displayed 18 H-bonds. Among them, the top two H-bonds were E58@O-S61@OH, situated at an

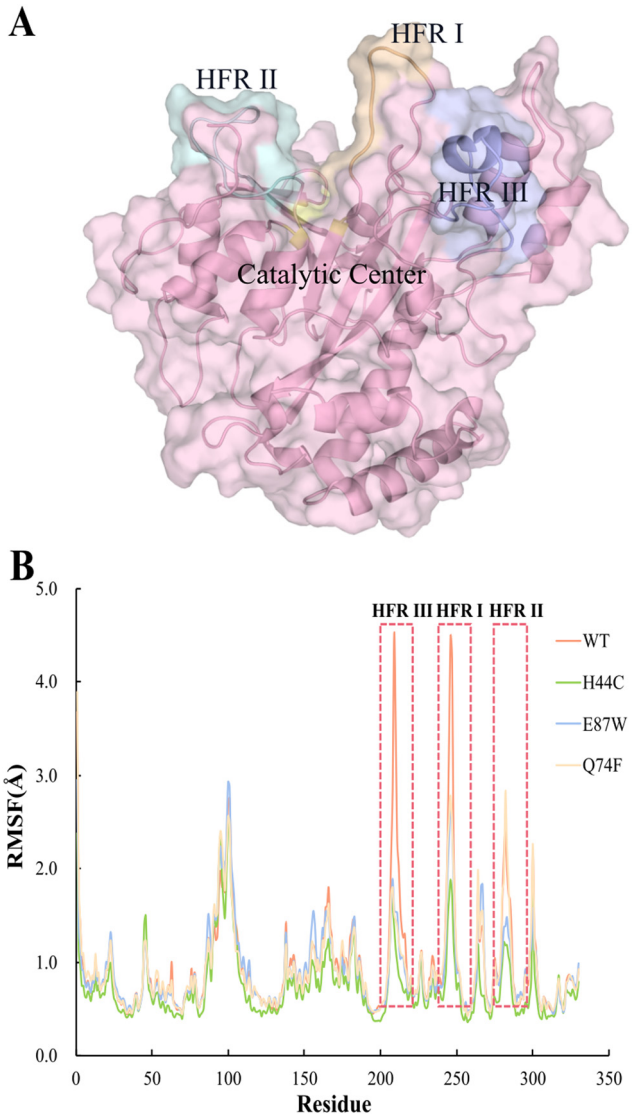

Fig. 5 RMSF of different models and the location of highly flexible regions (HFRs). (A) HFRs presented in the structure of smTG. (B) RMSF values of smTG and its mutants at 333 K. The positions of residues with large variations were annotated in the red boxes.

α-helix, and S292@O-W271@NH, located at β-sheets. They consistently occupied over 90.0% in all MD simulations of four systems. Both of them have been highlighted in orange in Fig. 6.

What's more, six H-bonds with high occupations in mutants, were found, namely R192@O-K317@NH, A173@O-T177@OH, D255@O-H201@NH, Q39@O-M52@NH, T68@O-S72@OH, and Y291@N-H274@OH. Especially, T68@O-S72@OH showed a significant increase in occupation by 15.0% in mutant H44C, and Y291@N-H274@OH exhibited an improvement of 10.0% in both mutants Q74F and E87W. These H-bonds are depicted in Fig. 6. Additionally, other H-bonds in MD trajectories performed enhancements in all mutants *versus* the wild-type smTG.

Apart from H-bonds, the newly emerging interactions near the mutation site attracted our attention. Fig. 7 illustrated the interactions around residues Q74F and E87W. Residue Q74 was located inside the smTG, and there were many hydrophobic residues around it. When Q74 was mutated to F, the benzene ring of F generated an R–π interaction with the side chain of P178. The distance between the center of mass of their side chains was utilized to describe the R–π interaction. Within 100 ns MD simulations, the distance was less than 5.0 Å in more than 75.5% of the trajectories (Fig. S7, ESI†), indicating the interaction between residues F74 and P178 had been enhanced, which might contribute to the stability of smTG. Besides, by clustering analysis, the conformation of residues P178, S179, F180, and K181 changed between helix and loop. According to DSSP analysis, these four residues near Q74 were found to remain at helix state more than the loop state by 10.0% after the mutation Q74F occurred (Table S9, ESI†). Furthermore, the benzene ring of F85 in mutant E87W formed a π–π interaction with the side chain of W87. Similarly, within 100 ns MD simulations, the proportion with a distance between the center of mass of the side chain of F85 and W87 within 5.0 Å was more than 66.1% in all trajectories (Fig. S8, ESI†). In summary, the stability of mutants had been apparently improved by critical interactions including H-bonds, R–π interaction, and π–π interaction.

3.7 Substrate binding in smTG mutants

The wild-type and mutant smTGs were docked with standard substrate CBZ-Gln-Gly. The best of 200 docking poses was selected by considering both scoring results and loading distance between C64 of smTG and Gln of CBZ-Gln-Gly (Fig. S9, ESI†). Following three times 50 ns MD simulations at 333 K, the RMSD values of the mutants H44C, Q74F as well as E87W, and the RMSF values of HFRs were observed to be lower than that of the wild type. This is consistent with the aforementioned results without substrate (Fig. S10, ESI†).

To understand the binding affinity, the MM-GBSA method was utilized to evaluate the binding energies of four systems. The binding energy in mutant H44C was found to be the lowest, and its average energy was -26.60 kcal mol^{-1}, suggesting that the substrate had a stronger affinity to mutant H44C than other systems (Table S10, ESI†). On the other hand, the energy contribution of individual amino acids was calculated in each system. Six amino acids, including V252, H277, F254, C64, Y278, and Y62 near the active pocket, were found with energy contributions larger than 0.50 kcal mol^{-1} in the wild type (Fig. 8). Among them, residues H277, F254, and Y62 exhibited great binding affinity with the substrate in all mutants, which surpassed their energy contributions in the wild type. Furthermore, according to the energy decomposition, Y302 (-0.65 kcal mol^{-1}) and Y291 (-0.53 kcal mol^{-1}) in the mutant H44C and R5 (-0.60 kcal mol^{-1}), and L285 (-1.02 kcal mol^{-1}) in Q74F were found critical for substrate binding, whereas their contributions to substrate affinity were almost zero in the wild type. Overall, residues Y302, Y291, R5, and L285 demonstrated a high substrate binding in mutants H44C and Q74F, suggesting the mutations might have an impact on substrate recognition and catalytic activity of smTG. The contribution of binding

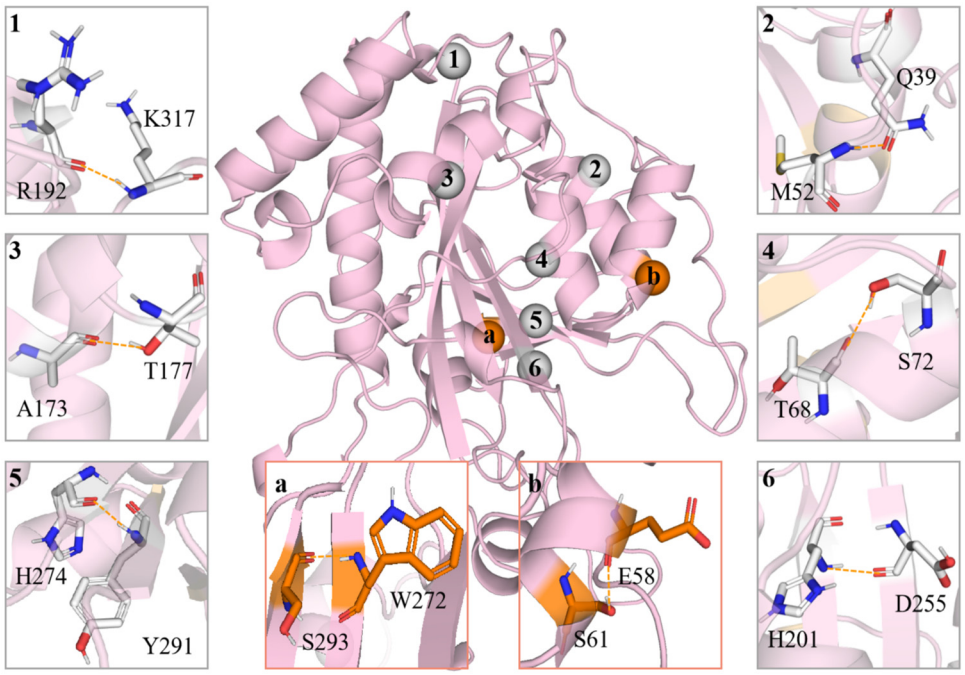

Fig. 6 H-bond interactions existed in mutant smTG. The top two H-bonds with the highest proportion of the four systems are highlighted in orange, and H-bonds with high occupation are depicted in grey.

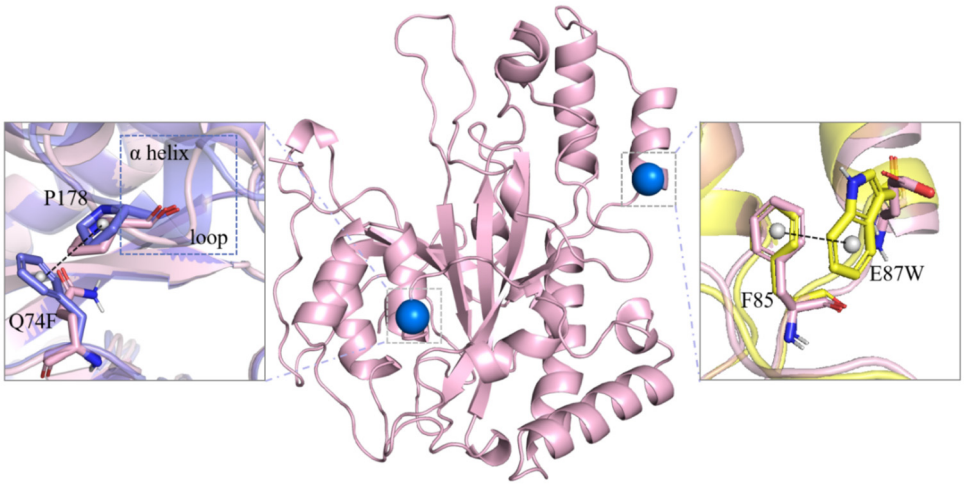

Fig. 7 Interactions around mutation Q74F and E87W. The wild type, Q74F mutant and E87W mutant were shown in pink, purple, and yellow, respectively.

energy to the improvement of catalytic activity of smTG in mutant E87W was not obvious.

4 Discussion

Given the importance of smTG in various applications, it is desirable to improve thermostability on the basis of maintaining or even enhancing enzyme activity.[2] In recent years, both protein engineering and rational design have been used to improve the diverse applications of smTG. Based on a combination of random and saturation mutagenesis, the S2P-S23Y-S24N-H289Y-K249L mutant was detected and showed great improvement in specific activity.[54] However, this technique is expensive and time-consuming. At present, FoldX, ELASPIC,[55] Amber TI, Rosetta and other modules have been developed to predict the change of folding free energy ($\Delta\Delta G$). With the rapid development of deep learning and graph neural network (GNN), GNN has become a popular choice in biological fields, especially in structural bioinformatics with respect to protein engineering and drug design.[56,57] In this study, based on the crystal structure database, a GNN model was constructed

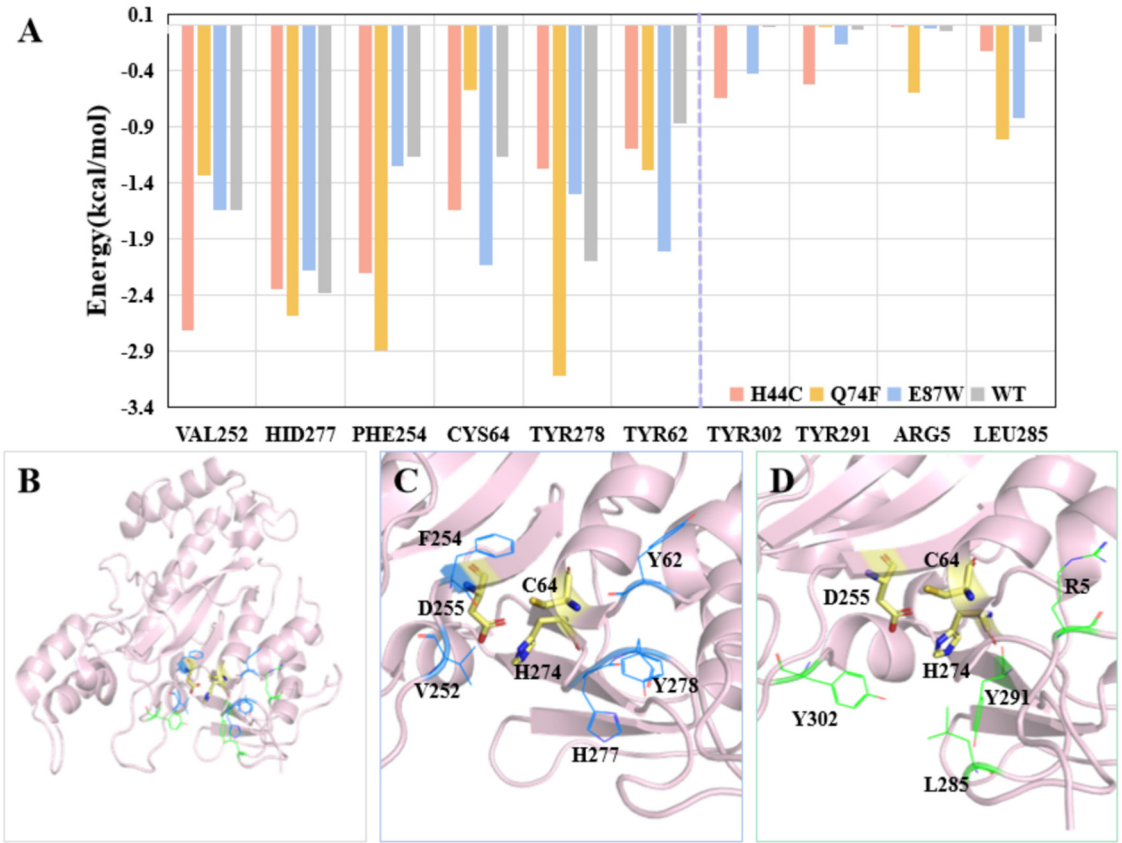

Fig. 8 Analysis of residues that played an important role in the interaction with substrates. (A) Energy decomposition. Residues that contributed significantly to substrate binding in all four systems are shown at the left side of the purple dashed line, and residues that had improved effects in mutants are shown at the right side; (B) the overall structure of smTG. (C) Key residues obtained by energy decomposition are shown in blue and the catalytic residues are shown in yellow. (D) Residues R5, L285, Y291, and Y302 are shown in green.

to estimate the $\Delta\Delta G$ upon site-directed mutation. The application of GNN enables the model to capture key residues and extra spatial information from network topology. It can provide an effective alternative in protein design to obtain favorable mutants with high thermostability. However, current methods do not guarantee to generalize to all protein targets by predicting general protein stability. For example, our model did not exceed the performance of ACDC-NN (a CNN-based method) on the p53 dataset. Nonetheless, the updated version of GNN further enhanced the prediction of protein thermostability now.[58] Beyond our model, multi-site mutations and the residue interaction network are still interesting questions. In addition, the protein pre-trained language model that learns global evolutionary representation for each amino acid can provide more effective sequence-based features.

In this work, the predictive mutants (H44C, Q74F, and E87W) were verified by the determination of the residual enzyme activity at 60 °C and they all exhibited increases in thermostability and activity. Further insight into the impact of the mutations on the structures of smTG was achieved through the performance of MD simulations. According to the computational analysis, these mutations could reduce the flexibility of smTG and make the structure more compact. Various factors, such as hydrophobic interactions, H-bonds, salt bridges, cation–π bond interactions, aromatic ring interactions, and disulfide bonds, were involved in the enhancement of thermostability. Here, in mutants H44C, Q74F and E87W, the critical factors were proposed to be H-bonding, R–π, and π–π interactions, respectively. Based on the results of MM-GBSA, in addition to improving the thermal stability, the residues around the catalytic pocket also enhanced the binding affinity to the substrate in mutants H44C and Q74F, thereby improving the catalytic activity.

According to the current results, we can combine the existing three mutation sites to test the effect of combined mutation in the future. The effective mutation sites also can be used to continue training the GNN model and predict the effective mutation sites of two-point, three-point and multi-point mutation at the same time. Guided by computation-aided site-directed mutations and structural analysis, our study improves both the thermostability and the enzymatic activity of smTG, which will be useful for many industrial applications.

5 Conclusions

In summary, we adopted a graph attention neural network as the foundational framework for our predictive model to detect

critical residues for the improvement of smTG thermostability. Through systematic residual enzyme activity experiments, we successfully confirmed the efficacy of three mutants (H44C, Q74F, and E87W) with the enhancement of thermostability and reactivity. Following a 10 min incubation at 60 °C, mutant H44C displayed 2-fold residual enzyme activity, respectively, in comparison to smTG. Furthermore, mutants Q74F and E87W, respectively, exhibited 3-fold and 2.3-fold activity after 20 min. Moreover, we delved into the molecular mechanisms of these mutants to uncover their structural changes. The thermostability was improved by enhancing hydrogen bond interactions, increasing additional residue interactions, and reducing loop flexibility. This exploration pointed us toward directions for identifying new mutation sites and designing innovative variants. Our work deeply probes molecular-level alterations and offers novel perspectives and strategies for enzyme engineering.

Author contributions

Yongzhen Li: investigation, formal analysis, visualization, writing – original draft, and writing – review & editing. Banghao Wu: investigation and writing – original draft. Yumeng Zhang: methodology, investigation, writing – original draft, and writing – review & editing. Lanxuan Liu: conceptualization and supervision. Linquan Bai: conceptualization and supervision. Ting Shi: conceptualization, supervision, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (no. 2021YFC2100600) and the National Natural Science Foundation of China (no. 32270038). The authors thanked the SJTU-HPC computing facility award for modeling and computation.

References

1 L. Parrotta, U. K. Tanwar, I. Aloisi, E. Sobieszczuk-Nowicka, M. Arasimowicz-Jelonek and S. Del Duca, Plant Transglutaminases: New Insights in Biochemistry, Genetics, and Physiology, *Cells*, 2022, **11**(9), 1529.

2 H. Fuchsbauer, Approaching Transglutaminase from *Streptomyces* Bacteria over Three Decades, *FEBS J.*, 2022, **289**(16), 4680–4703.

3 L. Duarte, C. R. Matte, C. V. Bizarro and M. A. Z. Ayub, Review Transglutaminases: Part II—Industrial Applications in Food, Biotechnology, Textiles and Leather Products, *World J. Microbiol. Biotechnol.*, 2020, **36**(1), 11.

4 P. Strop, Versatility of Microbial Transglutaminase, *Bioconjugate Chem.*, 2014, **25**(5), 855–862.

5 Y. Kim, J. I. Kee, S. Lee and S.-H. Yoo, Quality Improvement of Rice Noodle Restructured with Rice Protein Isolate and Transglutaminase, *Food Chem.*, 2014, **145**, 409–416.

6 M. F. Mazzeo, R. Bonavita, F. Maurano, P. Bergamo, R. A. Siciliano and M. Rossi, Biochemical Modifications of Gliadins Induced by Microbial Transglutaminase on Wheat Flour, *Biochim. Biophys. Acta, Gen. Subj.*, 2013, **1830**(11), 5166–5174.

7 J. Domagała, D. Najgebauer-Lejko, I. Wieteska-Śliwa, M. Sady, M. Wszołek, G. Bonczar and M. Filipczak-Fiutak, Influence of Milk Protein Cross-Linking by Transglutaminase on the Rennet Coagulation Time and the Gel Properties: Effect of Transglutaminase on Rennet Milk Coagulation, *J. Sci. Food Agric.*, 2016, **96**(10), 3500–3507.

8 M.-L. Shen, J.-Y. Ciou, L.-S. Hsieh and C.-L. Hsu, Recombinant Streptomyces Netropsis Transglutaminase Expressed in Komagataella Phaffii (Pichia Pastoris) and Applied in Plant-Based Chicken Nugget, *World J. Microbiol. Biotechnol.*, 2023, **39**(8), 200.

9 C. K. Marx, T. C. Hertel and M. Pietzsch, Random Mutagenesis of a Recombinant Microbial Transglutaminase for the Generation of Thermostable and Heat-Sensitive Variants, *J. Biotechnol.*, 2008, **136**(3–4), 156–162.

10 K. Buettner, T. C. Hertel and M. Pietzsch, Increased Thermostability of Microbial Transglutaminase by Combination of Several Hot Spots Evolved by Random and Saturation Mutagenesis, *Amino Acids*, 2012, **42**(2–3), 987–996.

11 T. Kashiwagi, K. Yokoyama, K. Ishikawa, K. Ono, D. Ejima, H. Matsui and E. Suzuki, Crystal Structure of Microbial Transglutaminase fromStreptoverticillium Mobaraense, *J. Biol. Chem.*, 2002, **277**(46), 44252–44260.

12 K. Yokoyama, D. Ogaya, H. Utsumi, M. Suzuki, T. Kashiwagi, E. Suzuki and S. Taguchi, Effect of Introducing a Disulfide Bridge on the Thermostability of Microbial Transglutaminase from Streptomyces Mobaraensis, *Appl. Microbiol. Biotechnol.*, 2021, **105**(7), 2737–2745.

13 X. Wang, J. Du, B. Zhao, H. Wang, S. Rao, G. Du, J. Zhou, J. Chen and S. Liu, Significantly Improving the Thermostability and Catalytic Efficiency of *Streptomyces Mobaraensis* Transglutaminase through Combined Rational Design, *J. Agric. Food Chem.*, 2021, **69**(50), 15268–15278.

14 N. C. Garbett and J. B. Chaires, Thermodynamic Studies for Drug Design and Screening, *Expert Opin. Drug Discovery*, 2012, **7**(4), 299–314.

15 S. Stefl, H. Nishi, M. Petukh, A. R. Panchenko and E. Alexov, Molecular Mechanisms of Disease-Causing Missense Mutations, *J. Mol. Biol.*, 2013, **425**(21), 3919–3936.

16 R. Guerois, J. E. Nielsen and L. Serrano, Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations, *J. Mol. Biol.*, 2002, **320**(2), 369–387.

17 K. T. Simons, C. Kooperberg, E. Huang and D. Baker, Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions, *J. Mol. Biol.*, 1997, **268**(1), 209–225.

18 F. Pucci, K. V. Bernaerts, J. M. Kwasigroch and M. Rooman, Quantification of Biases in Predictions of Protein Stability Changes upon Mutations, *Bioinformatics*, 2018, **34**(21), 3659–3665.
19 D. R. Usmanova, N. S. Bogatyreva, J. Ariño Bernad, A. A. Eremina, A. A. Gorshkova, G. M. Kanevskiy, L. R. Lonishin, A. V. Meister, A. G. Yakupova, F. A. Kondrashov and D. N. Ivankov, Self-Consistency Test Reveals Systematic Bias in Programs for Prediction Change of Stability upon Mutation, *Bioinformatics*, 2018, **34**(21), 3653–3658.
20 G. Wang, X. Liu, K. Wang, Y. Gao, G. Li, D. T. Baptista-Hon, X. H. Yang, K. Xue, W. H. Tai, Z. Jiang, L. Cheng, M. Fok, J. Y.-N. Lau, S. Yang, L. Lu, P. Zhang and K. Zhang, Deep-Learning-Enabled Protein–Protein Interaction Analysis for Prediction of SARS-CoV-2 Infectivity and Variant Evolution, *Nat. Med.*, 2023, **29**, 2007–2018.
21 L. F. Krapp, L. A. Abriata, F. Cortés Rodriguez and M. Dal Peraro, PeSTo: Parameter-Free Geometric Deep Learning for Accurate Prediction of Protein Binding Interfaces, *Nat. Commun.*, 2023, **14**(1), 2175.
22 P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, Graph Attention Networks, *arXiv*, 1710.10903v3, 2018.
23 J. Stourac, J. Dubrava, M. Musil, J. Horackova, J. Damborsky, S. Mazurenko and D. Bednar, FireProtDB: Database of Manually Curated Protein Stability Data, *Nucleic Acids Res.*, 2021, **49**(D1), D319–D324.
24 M. Olivier, R. Eeles, M. Hollstein, M. A. Khan, C. C. Harris and P. Hainaut, The IARC TP53 Database: New Online Mutation Analysis and Recommendations to Users, *Hum. Mutat.*, 2002, **19**(6), 607–614.
25 K. P. Kepp, Towards a ''Golden Standard'' for Computing Globin Stability: Stability and Structure Sensitivity of Myoglobin Mutants, *Biochim. Biophys. Acta, Proteins Proteomics*, 2015, **1854**(10), 1239–1248.
26 B. Faezov and R. L. Dunbrack, PDBrenum: A Webserver and Program Providing Protein Data Bank Files Renumbered According to Their UniProt Sequences, *PLoS One*, 2021, **16**(7), e0253411.
27 D. S. Goodsell, C. Zardecki, L. Di Costanzo, J. M. Duarte, B. P. Hudson, I. Persikova, J. Segura, C. Shao, M. Voigt, J. D. Westbrook, J. Y. Young and S. K. Burley, RCSB Protein Data Bank: Enabling Biomedical Research and Drug Discovery, *Protein Sci.*, 2020, **29**(1), 52–65.
28 M. D. Tyka, D. A. Keedy, I. André, F. DiMaio, Y. Song, D. C. Richardson, J. S. Richardson and D. Baker, Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping, *J. Mol. Biol.*, 2011, **405**(2), 607–618.
29 R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme and J. J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design, *J. Chem. Theory Comput.*, 2017, **13**(6), 3031–3048.
30 G. Lv, Z. Hu, Y. Bi and S. Zhang, Learning Unknown from Correlations: Graph Neural Network for Inter-Novel-Protein Interaction Prediction, *arXiv*, 2105.06709v3, 2021.
31 M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding and M. Steinegger, Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments, *Nucleic Acids Res.*, 2017, **45**(D1), D170–D176.
32 M. Remmert, A. Biegert, A. Hauser and J. Söding, HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment, *Nat. Methods*, 2012, **9**(2), 173–175.
33 M. Fey and J. E. Lenssen, Fast Graph Representation Learning with PyTorch Geometric, *arXiv*, 1903.02428v3, 2019.
34 G. Thiltgen and R. A. Goldstein, Assessing Predictors of Changes in Protein Stability upon Mutation Using Self-Consistency, *PLoS One*, 2012, **7**(10), e46084.
35 J. Zotzel, R. Pasternack, C. Pelzer, D. Ziegert, M. Mainusch and H.-L. Fuchsbauer, Activated Transglutaminase from Streptomyces Mobaraensis Is Processed by a Tripeptidyl Aminopeptidase in the Final Step: Tripeptidyl Aminopeptidase, *Eur. J. Biochem.*, 2003, **270**(20), 4149–4155.
36 N. E. Juettner, S. Schmelz, A. Kraemer, S. Knapp, B. Becker, H. Kolmar, A. Scrima and H. Fuchsbauer, Structure of a Glutamine Donor Mimicking Inhibitory Peptide Shaped by the Catalytic Cleft of Microbial Transglutaminase, *FEBS J.*, 2018, **285**(24), 4684–4694.
37 T. J. Dolinsky, J. E. Nielsen, J. A. McCammon and N. A. Baker, PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations, *Nucleic Acids Res.*, 2004, **32**, W665–W667.
38 X. Pan, H. Wang, C. Li, J. Z. H. Zhang and C. Ji, MolGpka: A Web Server for Small Molecule pK_a Prediction Using a Graph-Convolutional Neural Network, *J. Chem. Inf. Model.*, 2021, **61**(7), 3159–3165.
39 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility, *J. Comput. Chem.*, 2009, **30**(16), 2785–2791.
40 Y. Zhao and D. G. Truhlar, The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06 Functionals and 12 Other Functionals, *Theor. Chem. Acc.*, 2008, **119**(5–6), 525.
41 P. W. Payne, The Hartree–Fock Theory of Local Regions in Molecules, *J. Am. Chem. Soc.*, 1977, **99**(11), 3787–3794.
42 V. A. Rassolov, M. A. Ratner, J. A. Pople, P. C. Redfern and L. A. Curtiss, 6-31G* Basis Set for Third-Row Atoms, *J. Comput. Chem.*, 2001, **22**(9), 976–984.
43 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima,

Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *GAUSSIAN 16 (Revision C.01)*, Gaussian Inc., Wallingford, CT, 2016.

44 J. Wang, W. Wang, P. A. Kollman and D. A. Case, Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations, *J. Mol. Graphics Modell.*, 2006, **25**(2), 247–260.

45 T. Lu and F. Chen, Multiwfn: A Multifunctional Wavefunction Analyzer, *J. Comput. Chem.*, 2012, **33**(5), 580–592.

46 W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, *J. Am. Chem. Soc.*, 1995, **117**(19), 5179–5197.

47 D. A. Case, I. Y. Ben-Shalom, S. R. Rozell, D. S. Cerutti, T. E. I. I. I. Cheatham, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York and P. A. Kollman, *AMBER 18*, University of California, San Francisco, 2018.

48 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, *J. Chem. Theory Comput.*, 2015, **11**(8), 3696–3713.

49 T. Darden, D. York and L. Pedersen, Particle Mesh Ewald: An N·log (N) Method for Ewald Sums in Large Systems, *J. Chem. Phys.*, 1993, **98**(12), 10089–10092.

50 J. P. Ryckaert, G. Ciccotti and H. Berendsen, Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes, *J. Comput. Phys.*, 1997, **23**, 321–341.

51 D. R. Roe and T. E. Cheatham, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data, *J. Chem. Theory Comput.*, 2013, **9**(7), 3084–3095.

52 B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, *MMPBSA.Py*: An Efficient Program for End-State Free Energy Calculations, *J. Chem. Theory Comput.*, 2012, **8**(9), 3314–3321.

53 S. Benevenuta, C. Pancotti, P. Fariselli, G. Birolo and T. Sanavia, An Antisymmetric Neural Network to Predict Free Energy Changes in Protein Variants, *J. Phys. D: Appl. Phys.*, 2021, **54**(24), 245403.

54 B. Böhme, B. Moritz, J. Wendler, T. C. Hertel, C. Ihling, W. Brandt and M. Pietzsch, Enzymatic Activity and Thermoresistance of Improved Microbial Transglutaminase Variants, *Amino Acids*, 2020, **52**(2), 313–326.

55 D. K. Witvliet, A. Strokach, A. F. Giraldo-Forero, J. Teyra, R. Colak and P. M. Kim, ELASPIC Web-Server: Proteome-Wide Structure-Based Prediction of Mutation Effects on Protein Stability and Binding Affinity, *Bioinformatics*, 2016, **32**(10), 1589–1591.

56 S. Wang, H. Tang, P. Shan, Z. Wu and L. Zuo, ProS-GNN: Predicting Effects of Mutations on Protein Stability Using Graph Neural Networks, *Comput. Biol. Chem.*, 2023, **107**, 107952.

57 S. Wang, H. Tang, Y. Zhao and L. Zuo, BayeStab: Predicting Effects of Mutations on Protein Stability with Uncertainty Quantification, *Protein Sci.*, 2022, **31**(11), e4467.

58 H. Gong, Y. Zhang, C. Dong, Y. Wang, G. Chen, H. Li, L. Liu, J. Xu and G. Li, Unbiased Curriculum Learning Enhanced Global-Local Graph Neural Network for Protein Thermodynamic Stability Prediction, *Bioinformatics*, 2023, **39**(10), btad589.

