PIBB 生物化学与生物物理进展
*Progress in Biochemistry and Biophysics*
2021,48(10):1214~1232
www.pibb.ac.cn

# Integration of Machine Learning Improves The Prediction Accuracy of Molecular Modelling for *M. jannaschii* Tyrosyl−tRNA Synthetase Substrate Specificity*

DUAN Bing-Ya, SUN Ying-Fei**

(*School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing* 100190, China)

**Abstract** Design of enzyme binding pocket to accommodate substrates with different chemical structure is a great challenge. Traditionally, thousands even millions of mutants have to be screened in wet-lab experiments to find a ligand-specific mutant and large amount of time and resources are consumed. To accelerate the screening process, we propose a novel workflow through integration of molecular modeling and data-driven machine learning method to generate mutant libraries with high enrichment ratio for recognition of specific substrate. We collected all the *M. jannaschii* tyrosyl-tRNA synthetase (*Mj*. TyrRS) mutants reported in the literature to compare and analyze the sequence and structural feature and difference between mutant and wild type *Mj*. TyrRS. *Mj.* TyrRS is used as an example since the sequences and structures of many unnatural amino acid specific *Mj.* TyrRS mutants have been reported. Based on the crystal structures of different *Mj.* TyrRS mutants and Rosetta modeling result, we found D158G/P is the critical residue which influences the backbone disruption of helix with residue 158−163. Our results showed that compared with random mutation, Rosetta modeling and score function calculation can elevate the enrichment ratio of desired mutants by 2-fold in a test library having 687 mutants, while after calibration by machine learning model trained using known data of *Mj.* TyrRS mutants and ligand, the enrichment ratio can be elevated by 11-fold. This molecular modeling and machine learning-integrated workflow is anticipated to significantly benefit to the *Mj.* tyrRS mutant screening and substantially reduce the time and cost of wet-lab experiments. Besides, this novel process will have broad application in the field of computational protein design.

**Key words** tyrosyl-tRNA synthetase, genetic code expansion, enzyme substrate specificity, Rosetta, molecular modelling, machine learning

**DOI**: 10.16476/j.pibb.2020.0425

Genetic code expansion technology[1] has been widely used in biological research and can be applied in monitoring protein conformational change[2] caused by PTM[3] such as biophysical probe[4], improving enzyme activity[5] and designing proteins with novel catalytic functions[6-8]. Through this technology, we can incorporate artificially designed unnatural amino acid (UAA) into almost any specific site of target protein. Usually an orthogonal tRNA and amino acyl-tRNA synthetase (aaRS) mutant pair is necessary for recognition of specific UAA and subsequent acyl ligation to tRNA. There are more than 100 UAAs and their corresponding orthogonal aaRS mutants have been reported, UAAs with novel chemical structures, biophysical and biological function will significantly benefit to the biological research and protein therapeutics[9]. Identification of the UAA's corresponding aaRS by researchers has never stopped. Traditionally, an aaRS mutant recognizing specific UAA is selected by several rounds of positive and negative screening from large and diverse aaRS mutant libraries designed rationally from wild type

* This work was supported by a grant from The National Natural Science Foundation of China (61431017).

** Corresponding author.
Tel: 86-10-62626682, E-mail: yfsun@ucas.ac.cn

Received: December 4, 2020 Accepted: February 9, 2021

(WT) aaRS[1]. The screening process is very tedious, manpower-costing and error-prone. Thus, a more rapid and efficient method to identify aaRS mutant for specific UAA is in urgent need.

From the perspective of computational chemistry, finding aaRS mutant for specific UAA is a problem in protein design. The sequence space of WT aaRS as receptor is explored with the goal of finding the mutant having lowest binding free energy to the UAA ligand. In recent years, computational chemistry based-molecular modelling methods, such as protein homology modelling[10], protein structure prediction[11-12], molecular docking[13] and protein design[12, 14-16] have been rapidly developed. Large number of successful cases on enzyme design for substrate selectivity[17] are reported.

Recently there have been great advances in artificial intelligence (AI) technology, such as machine learning (ML) and deep learning. AI has been used in computer vision, speech recognition, machine translation, small-molecule drug design[18], protein engineering[14], antibody structure prediction[19], antibody design[20], protein structure prediction[11] and protein design[21]. In general, AI can extract representative features of protein sequence and structure from large amount of chemical molecule and protein data, learn internal patterns which can't explicitly spotted by human, and utilize experimentally validated properties such as binding affinity data, $IC_{50}$ and enzyme activity as labels to train a model which can be used to predict which sample has target property of interest from huge number of molecules and diversified protein mutant libraries. For example, improved score function of molecular docking software[22], improved protein design performance[23] and accurate prediction of thermostability for protein mutants[24] have been achieved through ML and deep learning. Combined with computational chemistry-based method, data-driven AI method is giving satisfactory solution and accurate prediction on many biological problems.

In this work, we collected all the *M. jannaschii* tyrosyl-tRNA synthetase (*Mj.* TyrRS) mutants reported in the literature to compare and analyze the sequence and structural feature and differences between mutants and WT *Mj.* TyrRS. Then we use Rosetta EnzymeDesign method[25] to model the structure and predict binding pose of UAA-*Mj.* TyrRS mutant pair. Finally, ML model is integrated to calibrate the score function for aaRS substrate selectivity prediction and mutant design to achieve better prediction accuracy. We think the improved *Mj.* TyrRS-UAA selectivity prediction model will be extremely useful for *in silico* screening of UAA-specific aaRS mutants.

## 1 Crystal structures comparison between *Mj.* TyrRS mutant and WT

Firstly, we collected the information of all UAAs and corresponding *Mj.* TyrRS mutants whose high-resolution X-ray crystal structures have been released, as shown in Figure 1, Table 1 and Supplementary material Table S1. In total there are 52 UAAs and 132 TyrRS mutants. The X-ray crystal structures of 18 TyrRS mutants have been reported (Table 1). The diversity of UAAs (Figure 1) is large. Both of small and large, polar and non-polar, hydrophilic and hydrophobic substitution groups of p-hydroxy of Tyr are included in these UAAs. In WT *Mj.* TyrRS, hydrogen bond network between tyrosine hydroxy group and pocket residues stabilizes the tyrosine ligand and lowers the binding free energy (Figure 2). In order to recognize other UAAs in Figure 1, mutations with different amino acid types and physicochemical properties have to be introduced to accommodate different chemical properties.

For a reliable model of the mutant structure, the backbone fluctuation introduced by mutation should be small compared with that introduced by WT. It has been reported that some mutations can change the backbone conformation[26], so next we aligned the crystal structures of *Mj.* TyrRS mutants and WT and then calculated the RMSD of backbone N, CA, C, O of binding pocket residues between mutant and WT to compare the structural difference. As shown in Table 1 and Figure 3a, the structures with PDB ID 1ZHO, 1ZH6, 2AG6, 3D6U, 3D6V have average backbone RMSD larger than 1.2, of which backbone conformation is changed by mutation compared with WT. Other mutants have small backbone fluctuation which can be ignored when building the homology model. Four of the 16 mutant structures have large backbone fluctuation (Table 1). In Figure 3b, the backbone RMSD of different binding pocket residues are compared. We can see residues 158−163 have

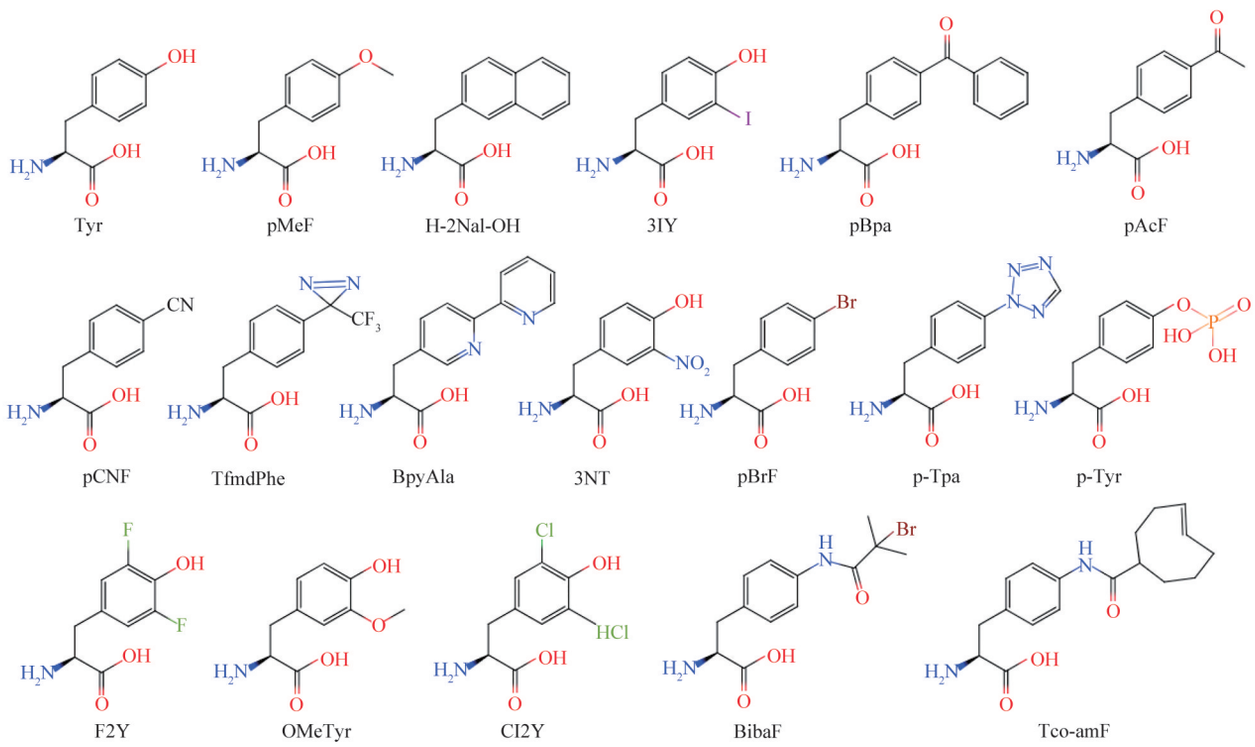


**Fig. 1　Chemical structures of the UAAs recognized by *Mj.* TyrRS mutants whose X−ray crystal structures have been solved**

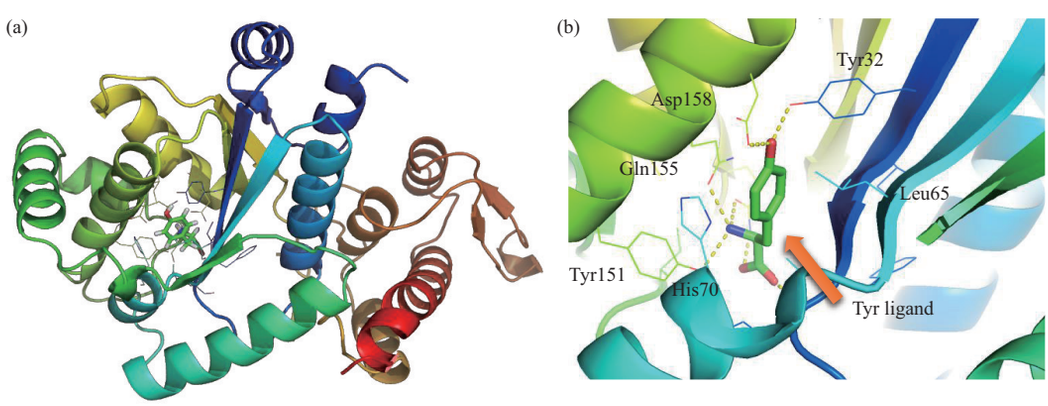


**Fig. 2　Crystal structures of WT *Mj.* TyrRS and tyrosine complex（PDB ID：1J1U）**

The figure is rendered by Pymol[27]. (a) *Mj.* TyrRS is shown as cartoon. Tyrosine ligand is shown as sticks. Surrounding residues are shown as lines. (b) The binding pocket of tyrosine ligand. Tyrosine forms hydrogen bond network with Tyr32, Asp158 and Gln155, which stabilizes the hydroxy group of tyrosine.

larger RMSD than other residues, which are on the alpha helix near UAA ligand. Four representative mutant structures are chosen and superimposed on WT structure, as shown in Figure 4. In the structures of 1ZH0 and 2AG6 the backbone change of alpha helix is large, while in those of 2HGZ and 4PBR the change is small. This implies it may be difficult to obtain accurate UAA and pocket residue binding pose for mutants bearing large backbone conformation change in homology model.

**Table 1　UAA index and name，aaRS mutation，PDB ID and X−ray crystal structure backbone RMSD of UAA binding pocket residues**[1)]

| UAA Index | Full name of UAA | Abbrevation of UAA | *Mj.* TyrRS mutation | PDB ID | Average backbone RMSD of pocket residues between mutant and WT | Large mutant backbone fluctuation compared with WT | Reference |
|---|---|---|---|---|---|---|---|
| 000 | L-tyrosine | Tyr | Wild type | 1J1U | 0 | No | ［28］ |
| 001 | p-Methoxy-L-phenylalanine | pMeF | Y32Q D158A E107T L162P | 1U7X | 1.02 | No | ［29］ |
| 003 | 3-(2-Naphthyl)-L-alanine | H-2Nal-OH | Y32L D158P I159A L162Q A167V | 1ZH0 | 2.20 | Yes | ［30］ |
| 004 | 3-Iodo-L-tyrosine | 3IY | H70A D158T | 2ZP1 | 0.36 | No | |
| 005 | p-Benzoyl-L-phenylalanine | pBpa | Y32G E107P D158G I159T | 2HGZ | 0.38 | No | ［31］ |
| 009 | p-Acetyl-L-phenylalanine | pAcF | Y32L D158G I159C L162R | 1ZH6 | 1.27 | Yes | ［32］ |
| 015 | p-Cyano-L-phenylalanine | pCNF | Y32L L65V F108W Q109M D158G I159A | 3QE4 | 0.68 | No | ［33］ |
| 018 | p-(3-Trifluoromethyl-3H-diazirin-3-yl)-phenylalanine | TfmdPhe | Y32I H70F E107S Q109M D158P I159L L162E | 3D6U | 2.15 | Yes | ［34］ |
| 020 | Bipyridylalanine | BpyAla | Y32G L65Y H70A Q155E D158G I159W L162S | 2PXH | 0.40 | No | ［35］ |
| 035 | 3-Nitro-L-tyrosine | 3NT | Y32H H70C D158S I159A L162R | 4NDA | 0.39 | No | ［36］ |
| 042 | p-Bromo-L-phenylalanine | pBrF | Y32L E107S D158P I159L L162E | 2AG6 | 2.12 | Yes | ［30］ |
| 049 | p-(2-Tetrazole)-L-phenylalanine | p-Tpa | Y32L L65I Q109M D158G L162V V164G | 3N2Y | 0.33 | No | ［37］ |
| 053 | 3,5-Difluoro tyrosine | F2Y | Y32R L65Y H70G F108N Q109C D158N L162S | 4HJX | 0.84 | No | ［3］ |
| 058 | 3-o-Methyl tyrosine | OMeTyr | Y32E L65S H70G Q109G D158N L162V | 4HPW | 0.88 | No | |
| 062 | 3,5-Dichloride tyrosine | Cl2Y | Y32L L65I H70G F108I Q109L Y114G D158S L162M | 4NX2 | 0.76 | No | ［38］ |
| 065 | 4-(2-bromoisobutyramido)-phenylalanine | BibaF | Y32G L65E F108W Q109M D158S L162K | 4PBR | 0.25 | No | ［39］ |
| 066 | 4-Trans-cyclooctene-amidopheylalanine | Tco-amF | Y32G L65E F108W Q109M D158S L162K | 4PBT | 0.57 | No | ［39］ |
| 068 | O-phosphotyrosine | pTyr | Y32S L65A F108K Q109H D158G L162K | 5U36 | 2.33 | Yes | ［26］ |

[1)] Crystal structures of mutants with the same sequence are only shown once.

How different mutations induce the backbone conformation change remains unclear. It was reported that in pTyr *Mj.* TyrRS, D158G mutation acted as a helix breaker and caused the rearrange of the helix and opened up the pocket to accommodate the bulky UAA[26]. Though pCNF *Mj.* TyrRS has the same D158G mutation, no obvious backbone conformation change is observed (Table 1). To further investigate the critical residues affecting the helix conformation, we calculated AAindex[40] (531 types of numerical indices representing various physicochemical and biochemical properties of amino acids) of residues 158−163 for the 17 *Mj.* TyrRS in Table 1. Totally there are 531×6=3 186 features for each TyrRS mutant. We performed analysis of variance (ANOVA) by sklearn[41] to find which residue and AAindex property will contribute most to the discrimination of TyrRS helix backbone disruption. The top largest 10 $f_$values are from residue 158 and have $P < 0.001$. The 10 AAindex types are related to the intrinsic secondary structure propensities of the amino acids (Table 2), which implies that residue 158 may

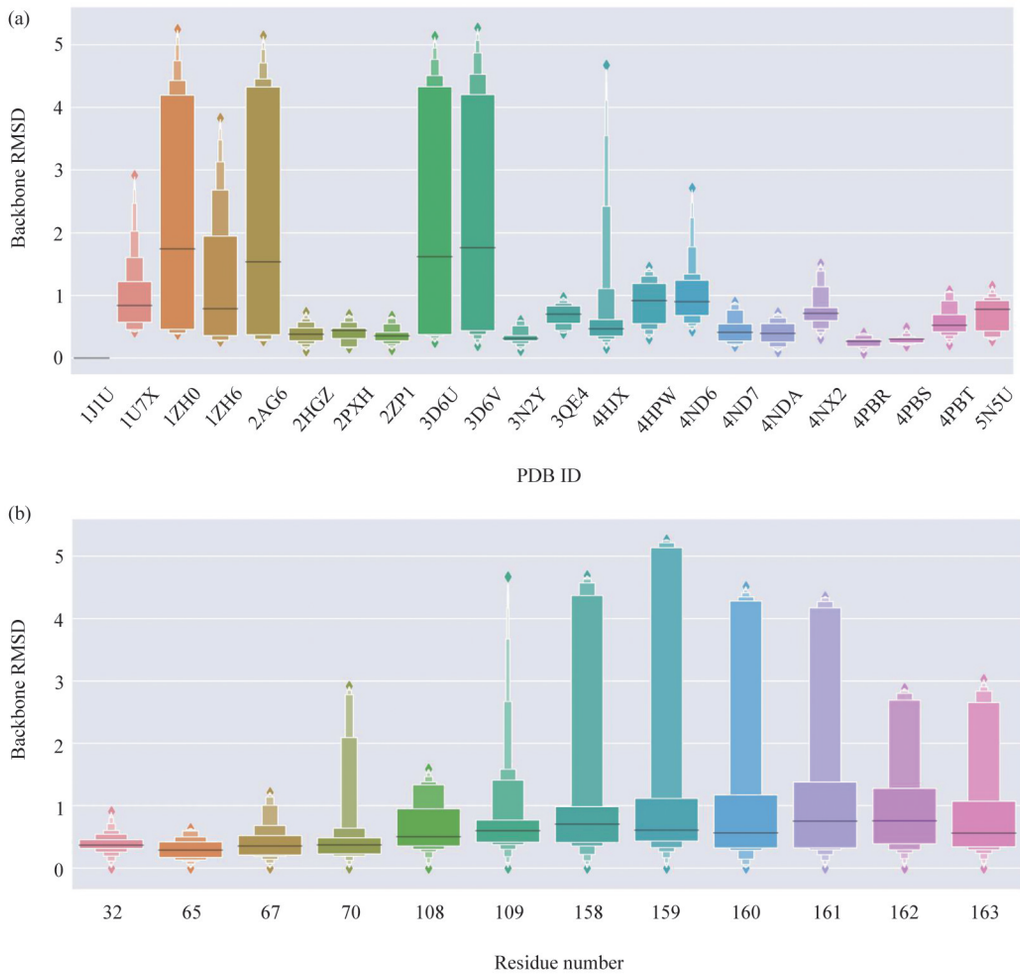


**Fig. 3　Crystal structure backbone heavy atom RMSD of UAA binding pocket residues（12 in total）between mutants（21 in total）and WT**

(a) For each mutant, backbone RMSD of pocket residues is calculated and plotted using Boxenplot showing different quantiles. (b) For each pocket residue, backbone RMSD from all mutants and WT is calculated and shown.

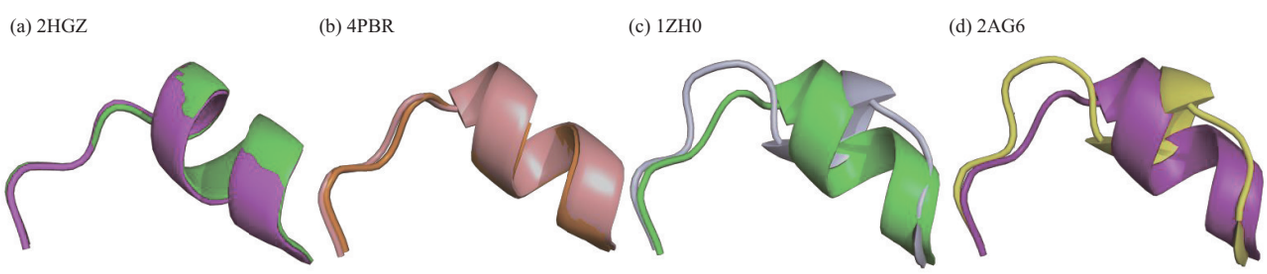


**Fig. 4　Comparison of the alpha helix containing residues 155−166 between 4 representative mutants and WT of *Mj.* TyrRS in X−ray crystal structure**

WT structure is colored in magenta (a), light pink (b), green (c) and magenta (d), respectively.

**Table 2　The top 10 AAindex types contributing most to the discrimination of TyrRS helix backbone disruption**

| Position | AAindex type | *f*_value | *P* |
|---|---|---|---|
| 158 | MUNV940104 | 27.745 47 | 0.000 063 |
| 158 | ROBB760104 | 27.262 113 | 0.000 069 |
| 158 | MUNV940105 | 26.026 208 | 0.000 089 |
| 158 | BLAM930101 | 25.948 507 | 0.000 09 |
| 158 | BUNA790103 | 25.905 193 | 0.000 091 |
| 158 | QIAN880134 | 25.607 837 | 0.000 097 |
| 158 | MUNV940101 | 24.323 819 | 0.000 126 |
| 158 | ONEK900101 | 23.726 363 | 0.000 144 |
| 158 | RACS820114 | 23.355 244 | 0.000 156 |
| 158 | QIAN880112 | 22.679 82 | 0.000 181 |

influences the secondary structure. AAindex type and value versus backbone disruption or not is shown in Figure 5. We can see that the distribution of the AAindex value is different for TyrRS with different helix backbone. Each of the 10 AAindex can be used to separate the TyrRS mutant with helix backbone disruption from others. This indicates that D158P/G mutation is a helix breaker and can disrupt the backbone conformation, which should be taken into consideration when building homology model for the TyrRS mutants to obtain accurate structure model.

## 2　Rosetta modelling of UAA−*Mj.* TyrRS mutant complex

To facilitate the TyrRS wet-lab screening process and reduce the time and resource cost, molecular modelling is carried out to predict which TyrRS mutant can recognize the target UAA. First we use Rosetta EnzymeDesign[25] to model the backbone change and AA side chain packing to accommodate specific UAA ligand for each of the 6 864 UAA-mutant pair complexes (52 UAAs times 132 TyrRS mutants). The crystal structure of WT *Mj.* TyrRS (PDB ID 1J1U) is used as input template. Representative modeling results are shown in Figure 6. For mutants with PDB ID 2HGZ and 4PBR, no obvious backbone disruption is observed in crystal structure (Table 1). In Figure 6a, the predicted binding pose for 2HGZ is accurate, while in Figure 6b, for 4PBR the predicted orientation of UAA ligand deviates from the true position in crystal structure and leads to inaccurate side chain packing of residue Lys162 and Ser158, which may be caused by wrong selection of BibaF conformation. For mutants with PDB ID 1ZH0 and 2AG6, large backbone disruption is observed in crystal structure (Table 1). Though the UAA ligand position and conformation (Figure 6c and 6d) is accurately modeled, the side chain of AA surrounding the binding pocket deviates a lot from the crystal structure mainly due to inaccurate modeling of the alpha helix with residues 158−163. The results

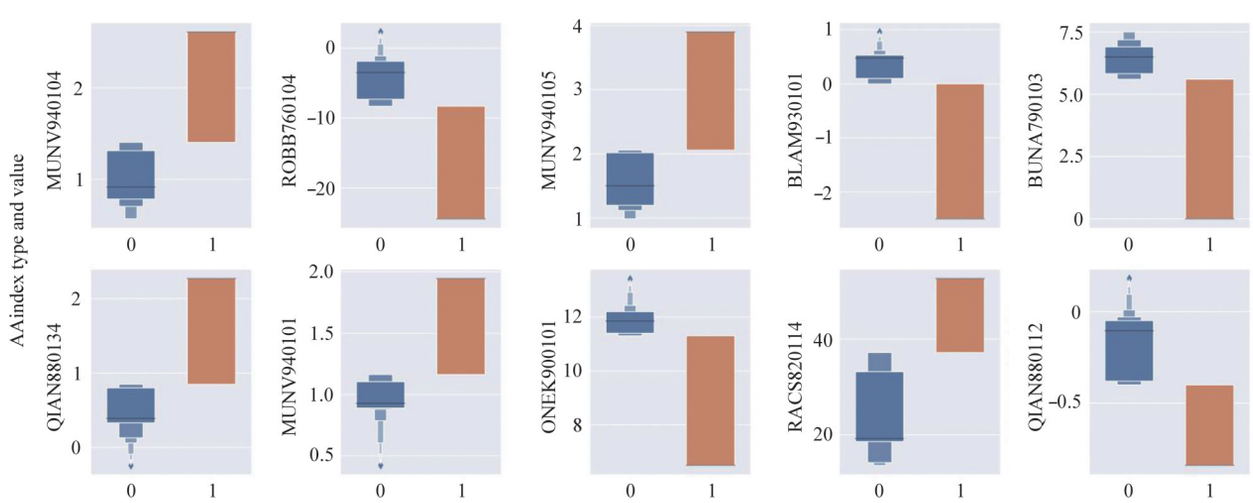


**Fig. 5　Boxplot of AAindex type and value versus backbone disruption or not**

Each subplot shows a different AAindex type. There are 17 data points in each subplot, which is the AAindex value of residue 158 from 17 *Mj*. TyrRS mutants. 1 is for backbone disruption while 0 is not.

indicate that for mutants with little disruption on helix 158−163, Rosetta model is accurate enough to predict the binding pose, but for mutants with large disruption, the opposite is true.

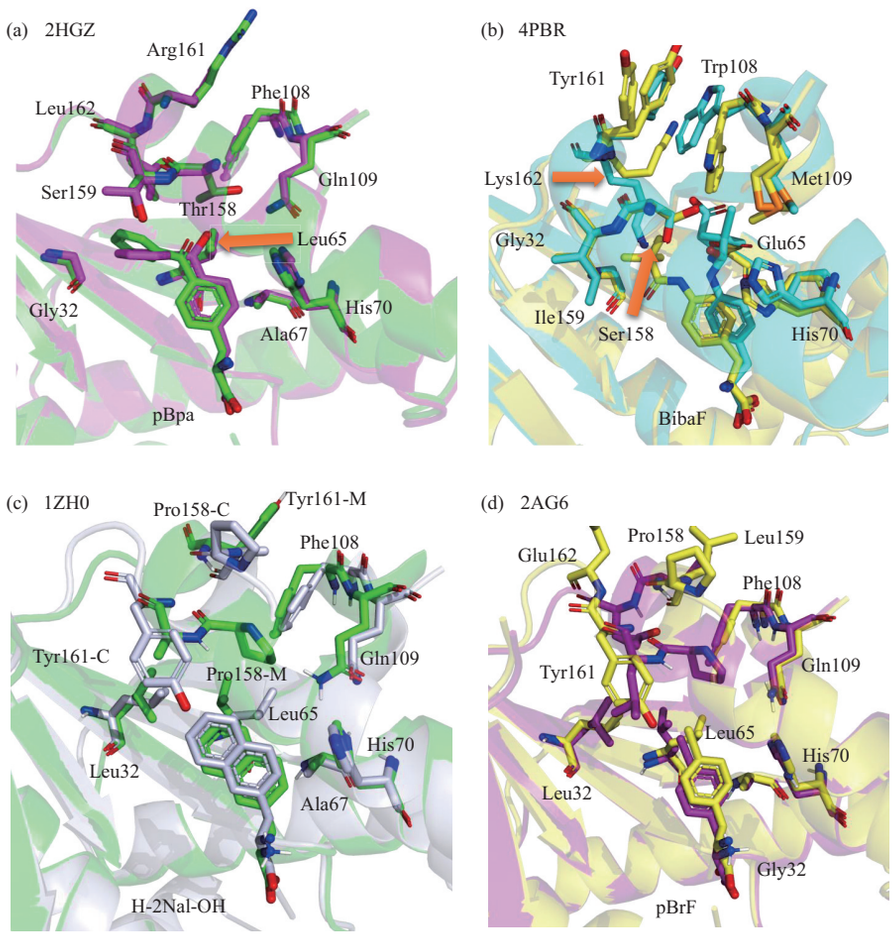


**Fig. 6 Comparison of the binding pose and binding pocket residues of 4 *Mj.* aaRS mutants from X−ray crystal structure and Rosetta model**

Corresponding UAA and binding site residues are shown as sticks. (a) PDB ID: 2HGZ. Crystal: green. Model: magenta. (b) PDB ID: 4PBR. Crystal: blue. Model: yellow. (c) PDB ID: 1ZH0. Crystal: white. Model: green. (d) PDB ID: 2AG6. Crystal: yellow. Model: magenta.

To test whether the correct backbone conformation of residues 158−163 helix having D158G/P mutation can be predicted, this segment is *de novo* remodeled using KIC loop modeling method[42] in Rosetta modelling. 1 000 models are generated for each of the 4 mutants with PDB ID 1ZH0, 1ZH6, 2AG6 and 3D6U using homology model in previous step as input. The backbone RMSD to crystal structure are calculated and plotted against total energy (Figure 7). Scatter plot in funnel shape can be observed for all 4 mutants, which indicates the amount of backbone conformation sampling is enough to find local minimum energy point. For the lowest energy structure, the RMSD to crystal structure is 3.65, 2.41, 3.74 and 4.39 Å in the 4 mutants. The lowest RMSD to crystal structure in 1 000 models of the 4 mutants is 0.94, 0.47, 0.49 and 0.48, respectively. The rmsd_to_crystal for the top 10 models ranked by total_energy_score and total_score_rank for the top 10 models ranked by rmsd_to_crystal on 4 TyrRS mutants are shown in Figure 8a and 8b, respectively. To further analyze the modeled structure, crystal structure, input Rosetta model, model with lowest RMSD to crystal, model with lowest energy are superimposed. We can see that model for 1ZH6 has the most ideal funnel plot (Figure 7b) and model with the lowest energy has least deviation from crystal structure (Figure 8d), though 2.41 Å RMSD may still not be accurate enough for UAA-TyrRS binding affinity prediction. For the remaining 3 mutants, model structure close to the native crystal structure has been generated but we can't pick it up due to the

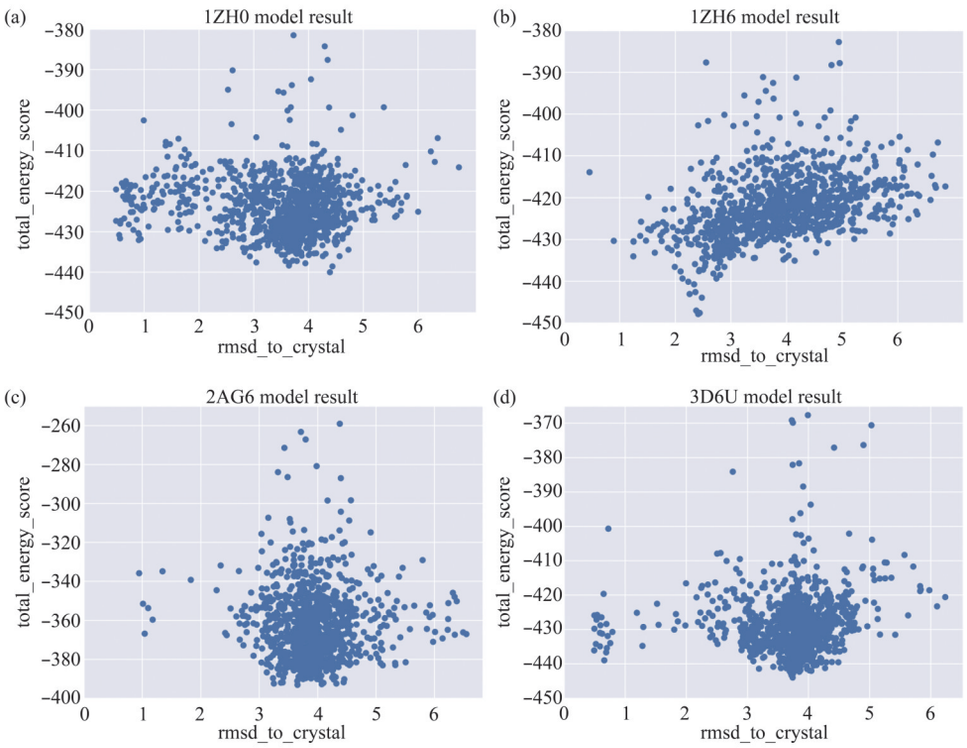


**Fig. 7　Scatter plot for rmsd_to_crystal *vs* total_energy_score in 1 000 models generated by Rosetta *de novo* loop modelling for TyrRS mutants with PDB ID 1ZH0，1ZH6，2AG6 and 3D6U**

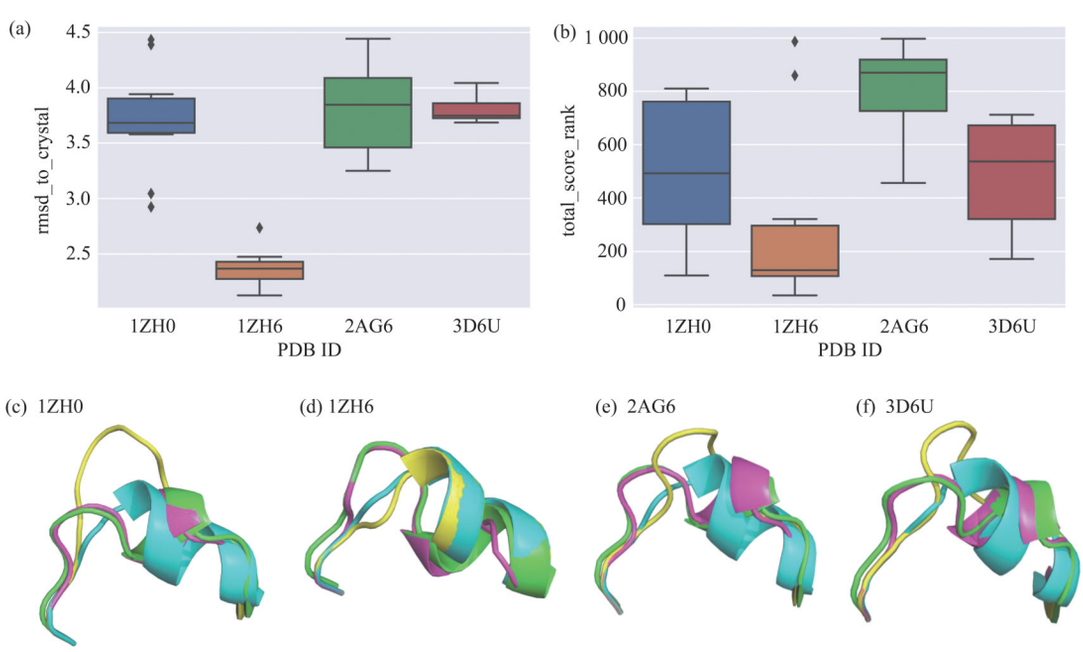


**Fig. 8　Structure analysis of Rosetta *de novo* loop modelling results on TyrRS mutants with PDB ID 1ZH0，1ZH6，2AG6 and 3D6U，respectively**

(a) Box plot of rmsd_to_crystal for the top 10 models ranked by total_energy_score on 4 TyrRS mutants. (b) Box plot of total_score_rank for the top 10 models ranked by rmsd_to_crystal on 4 TyrRS mutants. (c, d, e, f) Structure superimposition of crystal structure (in green), original Rosetta model using WT crystal structure as template (in blue), remodeled model with the lowest RMSD to crystal structure within the 1 000 models (in magenta) and remodeled model with the lowest energy within the 1 000 models (in yellow) on 4 TyrRS mutants.

error of Rosetta energy function (Figure 8).

## 3 Methods

### 3.1 Rosetta molecular modelling

Preparation of the UAA ligand was performed as described in the Rosetta tutorial for ligand preparation[43]. Briefly, the chemical structure of UAA is drawn using Marvin JS[44] and converted to smiles format. Cheminformatics tool RDKit[45] is used to generate low-energy conformations and perform energy minimization based on the smiles of UAA. The ligand parameter file used in Rosetta modelling is made by molfile_to_params.py[43].

The UAA-*Mj*. TyrRS mutant complex model is built using Rosetta EnzymeDesign[25] application with default parameters as described in Rosetta tutorial[46]. Briefly, the amino acid mutation in the mutants is written in resfile and passed into the application along with the UAA ligand parameter file. Catalytic residues of the aaRS formatting hydrogen bond with UAA is fixed using constraint file. For each complex, nstruct = 10 is used and the structure with the lowest total score is kept for further analysis.

The residues 158−163 segment of *Mj*. TyrRS is *de novo* modeled using KIC loop modeling method as described in Rosetta tutorial[47] with default parameters. 1 000 models are generated for each input structure. The RMSD between modeled structure and crystal structure is calculated by in-house script written using biopython[48].

### 3.2 Machine learning

PROFEAT, dpocket and rfscore descriptors of UAA ligand, *Mj*. TyrRS mutant protein and UAA- *Mj*. TyrRS mutant complex is calculated on http://www.descriptordb.com/ with default parameters.

Feature engineering, model selection and hyperparameter optimization are performed by PyCaret[49]. 90% of the full data is used in 5-fold CV (cross validation) and the rest 10% of data is used as test set. The feature space is transformed using 'zscore' method. 'ignore_low_variance' option is set to True and all categorical features with statistically insignificant variances are removed from the dataset. Feature selection is used and 'feature_selection_threshold' is set to 0.8.

The hyperparameters for the 15 ML models we used are listed below:

Light Gradient Boosting Machine (boosting_type = 'gbdt', class_weight=None, colsample_bytree=1.0, importance_type='split', learning_rate=0.1, max_depth =-1, min_child_samples=20, min_child_weight= 0.001, min_split_gain=0.0, n_estimators=100, n_jobs= -1, num_leaves=31, objective=None, random_state=5, reg_alpha=0.0, reg_lambda=0.0, silent=True, subsample=1.0, subsample_for_bin=200000, subsample_freq=0)

CatBoost Classifier (cat_features=None, text_features=None, sample_weight=None, baseline= None, use_best_model=None, eval_set=None, verbose =None, logging_level=None, plot=False, column_ description=None, verbose_eval=None, metric_period =None, silent=None, early_stopping_rounds=None, save_snapshot=None, snapshot_file=None, snapshot_ interval=None, init_model=None)

Extra Trees Classifier (bootstrap=False, ccp_alpha=0.0, class_weight=None, criterion= 'gini', max_depth=None, max_features= 'auto', max_leaf_ nodes=None, max_samples=None, min_impurity_ decrease=0.0, min_impurity_split=None, min_ samples_leaf=1, min_samples_split=2, min_weight_ fraction_leaf=0.0, n_estimators=100, n_jobs=None, oob_score=False, random_state=5, verbose=0, warm_start=False)

Extreme Gradient Boosting (base_score=0.5, booster= 'gbtree', colsample_bylevel=1, colsample_ bynode=1, colsample_bytree=1, gamma=0, learning_ rate=0.1, max_delta_step=0, max_depth=3, min_child_weight=1, missing=None, n_estimators= 100, n_jobs=-1, nthread=None, objective= 'binary: logistic', random_state=5, reg_alpha=0, reg_lambda= 1, scale_pos_weight=1, seed=None, silent=None, subsample=1, verbosity=0)

Logistic Regression (C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class= 'auto', n_jobs=None, penalty= 'l2', random_state=5, solver= 'lbfgs', tol=0.0001, verbose=0, warm_start=False)

Ridge Classifier (alpha=1.0, class_weight=None, copy_X=True, fit_intercept=True, max_iter=None, normalize=False, random_state=5, solver= 'auto', tol= 0.001)

Random Forest Classifier (bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion= 'gini', max_depth=None, max_features= 'auto', max_leaf_ nodes=None, max_samples=None, min_impurity_ decrease=0.0, min_impurity_split=None, min_ samples_leaf=1, min_samples_split=2, min_weight_

fraction_leaf=0.0, n_estimators=10, n_jobs=None, oob_score=False, random_state=5, verbose=0, warm_start=False)

Quadratic Discriminant Analysis (priors=None, reg_param=0.0, store_covariance=False, tol=0.0001)

K Neighbors Classifier (algorithm= 'auto', leaf_size=30, metric= 'minkowski', metric_params= None, n_jobs=None, n_neighbors=5, p=2, weights= 'uniform')

Ada Boost Classifier (algorithm= 'SAMME. R', base_estimator=None, learning_rate=1.0, n_estimators =50, random_state=5)

Gradient Boosting Classifier (ccp_alpha=0.0, criterion= 'friedman_mse', init=None, learning_rate= 0.1, loss= 'deviance', max_depth=3, max_features= None, max_leaf_nodes=None, min_impurity_decrease =0.0, min_impurity_split=None, min_samples_leaf= 1, min_samples_split=2, min_weight_fraction_leaf= 0.0, n_estimators=100, n_iter_no_change=None, presort='deprecated', random_state=5, subsample=1.0, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False)

Linear Discriminant Analysis (n_components= None, priors=None, shrinkage=None, solver= 'svd', store_covariance=False, tol=0.0001)

Decision Tree Classifier (ccp_alpha=0.0, class_weight=None, criterion= 'gini', max_depth= None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split= None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort= 'deprecated', random_state=5, splitter='best')

SVM-Linear Kernel Classifier (alpha=0.0001, average=False, class_weight=None, early_stopping= False, epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15, learning_rate= 'optimal', loss= 'hinge', max_iter=1000, n_iter_no_change=5, n_jobs=None, penalty= 'l2', power_t=0.5, random_state=5, shuffle= True, tol=0.001, validation_fraction=0.1, verbose=0, warm_start=False)

Naive Bayes (priors=None, var_smoothing= 1e-09)

## 4 ML model and performance analysis

In the previous part, we have modeled 6 864 UAA-TyrRS mutant pairs using Rosetta modelling and found that the energy function is not accurate enough to discriminate the true UAA binder from the false ones. To further improve the energy function, we use ML to train a model to learn important energy term and protein-ligand interactions that contribute most to the binding energy. The flowchart of the ML model and application is shown in Figure 9. Next each part of the flowchart will be described.

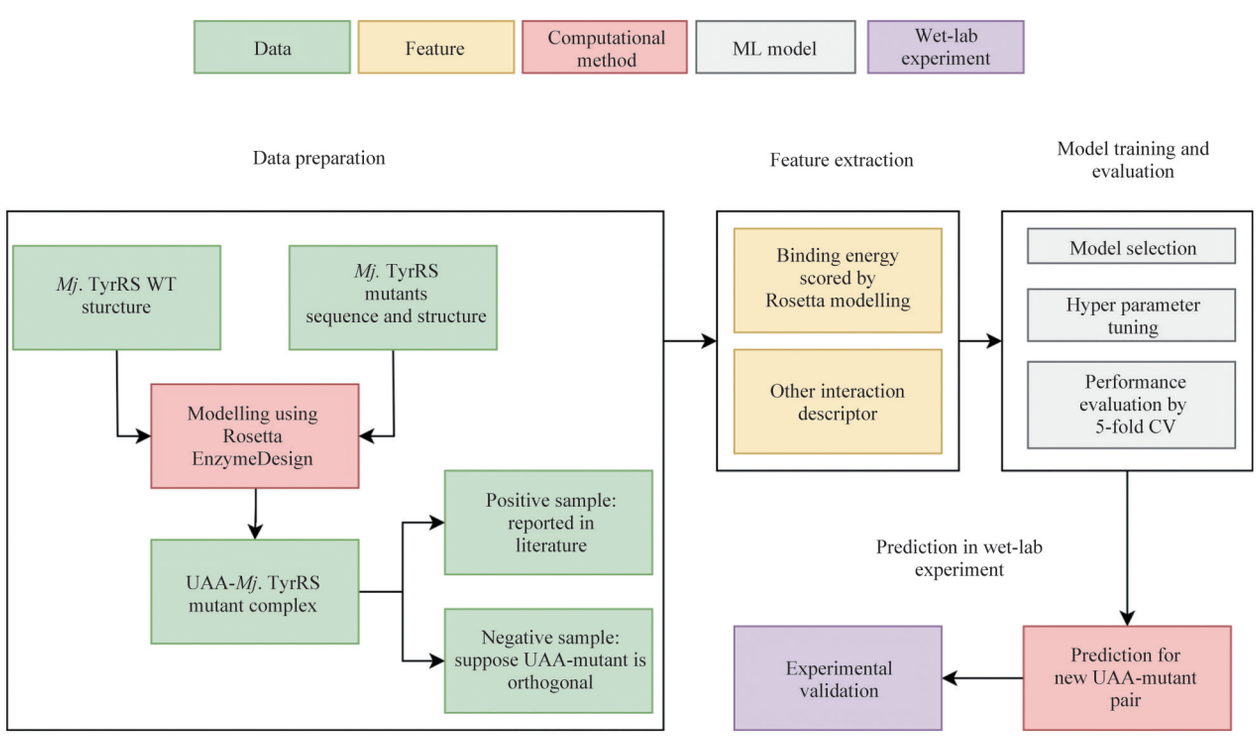


**Fig. 9 Flowchart of the data preparation， ML model training，prediction and application workflow**

### 4.1 Data preparation

We collected the chemical structure of 52 UAAs, the sequence and X-ray crystal structure of 132 TyrRS mutants from literature. Each UAA is paired to each TyrRS mutant to generate data set for ML model training. There are 6 864 UAA-TyrRS mutant pairs and the complex model is built in the previous section. For the specific UAA-mutant pair, if the UAA can be recognized by the mutant, this pair is labeled as positive and all the remaining pairs are taken as negative sample under the assumption that TyrRS mutant has high specificity and orthogonality for different UAA substrate. In the final dataset, we have 132 positive samples and 6 732 negative samples. The positive to negative ratio is 1∶51.

### 4.2 Feature extraction

The UAA-TyrRS mutant pairs can be described from 3 aspects: UAA ligand, mutant protein and UAA-mutant complex. The presentation and feature vectorization method are listed in Table 3. The feature is extracted from the chemical structure of UAA, the sequence of TyrRS mutant and the 3D structure of UAA-mutant complex model built using Rosetta modelling. The dimension of all features is 2 644. Dimension reduction was realized using PCA method. 100 dimensions were obtained while retaining 95% of the information.

### 4.3 ML model training

10% of the data is used as test set (687 samples). The remaining 90% data (6 177 samples) is used for ML model training, which is randomly splitted into training set and validation set with a ratio of 4∶1 for 5-fold CV. To overcome the problem of unbalanced positive and negative dataset, the positive samples are over sampled by 50-folds to make the dataset having 1∶1 ratio of positive and negative samples. Feature engineering, model selection and hyper parameter tuning are carried out through pycaret[49], a low-code ML library. Fifteen ML algorithms are used to train the model: Light Gradient Boosting Machine (lightGBM), CatBoost Classifier, Extra Trees Classifier, Extreme Gradient Boosting, Logistic Regression, Ridge Classifier, Random Forest Classifier, Quadratic Discriminant Analysis (QDA), K Neighbors Classifier, Ada Boost Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis (LDA), Decision Tree Classifier, SVM-Linear Kernel, Naive Bayes Classifier. Different metrics are used to evaluate the binary classification model performance: accuracy, AUC, recall, precision and F1. AUC, recall, F1 score are chosen to evaluate the performance of ML model because they are not sensitive to the unbalanced ratio of positive and negative samples.

### 4.4 ML model evaluation and explanation

Different feature combinations are explored in the model training process. The feature_type_index, values of best metrics and corresponding model name having the best performance are listed in Table 4. For different metrics, the algorithm having best performance is different. In general, the performance of model using all features is better than using single type of feature (Figure 10). When using the UAA dpocket descriptor only, Quadratic Discriminant Analysis and Extra Trees Classifier models have best recall and precision, respectively, while their F1 is not the highest. LightGBM model has the best accuracy, AUC and precision in some feature combinations.

**Table 3 Feature representation method for UAA small molecule ligand and TyrRS protein receptor used in this study**

| Method | Description | Dimension | Reference |
|---|---|---|---|
| PROFEAT-ligand | Small molecule 1D and 2D descriptors | 406 | [50] |
| PROFEAT-receptor | Protein sequence descriptors using amino acid biophysical properties and compositions | 1 437 | [50] |
| Rosetta energy score and decomposition | Rosetta interface_E total score and energy term (fa_atr, fa_rep, fa_sol, fa_elec, fa_pair, hbond_sr_bb, hbond_lr_bb, hbond_bb_sc, hbond_sc) decomposed into each binding pocket residue | 6+540 | [25, 51] |
| dpocket | Dpocket (describing pocket) extracts several descriptors using atom, amino acid, alpha sphere and volume information from the ligand binding pocket | 35 | [52] |
| rfscore | A machine learning-based score function for protein-ligand complex | 216 | [53] |

**Table 4　Comparison of 5−fold CV performance on different ML models and different performance metrics for UAA specificity prediction**

| Features and combination | Feature_ type_index | Model_ with_best_ accuracy | Best_ accuracy | Model_ with_best _AUC | Best_ AUC | Model_with_ best_recall | Best_ recall | Model_ with_best_ precision | Best_ precision | Model_with_ best_F1 | Best_ F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rosetta_6_score_ terms | 1 | SVM-Linear Kerne | 0.979 | CatBoost Classifier | 0.736 9 | Decision Tree Classifier | 0.061 5 | Naive Bayes | 0.176 | Quadratic Discriminant Analysis | 0.077 7 |
| Rosetta_energy_ decomposed | 2 | Extra Trees Classifier | 0.979 6 | Light Gradient Boosting Machine | 0.787 4 | Naive Bayes | 0.807 7 | Random Forest Classifier | 0.400 0 | Decision Tree Classifier | 0.146 5 |
| UAA_pocket_ descripotr | 3 | Extra Trees Classifier | 0.979 9 | CatBoost Classifier | 0.804 2 | Quadratic Discriminant Analysis | 1.000 0 | Extra Trees Classifier | 0.716 7 | Extra Trees Classifier | 0.161 2 |
| Rosetta_6_score_ terms+energy_ decomposed | 4 | Light Gradient Boosting Machine | 0.980 1 | CatBoost Classifier | 0.819 3 | Linear Discriminant Analysis | 0.184 6 | K Neighbors Classifier | 0.650 0 | Linear Discriminant Analysis | 0.183 6 |
| All_features_above | 5 | Light Gradient Boosting Machine | 0.979 4 | CatBoost Classifier | 0.822 2 | Naive Bayes | 0.476 9 | Light Gradient Boosting Machine | 0.400 0 | Linear Discriminant Analysis | 0.165 4 |

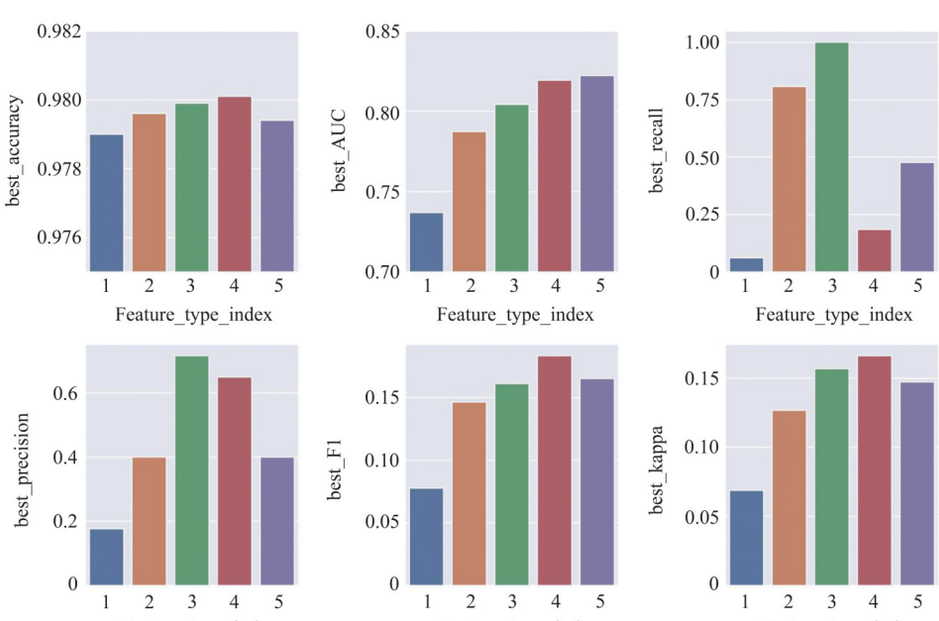


**Fig. 10　Barplot of the best model performance on different feature combinations and different metrics**

Feature_type_index is the same as that in Table 4.

The metrics to evaluate the model performance of 15 different ML models are shown in Figure 11. For AUC metric, the performances of ML models have no significant difference. For precision metric, Extra Trees Classifier and LightGBM models have relatively higher performance. For recall metric, Naive Bayes method has higher performance. For F1 score, LDA and Decision Tree Classifier models have higher performance.

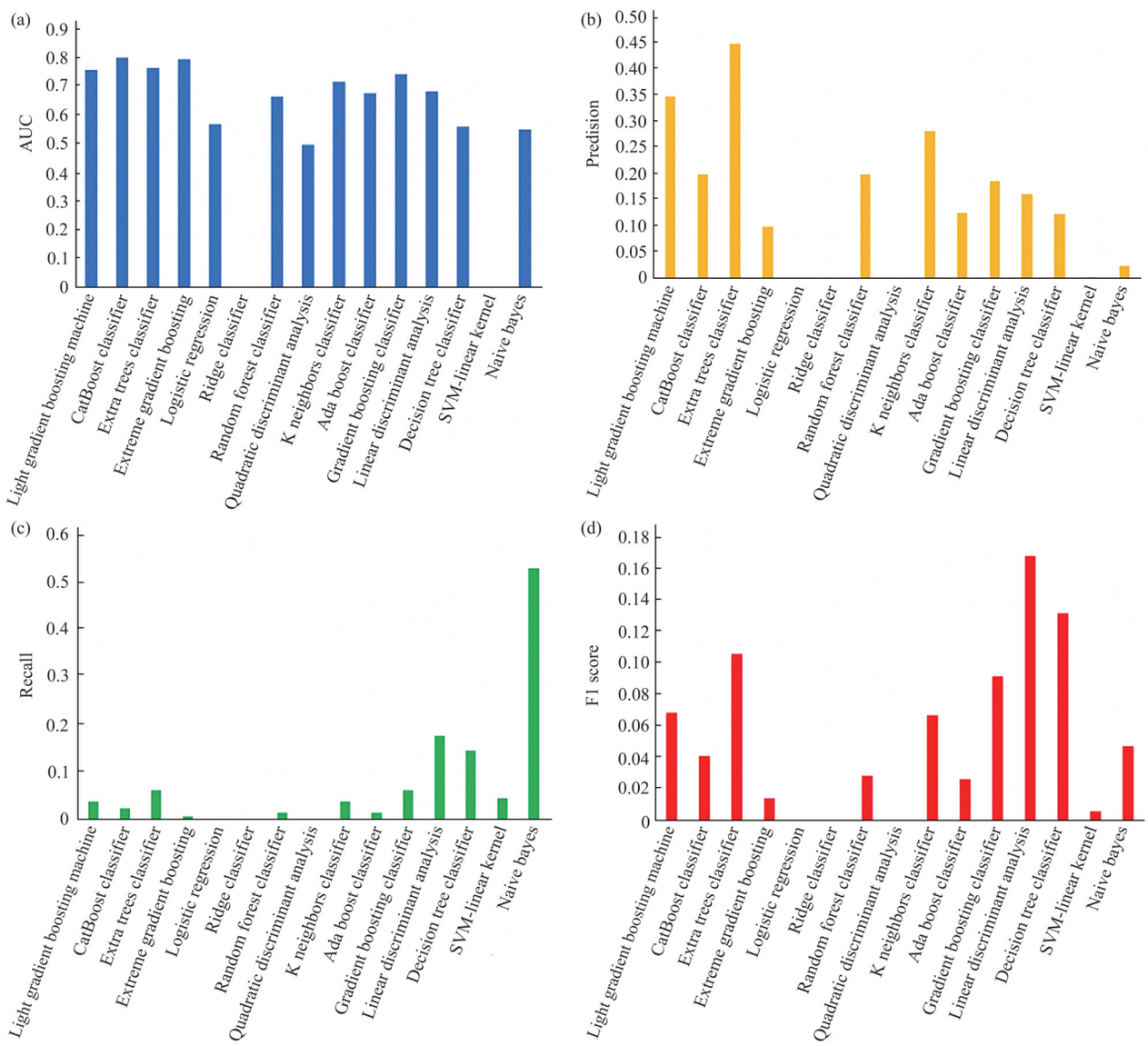


**Fig. 11　Performance of different ML models**

The AUC, precision, recall and F1 score metrics are shown for 15 types of different ML models in a, b, c and d, respectively.

We choose lightGBM[54] model for further analysis since it's a tree-based ensemble model which can avoid overfitting and has been widely used in other ML model applications. Feature importance of the model can be easily explained. The performance of lightGBM model on 5-fold CV and test set is compared (Table 5). We can see the metrics are better on test set than that on 5-fold CV.

**Table 5　Performance of lightGBM model on 5−fold cross validation set and test set using all features (feature_type_index 5)**

| LightGBM | Accuracy | AUC | Recall | Precision | F1 score |
|---|---|---|---|---|---|
| 5-fold CV | 0.979 4 | 0.759 3 | 0.038 5 | 0.35 | 0.068 6 |
| Test set | 0.979 6 | 0.854 9 | 0.066 7 | 1 | 0.125 |

SHAP method[55] is used to calculate the feature importance and its contribution to the binary class label (Figure 12). According to the result, residue_70_fa_sol and residue_34_fa_sol are considered as the most important features by the lightGBM model which indicates that the solvation energy of residue 34 and 70 may be important. Rosetta_interface_E and residue_158_fa_elec are also important which is in accordance with our domain knowledge that score function is useful to discriminate the positive sample to some extent.

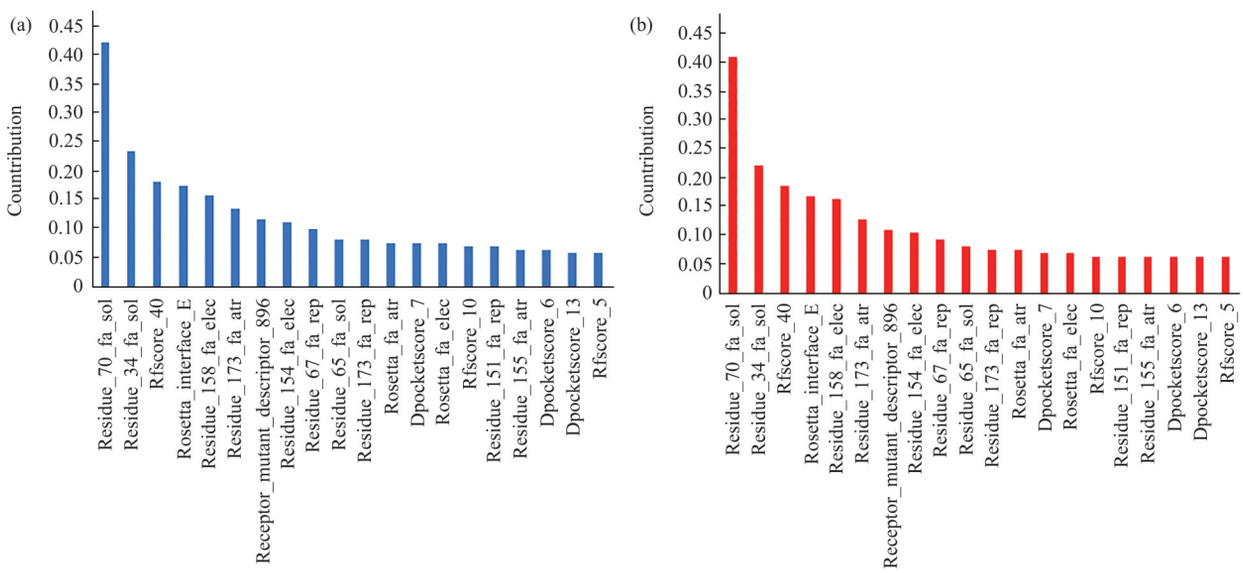


**Fig. 12 SHAP values of the important features from the lightGBM model**

The SHAP value, which is a measurement of average impact/contribution on model output of negative (a) and positive (b) samples for the important features are shown. The impact on negative label is in blue and the impact on positive label is in red. The features are sorted by the SHAP value.

To compare the prediction accuracy between lightGBM model and Rosetta score, ROC and PR curve on test set (Figure 13a and 13b) are plotted for both methods. Better ROC and PR curve are observed for lightGBM ML model. AUC of lightGBM model is 0.84, higher than 0.77 of Rosetta score. As a simulation test for real wet-lab experiment, we calculate when the number of mutants k for web-lab experiment is given and fixed, and how many true positive samples exist in the k mutants (Figure 13c and 13d). For example, if we want to test 50 mutants in wet-lab experiment out of 687 mutants, there will be 1, 2 and 11 true positive mutants for random sample, Rosetta score prediction and lightGBM model prediction, respectively. The success rate is 2%, 4% and 22%, respectively, which means Rosetta score prediction has 2-fold elevation on enrichment ratio of true positive mutants, while lightGBM model has 11-fold elevation. The TyrRS mutant screening will significantly benefit from the improvement of prediction accuracy using ML model.

## 5 Discussion

### 5.1 Significance of the work for genetic code expansion and computational protein design

In the field of computational protein design, it is a great challenge to introduce mutations into proteins to change the substrate specificity for different ligands in a particular ligand-receptor complex system. There have been many reports, such as using Rosetta[56] and OSPREY[57] to design proteins to switch the substrate and accommodate specific ligands. As a model system, *Mj.* TyrRS has be mutated and designed to bind UAA with different chemical structures as reported before[58]. This system can be used as a benchmark for substrate-specific protein design, as a large number of mutants have been reported. Comprehensive study of this system will be of great significance to the field of protein design.

This work focuses on UAA-*Mj.* TyrRS complex system which has been widely used in genetic code

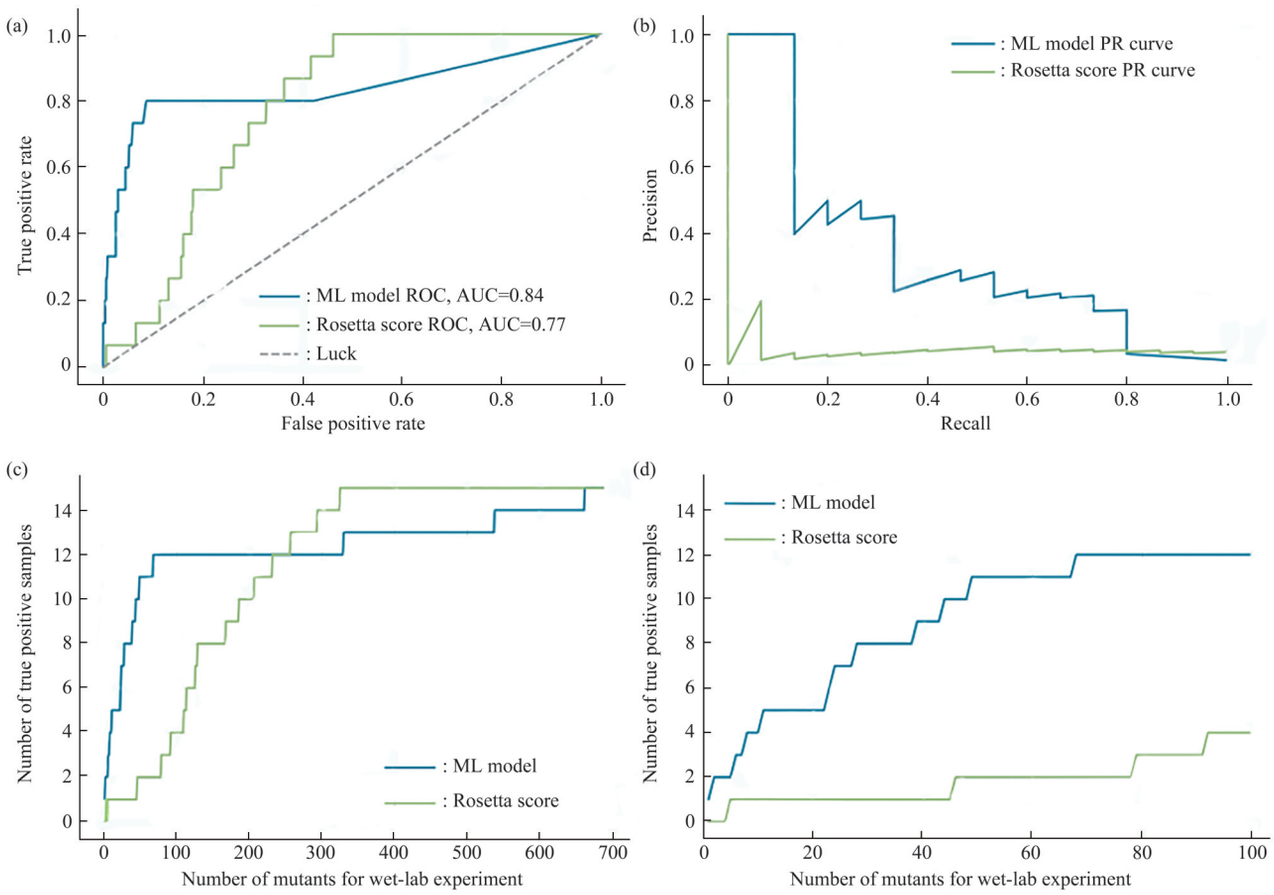


**Fig. 13 Comparison of lightGBM ML model and Rosetta score predictions**

(a) Receiver operating characteristic curve, (b) Precision-recall curve and (c, d) Number of true positive samples *vs* number of mutants selected for wet-lab experiment of lightGBM model and Rosetta interface_E score only prediction on test set. The scale of *x* axis is limited to 0−100 in (d).

expansion research field. Through the efforts of many research groups over 15 years in the past, a large number of UAA and its corresponding TyrRS mutants have been identified after spending a lot of human and material resources, so far, there is no systematic study to integrate these data for the development and test of UAA virtual screening method. This paper summarizes all the reported UAA and *Mj*. TyrRS mutants, systematically analyzes them, and attempts to establish a relatively accurate model to predict the UAA-specific recognition of different mutants, establishing the foundation for large-scale, high-throughput virtual screening.

The existing methods for predicting protein ligand interaction are probably not suitable for this system. The reason is that the adaptability and accuracy of score functions for different ligand-receptor systems are different. For example, the importance of hydrogen bond in one system may be higher than another. Method developed for affinity prediction of drug-target protein interaction cannot be applied on the *Mj*. TyrRS system without modification. In this paper, we calibrate the Rosetta score function for this specific system to get better prediction performance. Besides Rosetta molecular modeling, other methods such as molecular docking, MM-PBSA[59], TI[60], and FEP[61] can also be integrated with ML in the similar way to give more accurate prediction results in the future.

### 5.2 Previous work for TyrRS mutant substrate specificity prediction and computational library design

Several methods aiming to accelerate the screening of aaRS mutants or design more focusing library using computational chemistry method such as molecular modeling and MM-PBSA[62-64] have been reported. These results show that molecular modeling

can be used to predict TyrRS substrate specificity, but the amount of UAA and TyrRS mutants used in previous studies is too small to give convincing conclusion. In this work, we considered all UAA and TyrRS mutants reported before. With thousands of UAA-mutant pairs as training and validation data, a more solid conclusion can be made that molecular modeling integrated with ML can be useful in the virtual screening of TyrRS mutant.

### 5.3 Discussion on the prediction result

Though the backbone disruption of helix 158−163 can't be precisely modeled, we think its influence on the prediction accuracy may not be large because there are only 58 D158G mutants and 13 D158P mutants with a fraction of 42% in total 171 mutants (Supplementary file Table S1). Besides, descriptors extracted from protein sequence are not affected by the backbone disruption of structure. Knowledge and information learned from mutant sequence by ML model could correct the error in the homology model and increase the prediction accuracy.

### 5.4 Limitation of the work

On the one hand, for the homology modeling of *Mj.* TyrRS mutants, the position of amino acid side chain and rotamer in the binding pocket can be recovered well, but for those with large backbone changes, it is hard to model the backbone conformation aligned well to crystal structure. Though the close conformation can be sampled by Rosetta *de novo* loop remodeling, the score function is not precise enough to pick it out. The inaccuracy of structure prediction decreases the prediction accuracy of UAA-specific mutant. On the other hand, there are water molecules in the UAA binding pocket of *Mj.* TyrRS. The influence of water is not considered by current molecular modeling method.

The generalization ability of ML model may be low because the number of reported UAA and mutant pair is small compared to the huge chemical space of UAA and sequence space of *Mj.* TyrRS mutants. Currently it is difficult to cover the diversified chemical structure and protein sequence. One of the solutions is deep mutational scanning[65], which uses next generation sequencing to get phenotype of over 1 million mutants in a single experiment and generates enough data for ML and deep learning model training and prediction. Still, wet-lab experiments are needed to validate the model for practical usage.

## 6 Conclusion

To get further knowledge of the *Mj.* TyrRS structure and accelerate the time-costly screening process of mutant for specific UAA, we collected all the UAAs and *Mj.* TyrRS reported before, analyzed the structure and sequence difference between the mutants and found that some mutants have alternative backbone conformation on alpha helix residue 158−163 with D158G/P mutation, which makes accurate mutant modelling more difficult. Rosetta modeling and ML are integrated to give more accurate prediction results for mutant selectivity towards different UAAs. Different feature combinations and ML algorithms are tested for higher model performance. LightGBM model is chosen and the feature importance and the contribution to the binary class label is calculated to explain the knowledge learned by the model. After the calibration of Rosetta score function using lightGBM model, the enrichment ratio of target mutant is elevated by 11-fold compared with random mutation. Wet-lab experiment is in progress to validate the model. We anticipate that this proof-of-concept workflow will be of great help in the screening of *Mj*. TyrRS and protein design field.

**Supplementary material** PIBB20200425_TableS1.xlsx is available at paper online (http://www.pibb.ac.cn, http://www.cnki.net)

## References

[1] Chin J W. Expanding and reprogramming the genetic code of cells and animals. Annu Rev Biochem. 2014, **83**: 379-408

[2] Yang F, Yu X, Liu C, *et al*. Phospho-selective mechanisms of arrestin conformations and functions revealed by unnatural amino acid incorporation and (19)F-NMR. Nat Commun. 2015, **6**: 8202

[3] Li F, Shi P, Li J, *et al*. A genetically encoded 19F NMR probe for tyrosine phosphorylation. Angew Chem Int Ed Engl. 2013, **52**(14): 3958-3962

[4] Yokoyama K, Uhlin U, Stubbe J. Site-specific incorporation of 3-nitrotyrosine as a probe of pk(a) perturbation of redox-active tyrosines in ribonucleotide reductase. J Am Chem Soc. 2010, **132**(24): 8385-8397

[5] Ugwumba I N, Ozawa K, Xu Z Q, *et al*. Improving a natural enzyme activity through incorporation of unnatural amino acids. J Am Chem Soc. 2011, **133**(2): 326-333

[6] Drienovska I, Roelfes G. Expanding the enzyme universe with genetically encoded unnatural amino acids. Nat Catal, 2020, **3**:193-202

[7] Liu X H, Kang F Y, Hu C, *et al*. A genetically encoded photosensitizer protein facilitates the rational design of a miniature photocatalytic $CO_2$-reducing enzyme. Nat Chem, 2018, **10**(12): 1201-1206

[8] Drienovska I, Alonso-Cotchico L, Vidossich P, *et al*. Design of an enantioselective artificial metallo-hydratase enzyme containing an unnatural metal-binding amino acid. Chem Sci, 2017, **8**(10): 7228-7235

[9] Li Q, Chen Q, Klauser P C, *et al*. Developing covalent protein drugs *via* proximity-enabled reactive therapeutics. Cell, 2020, **182**(1): 85-97

[10] Vyas V K, Ukawala R D, Ghate M, *et al*. Homology modeling a fast tool for drug discovery: current perspectives. Indian J Pharm Sci, 2012, **74**(1): 1-17

[11] Senior A W, Evans R, Jumper J, *et al*. Improved protein structure prediction using potentials from deep learning. Nature, 2020, **577**(7792): 706-710

[12] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Bio, 2019, **20**(11): 681-697

[13] Meng X Y, Zhang H X, Mezei M, *et al*. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des, 2011, **7**(2): 146-157

[14] Wu Z, Kan S B J, Lewis R D, *et al*. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci USA, 2019, **116**(18): 8852-8858

[15] Silva D A, Yu S, Ulge U Y, *et al*. De novo design of potent and selective mimics of IL-2 and IL-15. Nature, 2019, **565**(7738): 186-191

[16] Sesterhenn F, Yang C, Bonet J, *et al*. *De novo* protein design enables the precise induction of RSV-neutralizing antibodies. Science, 2020, **368**(6492): easy5051

[17] Li R F, Wijma H J, Song L, *et al*. Computational redesign of enzymes for regio- and enantioselective hydroamination. Nat Chem Biol, 2018, **14**(7): 664-670

[18] Zhavoronkov A, Ivanenkov Y A, Aliper A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol, 2019, **37**(9): 1038-1040

[19] Ruffolo J A, Guerra C, Mahajan S P, *et al*. Geometric potentials from deep learning improve prediction of CDR H3 loop structures. Bioinformatics, 2020, **36**(Suppl_1): i268-i275

[20] Graves J, Byerly J, Priego E, *et al*. A review of deep learning methods for antibodies. Antibodies(Basel), 2020, **9**(2): 12

[21] Xiong P, Hu X H, Huang B, *et al*. Increasing the efficiency and accuracy of the ABACUS protein sequence design method. Bioinformatics, 2020, **36**(1): 136-144

[22] Wang C, Zhang Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. J Comput Chem, 2017, **38**(3): 169-177

[23] Wang J X, Cao H L, Zhang J Z H, *et al*. Computational protein design with deep learning neural networks. Sci Rep, 2018, **8**(1): 6349

[24] Cao H, Wang J, He L, *et al*. DeepDDG: predicting the stability change of protein point mutations using neural networks. J Chem Inf Model, 2019, **59**(4): 1508-1514

[25] Richter F, Leaver-Fay A, Khare S D, *et al*. *De novo* enzyme design using Rosetta3. PLoS One, 2011, **6**(5): e19230

[26] Luo X, Fu G, Wang R E, *et al*. Genetically encoding phosphotyrosine and its nonhydrolyzable analog in bacteria. Nat Chem Biol, 2017, **13**(8): 845-849

[27] Lilkova E. The pyMOL molecular graphics system, Version 1.8. New York: LLC, 2015

[28] Kobayashi T, Nureki O, Ishitani R, *et al*. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. Nat Struct Biol, 2003, **10**(6): 425-432

[29] Zhang Y, Wang L, Schultz P G, *et al*. Crystal structures of apo wild-type *M*. jannaschii tyrosyl-tRNA synthetase (TyrRS) and an engineered TyrRS specific for O-methyl-L-tyrosine. Protein Sci, 2005, **14**(5): 1340-1349

[30] Turner J M, Graziano J, Spraggon G, *et al*. Structural plasticity of an aminoacyl-tRNA synthetase active site. Proc Natl Acad Sci USA, 2006, **103**(17): 6483-6488

[31] Liu W, Alfonta L, Mack A V, *et al*. Structural basis for the recognition of para-benzoyl-L-phenylalanine by evolved aminoacyl-tRNA synthetases. Angew Chem Int Ed Engl, 2007, **46**(32): 6073-6075

[32] Turner J M, Graziano J, Spraggon G, *et al*. Structural characterization of a p-acetylphenylalanyl aminoacyl-tRNA synthetase. J Am Chem Soc, 2005, **127**(43): 14976-14977

[33] Young D D, Young T S, Jahnz M, *et al*. An evolved aminoacyl-tRNA synthetase with atypical polysubstrate specificity. Biochemistry, 2011, **50**(11): 1894-1900

[34] Tippmann E M, Liu W, Summerer D, *et al*. A genetically encoded diazirine photocrosslinker in *Escherichia coli*. Chembiochem, 2007, **8**(18): 2210-2214

[35] Xie J, Liu W, Schultz P G. A genetically encoded bidentate, metal-binding amino acid. Angew Chem Int Ed Engl, 2007, **46**(48): 9239-9242

[36] Cooley R B, Feldman J L, Driggers C M, *et al*. Structural basis of improved second-generation 3-nitro-tyrosine tRNA synthetases. Biochemistry, 2014, **53**(12): 1916-1924

[37] Wang J, Zhang W, Song W, *et al*. A biosynthetic route to photoclick chemistry on proteins. J Am Chem Soc, 2010, **132**(42): 14812-14818

[38] Liu X, Jiang L, Li J, *et al*. Significant expansion of fluorescent protein sensing ability through the genetic incorporation of superior photo-induced electron-transfer quenchers. J Am Chem Soc, 2014, **136**(38): 13094-13097

[39] Cooley R B, Karplus P A, Mehl R A. Gleaning unexpected fruits from hard-won synthetases: probing principles of permissivity in non-canonical amino acid-tRNA synthetases. Chembiochem, 2014, **15**(12): 1810-1819

[40] Kawashima S, Pokarowski P, Pokarowska M, *et al*. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res, 2008, **36**(Database issue): D202-D205

[41] Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. J Mach Learn Res, 2011, **12**: 2825-2830

[42] Stein A, Kortemme T. Improvements to robotics-inspired conformational sampling in rosetta. PLoS One, 2013, **8**(5): e63090

[43] Hosseinzadeh P. Preparing Ligands, https://www. rosettacommons. org/demos/latest/tutorials/prepare_ligand/prepare_ligand_tutorial. 2016

[44] Marvin P. ChemAxon marvinjs software. https://marvinjs-demo. chemaxon.com/latest/

[45] Landrum G. RDKit: Open-source Cheminformatics, http://www. rdkit.org. 2020

[46] Richter F. Enzyme Design Application, https://www. rosettacommons. org/docs/latest/application_documentation/design/enzyme-design. 2010

[47] Stein A. Next-generation Kinematic Loop Modeling and Torsion-restricted Sampling, https://www. rosettacommons. org/docs/latest/application_documentation/structure_prediction/loop_modeling/next-generation-KIC. 2015

[48] Cock P J, Antao T, Chang J T, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 2009, **25**(11): 1422-1423

[49] Ali M. PyCaret: an open source, low-code machine learning library in Python. 2020

[50] Li Z R, Lin H H, Han L Y, *et al*. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res, 2006, **34**(Web Server issue): W32-W37

[51] Alford R F, Leaver-Fay A, Jeliazkov J R, *et al*. The rosetta all-atom energy function for macromolecular modeling and design. J Chem Theory Comput, 2017, **13**(6): 3031-3048

[52] Schmidtke P, Le Guilloux V, Maupetit J, *et al*. fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic Acids Res, 2010, **38**(Web Server issue): W582-W589

[53] Ballester P J, Mitchell J B. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics, 2010, **26**(9): 1169-1175

[54] Ke G, Meng Q, Finley T W, *et al*. LightGBM: a highly efficient gradient boosting decision tree//Guyon I, Luxburg U V, Bengio S, *et al*. IEEE. Neural Information Processing Systems Conference. Los Angeles: Curran Associates Inc., 2017, 3149-3157

[55] Lundberg S, Erion G G, Chen H, *et al*. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell, 2020, **2**(1): 56-67

[56] Moretti R, Bender B J, Allison B, *et al*. Rosetta and the design of ligand binding sites. Methods Mol Biol, 2016, **1414**: 47-62

[57] Hallen M A, Martin J W, Ojewole A, *et al*. OSPREY 3.0: open-source protein redesign for you, with powerful new features. J Comput Chem, 2018, **39**(30): 2494-2507

[58] Chin J W. Expanding and reprogramming the genetic code. Nature, 2017, **550**(7674): 53-60

[59] Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov, 2015, **10**(5): 449-461

[60] Lee T S, Hu Y, Sherborne B, *et al*. Toward fast and accurate binding affinity prediction with pmemdGTI: an efficient implementation of GPU-accelerated thermodynamic integration. J Chem Theory Comput, 2017, **13**(7): 3077-3084

[61] Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. J Chem Inf Model, 2017, **57**(12): 2911-2937

[62] Ren W, Truong T M, Ai H W. Study of the binding energies between unnatural amino acids and engineered orthogonal tyrosyl-tRNA synthetases. Sci Rep, 2015, **5**: 12632

[63] Opuu V, Nigro G, Schmitt E, *et al*. Adaptive landscape flattening allows the design of both enzyme: substrate binding and catalytic power. PLoS Comput Biol, 2020, **16**(1): e1007600

[64] Baumann T, Hauf M, Richter F, *et al*. Computational aminoacyl-tRNA synthetase library design for photocaged tyrosine. Int J Mol Sci, 2019, **20**(9): 2343

[65] Fowler D M, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods, 2014, **11**(8): 801-807

# 利用机器学习提高*M. jannaschii*酪氨酰tRNA合成酶底物特异性分子建模预测的准确度*

段秉亚 孙应飞**

（中国科学院大学电子电气与通信工程学院，北京 100190）

**摘要** 设计结合不同化学结构底物的酶结合袋是一个巨大的挑战. 传统的湿实验要筛选成千上万甚至上百万个突变体来寻找对特定配体结合的突变体，此过程需要耗费大量的时间和资源. 为了加快筛选过程，我们提出了一种新的工作流程，将分子建模和数据驱动的机器学习方法相结合，生成具有高富集率的突变文库，用于高效筛选能识别特定底物的蛋白质突变体. *M. jannaschii* 酪氨酰 tRNA 合成酶（*Mj*. TyrRS）能识别特定的非天然氨基酸并催化形成氨酰 tRNA，其不同的突变体能够识别不同结构的非天然氨基酸，并且已经有了许多报道和数据的积累，因此我们使用 TyrRS 作为一个例子来进行此筛选流程的概念验证. 基于已知的多个 *Mj*. TyrRS 的晶体结构及分子建模的结果，我们发现 D158G/P 是影响残基 158~163 位 α 螺旋蛋白骨架变化的关键突变. 我们的模拟结果表明，在含有 687 个突变体的测试数据中，与随机突变相比，分子建模和打分函数计算排序可以将目标突变体的富集率提高 2 倍，而使用已知突变体和对应的非天然氨基酸数据训练的机器学习模型进行校准后，筛选富集率可提高 11 倍. 这种分子建模和机器学习相结合的计算和筛选流程非常有助于 *Mj*.TyrRS 的底物特异性设计，可以大大减少湿实验的时间和成本. 此外，这种新方法在蛋白质计算设计领域具有广泛的应用前景.

**关键词** 酪氨酰tRNA合成酶，遗传密码扩展，酶底物特异性，Rosetta，分子建模，机器学习

**中图分类号** Q559，Q518.2 **DOI**：10.16476/j.pibb.2020.0425

* 国家自然科学基金（61431017）资助项目.

** 通讯联系人.

Tel：010-62626682，E-mail：yfsun@ucas.ac.cn

收稿日期：2020-12-04，接受日期：2021-02-09