nature communications

Article

https://doi.org/10.1038/s41467-025-61952-2

Machine learning-guided evolution of pyrrolysyl-tRNA synthetase for improved incorporation efficiency of diverse noncanonical amino acids

Received: 12 November 2024

Accepted: 7 July 2025

Published online: 19 July 2025

Check for updates

Qunfeng Zhang[1,3], Ling Jiang[1,2,3], Yadan Niu[1], Yujie Li[1], Wanyi Chen[1], Jingxi Cheng[1], Haote Ding[1], Binbin Chen[1,2], Ke Liu[1], Jiawen Cao[1], Junli Wang[1], Shilin Ye[1], Lirong Yang[1,2], Jianping Wu[1,2], Gang Xu[1], Jianping Lin[1] & Haoran Yu[1,2]

The pyrrolysyl-tRNA synthetase (PylRS) is widely used to incorporate noncanonical amino acids (ncAAs) into proteins. However, the yields of most ncAA-containing protein remain low due to the limited activity of PylRS variants. Here, we apply machine learning to engineer the tRNA-binding domain of PylRS. The FFT-PLSR model is first applied to explore pairwise combinations of 12 single mutations, generating a variant Com1-IFRS with an 11-fold increase in stop codon suppression (SCS) efficiency. Deep learning models ESM-1v, Mutcompute, and ProRefiner are then used to identify additional mutation sites. Applying FFT-PLSR on these sites yields a variant Com2-IFRS showing a 30.8-fold increase in SCS efficiency, and up to 7.8-fold improvement in the catalytic efficiency (k_{cat}/K_m^{tRNA}). Transplanting these mutations into 7 PylRS-derived synthetases significantly improves the yields of proteins containing 6 types of ncAAs. This paper presents improved PylRS variants and a machine learning framework for optimizing the enzyme activity.

Genetic code expansion enables site-specific incorporation of noncanonical amino acids (ncAAs) into proteins by suppressing an amber stop codon (UAG) using a suppressor tRNA paired with an engineered aminoacyl-tRNA synthetase (aaRS)[1]. The pyrrolysyl-tRNA synthetase (PylRS)/$tRNA^{Pyl}_{CUA}$ pair from *Methanosarcina barkeri* and *Methanosarcina mazei* is one of the most widely used systems for incorporating ncAAs in bacteria and eukaryotes[2]. Over 300 ncAAs have been incorporated into proteins by utility of *Mm*PylRS and *Mb*PylRS variants[3]. However, the incorporation efficiency for most ncAAs remains low, leading to reduced expression of ncAA-containing proteins compared to the wild-type proteins. A powerful directed evolution strategy involving positive and negative selection has been developed to evolve PylRS for new or more efficient ncAA incorporation[4]. Despite its effectiveness, the method is time- and labor-intensive, requiring multiple rounds of selection for each target ncAA.

PylRS catalyzes a two-step reaction process[5]. In the first step, pyrrolysine is activated through adenylation with ATP. In the second step, the resulting aminoacyl-adenylate acts as a substrate for the formation of aminoacyl-$tRNA^{Pyl}$. *Mm*PylRS or *Mb*PylRS is organized into two conserved domains connected by a variable linker, including an N-terminal domain (NTD) of around 90 residues and a C-terminal domain (CTD) of around 270 residues, which harbors the catalytic sites. The tRNA-binding domain (TBD) includes NTD, the linker, and part of CTD, around 240 amino acids in total. Both TBD and the

[1] Institute of Bioengineering, College of Chemical and Biological Engineering, Zhejiang University, Hangzhou, Zhejiang 310058, China. [2] ZJU-Hangzhou Global Scientific and Technological Innovation Centre, Hangzhou, Zhejiang 311200, China. [3] These authors contributed equally: Qunfeng Zhang, Ling Jiang. e-mail: yuhaoran@zju.edu.cn

catalytic domain (CD), the remaining part in CTD, are required to create a functional PylRS/tRNA^{Pyl} pair[6]. When carrying out the directed evolution of PylRS for the incorporation of structurally diverse ncAAs, mutations were commonly constructed in the CD to expand the substrate scope. Although mutations in the TBD are not directly involved in catalysis, they have been shown to affect the efficiency of ncAA incorporation in variants of *Mm*PylRS, *Mb*PylRS, and their chimeras. For example, a chimeric PylRS variant, IPYE (V31I/T56P/H62Y/A100E), was obtained through a phage-assisted continuous evolution (PACE) selection, and exhibited a 45.2-fold improvement in catalytic efficiency ($k_{cat}/K_m{}^{tRNA}$) compared to the parent enzyme[7]. Transplanting the evolved mutations into *Mb*PylRS AcK3RS and *Mm*PylRS AcK3RS variants increased expression of reporter proteins by 9.7-fold and 2.2-fold, respectively, compared to unmodified PylRS counterparts. In another study, a N-terminal mutation of *Mm*PylRS R19H/H29R/T122S was found to enhance the incorporation of ncAA and the yield of ncAA-containing proteins by almost 4-fold compared to wild-type PylRS[8]. Interestingly, since the mutations located in the TBD of PylRS were separated from its CD, they could be introduced to other PylRS mutants to improve the incorporation efficiency of their corresponding ncAAs. This makes the N-terminal engineering of PylRS a general strategy to enhance the incorporation efficiency of various ncAAs[9]. However, available N-terminal mutations of PylRS are still relatively rare, and their impact on enhancing the efficiency of ncAA incorporation remains limited.

Recently, a class of highly active PylRS enzymes lacking an NTD has been identified and characterized[10,11]. These PylRS enzymes function with their cognate ^{Pyl}tRNA and could also be used to direct the incorporation of ncAAs. One of such enzymes is the PylRS from *Methanomethylophilus alvus* (Ma), which showed a similar amino acid-binding pocket with *Mm*PylRS, and transplanting mutations from *Mm*PylRS to *Ma*PylRS expanded the substrate specificity[7,12]. *Ma*PylRS/*Ma*^{Pyl}tRNA was also engineered to be mutually orthogonal to the *Mb*/*Mm*PylRS system[11]. However, since these PylRS enzymes were only recently identified, the number of ncAAs incorporated by them was significantly lower than the number directed by *Mb*/*Mm*PylRS.

Various strategies have been developed to engineer enzymes for improved activity, including directed evolution, rational design, computational design using Rosetta, and machine learning. In the enzyme engineering process, some single mutants are combined with the goal of creating variants with enhanced performance. However, the contributions of these single mutants are affected by mutations made at other sites in the protein, which is known as epistasis[13]. When mutations that contribute positively on their own are combined into a single protein, two or more mutations often interact in a non-additive manner. Epistatic effects can decrease the efficiency of enzyme engineering by altering the shape of the protein fitness landscape, and it remains challenging to predict this kind of epistatic behavior[14]. Recently, machine learning has been demonstrated to be useful for navigating an epistatic fitness landscape that covers a small sequence space[15]. A machine learning-assisted directed evolution strategy has been developed to evolve an enzyme to enhance and invert the stereoselectivity of a putative nitric oxide dioxygenase, in which the machine learning method avoids some local fitness traps or long paths to the global optimum for the combinatorial library[16]. Additionally, an innov'SAR model was developed to predict the enantioselectivity of 512 combinatorial variants based on combinations of 9 single-point mutations for an epoxide hydrolase[17]. Innov'SAR uses the indexes of the AAindex database to encode the primary protein sequence into a numerical chain, which is then converted to a protein spectrum using Fast Fourier Transform (FFT). With the protein spectrum as the encodings of the protein sequence, along with the experimental values, regression models were then trained. Innov'SAR has also been used to predict the thermostability of combinatorial variants[18], including a limonene epoxide hydrolase and a transaminase. Despite these advancements, it remains unclear whether these machine learning methods are broadly effective for engineering different enzymes. This uncertainty arises because the protein fitness landscape can vary significantly across different proteins and even among different mutation sites within a single protein.

In this study, we apply machine learning to engineer the TBD of PylRS, aiming to develop highly improved variants. We anticipate that these variants could be broadly applicable to other catalytic domain variants, thereby enhancing the incorporation efficiency of corresponding ncAAs. A *Mm*PylRS variant named IFRS (N346I/C348S) is selected as the model for protein engineering, as it was obtained by screening libraries against 3-iodo-Phe (3IF) and also shows a broad substrate range. Supervised machine learning is first applied to predict the highly active combinatorial variants of 12 single-point mutations. The improved variant is then used as the input for three deep learning models to predict additional single variants, which are then combined again to obtain the combinatorial variant with the highest activity. The best mutations are then combined with various C-terminal variants to test the generality of these mutations in improving the incorporation efficiency of various ncAAs. We also carry out molecular dynamics (MD) simulations to explore the mechanism behind the enhanced performance of these variants.

Results

Design of combinatorial variants of PylRS using the FFT-PLSR model

Mutations in the TBD of *Mm*PylRS have been demonstrated to improve the efficiency of ncAA incorporation. It was also shown that N-terminal mutations did not significantly influence the substrate specificity of the PylRS and could be transferred to different variants[19]. We here tested four sets of mutations obtained previously in the TBD of *Mm*PylRS to improve catalytic efficiency of IFRS, which include R61K/H63Y/S193R, R19H/H29R/T122S, D2N/K3N/T56P/H62Y, and V31I/T56P/H62Y/A100E[6–8,20] (Supplementary Table 1). Based on sequence alignment, we introduced these N-terminal mutations into IFRS and tested their activity for incorporating 3-bromo-Phe (3BrF), a cheaper substrate than 3IF (Fig. 1a, Supplementary Fig. 1). Expression of IFRS was driven by the constitutive, mid-strength *E. coli* glutaminyl-tRNA synthetase (*glnS*) promoter, and expression of PylT was controlled by the *E. coli* lpp gene promoter. Amber suppression of the sfGFPS2TAG gene by 3BrF was investigated by measuring the fluorescence intensity, and the ncAA-containing protein yield was presented by the ratio of fluorescence intensity to optical density OD_{600} (Flu/OD) of cells expressing sfGFP and PylRS. The normalized protein yield was calculated by subtracting the Flu/OD ratio of cells cultured in the presence of ncAA with that of in the absence of ncAA (Supplementary Fig. 2). We found that R19H/H29R/T122S did not achieve the expected increase in stop codon suppression (SCS) efficiency, and the IPYE (V31I/T56P/H62Y/A100E) from chPylRS did not increase the SCS efficiency of IFRS either. Although R61K/H63Y/S193R in IFRS did achieve a modest increase in sfGFPS2TAG expression yield compared to the IFRS alone, the effect was still lower than that observed for the mutations in wild-type *Mm*PylRS. Interestingly, the D2N/K3N/T56P/H62Y from chPylRS increased the SCS efficiency of IFRS by around 7-fold, confirming that tRNA binding domain mutations can indeed enhance the activity of catalytic domain mutants (Supplementary Fig. 2b). We also tested the activity of 12 single-point mutations and found that only D2N, R61K, and H62Y enhanced the activity of IFRS, with D2N being the most active, exhibiting 3-fold improved SCS efficiency compared to the wild type (Fig. 1b, Supplementary Fig. 2a). In the combinatorial mutant D2N/K3N/T56P/H62Y, D2N and H62Y enhanced the SCS efficiency of IFRS, while K3N and T56P reduced the SCS efficiency, indicating a positive sign epistasis among the mutations (Supplementary Table 2).

We then attempted to use machine learning to explore the combinatorial space of these 12 single-point mutations, a total of 4096 (2^{12})

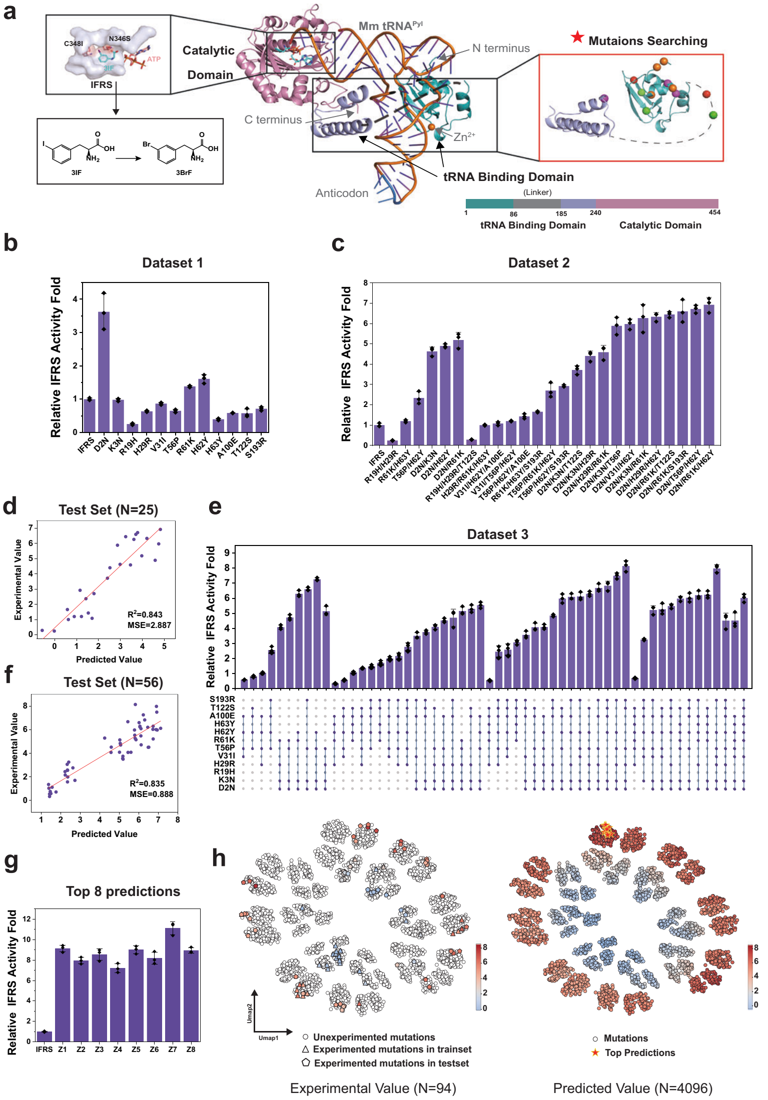

Fig. 1 | Machine learning to guide the design of high-activity combinatorial variants. **a** Structure of IFRS used for engineering TBD of PylRS, and 3BrF was selected as the substrate of IFRS. **b** Dataset 1 is composed of activities of 12 single-point mutants. **c** Dataset 2 is composed of activities of double and triple mutants used as a test set. **d** Accuracy of ML model in Dataset 2. **e** Dataset 3 is composed of activities of quadruple and multiple-point mutants used as a test set. **f** Accuracy of ML model in Dataset 3. **g** The SCS efficiency of the top 8 variants predicted by the ML model. **h** Experimental data and predicted data presented in a 2-dimensional sequence space. Error bars represent ±standard deviation of the mean over three independent replicates. Source data are provided as a Source Data file.

variants. The FFT-partial least squares regression (PLSR) approach was applied for predicting the fitness of combinatorial variants, which uses FFT for the protein sequence encoding and PLSR as the algorithm of the ML model. The FFT-PLSR model has demonstrated the ability to be trained on a small dataset of enzyme mutant activity data to accurately predict the activity across the entire combinatorial space. The variant sequences were first transferred into numerical features and then transformed into a two-dimensional energy versus frequency representation using FFT (Supplementary Fig. 3). Based on the transformed data, a PLSR model was trained to predict the activities of other mutants. We first trained the FFT-PLSR model using dataset 1, composed of 12 single variants datasets. By scoring with leave-one-out cross-validation (LOOCV), we screened 566 amino acid encodings from the AAindex database and selected the index with the best score to encode the amino acid sequences and build the PLS regression model (Supplementary Figs. 4 and 5). We constructed 6 double and 19 triple mutants, and measured their SCS efficiency to form dataset 2 as a test set ($N = 25$) (Fig. 1c). The trained model achieved an R^2 of 0.843 and an MSE of 2.887 on the test set, indicating that the model showed good prediction ability for high-activity combinatorial mutants (Fig. 1d). Interestingly, the best combinatorial mutant predicted by the model was D2N/R61K/H62Y, which was also the variant with highest activity in the test set (Supplementary Table 3).

In dataset 2, we observed epistasis between mutations. For example, T56P showed a decreased amber codon suppression efficiency, while H62Y improved the amber codon suppression efficiency compared to the IFRS. However, the T56P/H62Y showed a normalized fluorescence intensity significantly higher than that of H62Y, which led to a positive sign epistasis between T56P and H62Y. The positive sign epistasis effect was also found for R61K and H63Y, and D2N and K3N (Supplementary Table 2). By contrast, R19H and H29R showed a positive reciprocal sign epistasis (Supplementary Table 2). To help the model gain a more thorough understanding of epistasis between mutants, we added dataset 2 to the training set, which raised the total number of mutants in the training set to 38. Additionally, we rationally designed and constructed an additional test dataset 3 ($N = 56$) including 56 combinatorial mutants, mainly based on combination of improved variants (Fig. 1e). The retrained model achieved an R^2 of

0.835 on the test set, still showing high accuracy for predicting the activities of combinatorial mutants (Fig. 1f). We then constructed the top 8 mutants predicted by the model and tested their SCS efficiency (Fig. 1g and Supplementary Table 4). All eight variants showed a significant increase in the activity compared to the IFRS, with the mutant Z7 (D2N/V31I/T56P/R61K/H62Y/T122S/S193R, Com1-IFRS) exhibiting the highest SCS efficiency, which was 11-fold higher than the IFRS (Fig. 1g). We then attempted to apply the ML model to predict single-point variants in the TBD to improve the activity of Com1-IFRS. Fifteen single-point variants were predicted by the model to be more active than Com1-IFRS, with 12 of them located at previously unseen positions. However, all 15 variants failed to further improve the activity of Com1-IFRS, indicating the model's limited ability to predict effects of mutations at unseen positions (Supplementary Fig. 6). The outcome is reasonable, given that the training set was small, only containing 38 variants across the 12 positions.

We utilized the Uniform Manifold Approximation and Projection algorithm, along with one-hot encoding for sequence representation, to reduce the dimensionality of the 12 single-point mutations combinational space and visualized it as a two-dimensional scatter plot. Our analysis revealed that mutations with similar activities clustered closely together (Fig. 1h). The mutants from the training set were distributed across the entire sequence space, which was crucial for developing an accurate machine learning model (Fig. 1h). Furthermore, our model's predictions aligned well with the experimental data with high-activity variants in the outer clusters and low-activity variants in the inner clusters (Fig. 1h). This indicates that the model accurately mapped the fitness landscape, allowing us to identify mutants with significantly increased activity in the high-activity clusters.

Further improvement of IFRS activity by deep learning models

To further improve IFRS activity, we explored additional mutation sites in TBD with Com1-IFRS as the template, by employing three deep learning models that enable zero-shot prediction of high-fitness variants, including ESM-1v, MutCompute, and ProRefiner (Fig. 2a, Supplementary Fig. 7 and Table 5). EMS-1v is a protein language model that was trained on extensive datasets of protein sequences spanning the evolutionary tree of life, and hence learned the fundamental principles of protein structure and function[21]. Given the sequence of Com1-IFRS, each amino acid in the TBD region was individually masked and analyzed by the ESM-1v model to predict the impact of potential mutations at that site. We constructed and characterized 16 single-point mutants that were predicted to have a higher likelihood than wild type (Supplementary Figs. 7 and 8). MutCompute is a structure-based three-dimensional self-supervised convolutional neural network model that was trained to associate local protein micro-environments with their central amino acid[22]. Given the N-terminal and C-terminal structures of *Mm*PylRS (5UD5.pdb & 4TQD.pdb), MutCompute predicted single-point mutations optimizing the protein structure, and we characterized the top 44 mutants based on the probability predicted (Supplementary Figs. 7 and 8). ProRefiner is a global graph attention model for inverse protein folding that designs sequences compatible with a given backbone structure[23]. We restrict the candidate mutation sites to the TBD of PylRS and leverage sequence design models to compute a quality score for every candidate site. For each site to be examined, we masked this site in the sequence to get the input partial sequence, and the input backbone structure is from the Com1-IFRS structure predicted by Alphafold 3[24]. The model then predicted the identity of the masked site in the form of a probability distribution over all amino acid types, and the top 42 single-point mutants were selected for characterization (Supplementary Figs. 7 and 8). Interestingly, 7 mutants were predicted to have improved activities by both Mutcompute and ProreRiner (Supplementary Fig. 7). Based on three methods, a total of 95 single-point mutants were constructed on 85 amino acids of COM1-IFRS and assayed for enzyme activity (Fig. 2b and Supplementary Fig. 7).

Since the enzyme activity is represented by the fluorescence intensity of sfGFP incorporated with 3BrF, the maximum activity that could be detected is the fluorescence intensity of wild-type sfGFP. The fluorescence intensity of sfGFPS2TAG in the presence of Com-1-IFRS reached 45% of wild-type sfGFP. Among all the single-point variants constructed, we did obtain several variants with higher activities than the Com1-IFRS (Fig. 2b), which were I176S predicted by ESM-1v, D76A, T68V, H28K, T20S predicted by MutCompute, K67S, N7S, V74I predicted by ProRefiner, and H63N predicted by both MutCompute and ProRefiner. The best variant D76A, designed by MutCompute, showed a 31% improvement in activity compared to Com-1-IFRS. However, most of the variants designed by the three approaches exhibited reduced or even lost activity, such as W16E, showing a 99.7% decrease in activity compared to IFRS.

We then wondered if these data are useful for training a supervised ML model to predict high-fitness single-point variants across the protein. With the above single-point mutants' activity data as the training set ($N = 96$, including Com1-IFRS), the FFT-PLSR model was built. There are 566 amino acid encodings in the AAindex database, and we tested the performance of the models trained with different numbers of amino acid encodings. To optimize the amino acid encoding, we first screened the single AAindex encoding, which was then fixed for optimizing the second encoding. The third encoding was optimized with the first two encodings fixed. When screening two or three indices, the protein sequence was first subjected to FFT separately, and then the results were combined to train the model. When one index was used, the R^2 of the model was only 0.452, while when three amino acids encodings were used, the R^2 of the model increased to 0.926 (Supplementary Fig. 9). We then used the three-index model to predict fitness of all single-point mutations in the PylRS TBD region, and the top 20 variants were constructed and characterized for enzyme activity (Fig. 2b, Supplementary Fig. 10). The best variant, K67G, showed a 31.9% improvement in activity, which was even higher than D76A predicted by MutCompute and K67S predicted by ProRefiner, indicating that the FFT-PLSR model was effective in exploring sequence space (Fig. 2c).

To further explore the sequence space, we constructed the saturation mutagenesis on the nine sites where the improved variants were obtained, to build a mutability landscape (Fig. 2d). The mutability landscape is defined by the impact of all possible point mutations on protein function by substituting the native amino acid at each residue position with each of the 19 non-native amino acids, one at a time[25,26]. The mutability landscape showed that most of the improved variants were found at positions D76, H63, and K67, and the mutations at sites T20, H28, and I176 were mostly detrimental. We screened the variants that showed over a 10% improvement in SCS efficiency compared to Com1-IFRS (Fig. 2e). There were 3, 7, 5, 1, 2, 9 mutations at positions N7, H63, K67, T68, V73, D75, respectively, meeting this criterion. We hence attempted to use the FFT-PLSR model to explore this sequence space containing 11,520 ($4 \times 8 \times 6 \times 2 \times 3 \times 10$) mutations. To achieve epistasis information, the top 3 variants at each mutation site were combined to construct double variants, resulting in a total of 92 combined variants to be used for model training (Fig. 2f). Combination of improved single variants did generate further improved variants, with the best double mutant of N7E/T68F showing a SCS 70% higher than the Com1-IFRS (Fig. 2f). However, the variants with decreased activities were also observed, indicating a strong epistasis among several mutations. For example, both V74W and K67G improved the SCS efficiency, while V74W/K67G showed a significant decrease compared to the Com1-IFRS, which resulted in a strong negative reciprocal sign epistasis effect between V74W and K67G. The antagonistic effect was also observed for V74W and single variants including K67L, D76F, D76L, and D76Y (Supplementary Table 6).

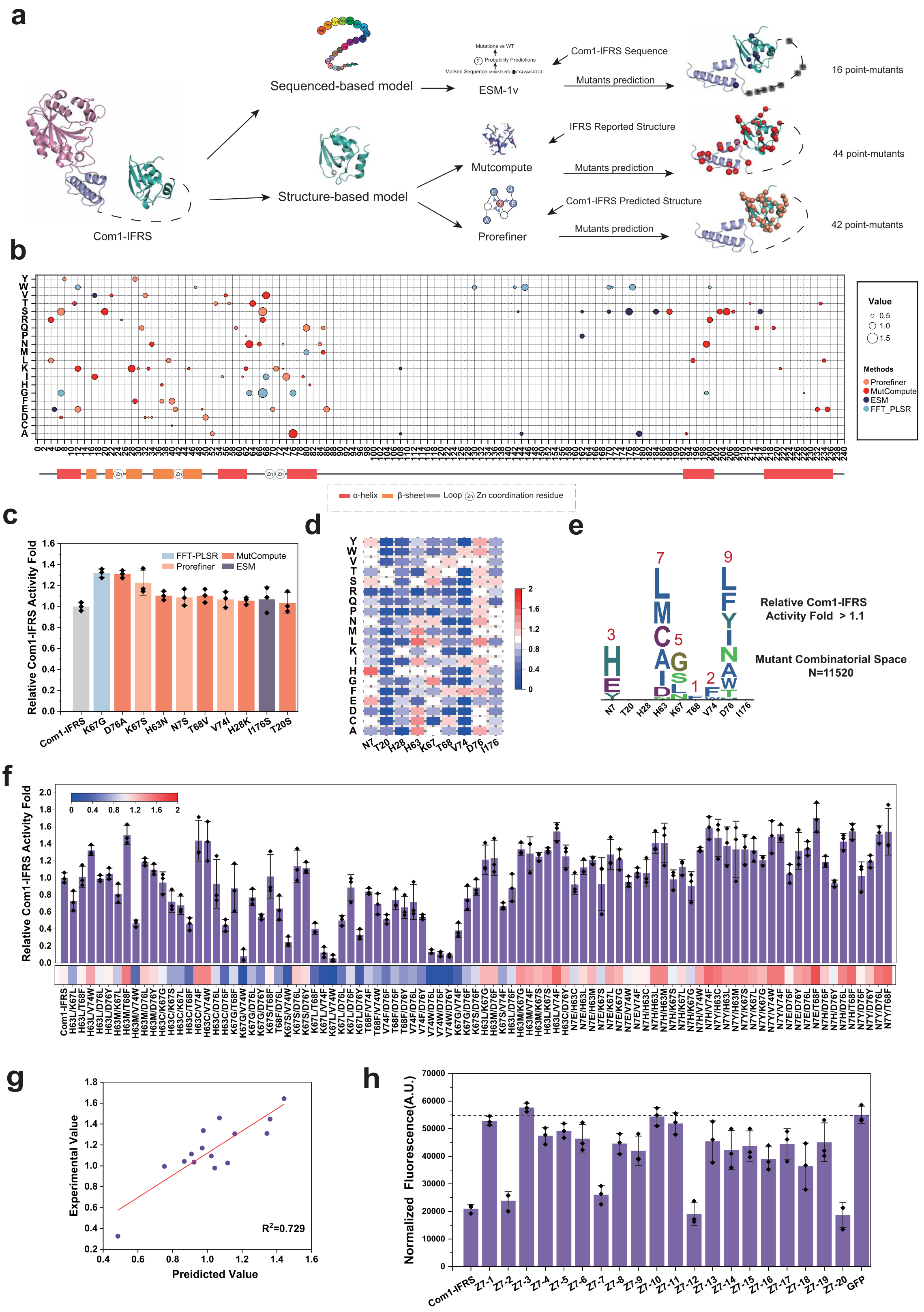

Fig. 2 | Design of variants to further improve the activity of IFRS. a Design of single variants using models including ESM-1v, MutCompute, and ProRefiner. **b** Fitness of single variants is designed. The size of the circle indicates the activity ratio of mutants to Com1-IFRS. The mutants designed by different methods are colored differently. **c** Activities of nine improved single variants. **d** The mutability landscape constructed by saturation mutagenesis at nine amino acid sites. The color bar indicates the activity of mutants relative to Com1-IFRS. **e** The sequences of mutants with at least 10% improved SCS efficiency are higher than Com1-IFRS. The height of each character indicates the relative fitness of the mutant. **f** The relative activities of double variants constructed by combining the top 3 single variants at each mutation site. The mutants include N7H, N7E, N7Y, T68F, K67G, K67S, K67L, H63L, H63M, H63C, V74F, V74W, D76L, D76F, D76Y. **g** Accuracy of the FFT-PLSR model built. The R^2 value was calculated on the test set, which consisted of 15 double-point mutants. **h** The SCS efficiency of the top 20 combinatorial variants predicted by the FFT-PLSR model. Error bars represent ±standard deviation of the mean over three independent replicates. Source data are provided as a Source Data file.

We then conducted another round of FFT-PLSR building for predicting high-activity combinatory mutations. There were 27 single-point mutations and 92 combinatorial mutations, plus Com1-IFRS, 120 data points in total in the training dataset. These were used to train the FFT-PLSR model. We first used 10-fold cross-validation to evaluate and identify the optimal single-index model. However, this model only achieved an R^2 score of 0.558 on the training set, indicating a poor performance. To improve the model performance, multiple indices were used for encoding amino acids, which improved the R^2 score to 0.668 for the two-index model and 0.677 for the three-index model, respectively (Supplementary Fig. 11 and Table 7). The retrained FFT-PLSR model achieved an R^2 of 0.729 on the test set, which consisted of 15 double-point combinatory mutants not included in the training set (Fig. 2g, Supplementary Fig. 12). Based on the three-index model, we predicted the activity data of 11520 mutations, and the top 20 combinatorial variants were selected for experimental verification (Supplementary Table 8). All the variants showed comparable or higher activities than the Com1-IFRS, with 15 of them possessing activities improved by more than 2-fold, suggesting the great effect of the ML model. The best Z7-3 variant (N7Y/H63L/K67N/V74W, Com2-IFRS) showed a 2.8-fold increase in SCS efficiency compared to Com1-IFRS, more than 30-fold higher than IFRS, reaching fluorescence intensity comparable to wild-type sfGFP (Fig. 2h). Biochemical characterization of IFRS, Com1-IFRS and Com2-IFRS using 3BrF confirmed that the catalytic efficiency (k_{cat}/K_m) of Com1-IFRS and Com2-IFRS for tRNA was improved by 1.4-fold and 5.6-fold, respectively, compared to IFRS (Supplementary Table 9). Additionally, we tested if the mutations H20S, H28K, I176N, and I176S, which were not selected for making combinatory mutations, could further improve the activity of Com2-IFRS. It was found that none of these single-point mutations improved the activity of Com2-IFRS (Supplementary Fig. 13).

Combination of tRNA-binding domain mutations with catalytic domain mutations to generally enhance the incorporation efficiency of diverse ncAAs

As the IFRS is polyspecific and could accept various phenylalanine derivatives[27], we hence tested if Com1-IFRS and Com2-IFRS improved the incorporation efficiency of other ncAA substrates. It was found that normalized fluorescence of sfGFPS2TAG incorporated with 12 diverse ncAAs was significantly increased by the two variants compared to IFRS, with the largest 101.9-fold improvement for 3FF enabled by Com2-IFRS (Fig. 3a). However, since IFRS is a promiscuous enzyme with a certain degree of misincorporation of canonical amino acids (cAAs), Com1-IFRS and Com2-IFRS also increased the incorporation efficiency of cAAs, and different extent of misincorporation was observed in presence of different ncAAs (Fig. 3a). Subtracting fluorescence intensity of sfGFP2TAG in presence of 3FF with that in absence of 3FF revealed a 3944.8-fold improvement in SCS efficiency for Com2-IFRS compared to IFRS (Supplementary Fig. 14a). Biochemical characterization of IFRS, Com1-IFRS and Com2-IFRS using 3FF confirmed that the catalytic efficiency (k_{cat}/K_m) of Com1-IFRS and Com2-IFRS for tRNA was improved by 1.8-fold and 8.8-fold, respectively, compared to IFRS (Supplementary Table 9).

CTD of PylRS containing the catalytic sites has evolved to accept various ncAAs. The mutations Com1 and Com2 were then combined with different catalytic domain (CD) mutations to enhance the incorporation efficiency of the corresponding ncAAs. To test the universality of Com1 and Com2 in enhancing the incorporation efficiency of various ncAAs, we selected CD mutations with large sequence diversity that could accept ncAAs with significantly different side chains, including Phe derivatives, Tyr derivatives, Trp derivatives, Cys derivatives, His derivatives, and Lys derivatives (Supplementary Table 10). CD mutant NACA was combined with Com1 and Com2 to test the incorporation of three Phe derivatives, including 3-bromo-L-phenylalanine (3BrF),2-chloro-L-phenylalanine (2ClF), and 3-L-phenyllactic acid (PLA). The two variants significantly improved efficiency of NACA to incorporate these three ncAAs, and the Com2 led to improvement of 38.9-fold, 29.2-fold and 7.7-fold, for 3BrF, 2ClF, and PLA respectively (Fig. 3b). When the fluorescence intensity was normalized to cAA incorporation, the fold change was 61.1-fold, 68.1-fold and 9.9-fold, for 3BrF, 2ClF, and PLA, respectively (Supplementary Fig. 14b). IFRS could also accept 3BrF and 2ClF, and Com1, Com2 mutations significantly improved the misincorporation of cAAs in presence of 3BrF and 2ClF (Fig. 3a). Interestingly, when combined with NACA, the two variants did not improve the incorporation efficiency of cAAs in presence of 3BrF and 2ClF, compared to IFRS (Fig. 3b). This could be due to that NACA exhibited lower activity against cAAs than IFRS, and Com1, Com2 did not influence substrate specificity and generally enhanced the activity of CD mutations against all substrates.

This was further confirmed by the performance of Com1 and Com2 introduced in wild-type *Mm*PylRS. *Mm*PylRS showed limited misincorporation of cAAs, and the two variants did not increase the misincorporation of cAAs either, compared to the wild type. On the other hand, the two variants significantly increased the incorporation efficiency of N6-(tert-butoxycarbonyl)-L-lysine (BocK) and N6-((allyloxy)carbonyl)-L-lysine (AlocK) and the best Com2 resulted in 40.2-fold and 32.7-fold improvement for BocK and AlocK, respectively (Fig. 3b, Supplementary Fig. 14b). This was further confirmed by the huge increase in expression level of sfGFP incorporated with BocK in presence of Com2-WT compared to WT (Supplementary Fig. 15). Mass-spectra analysis also confirmed the correct incorporation of BocK and no misincorporation of cAAs was observed (Supplementary Fig. 16 and Table 11). Interestingly, the incorporation efficiency of BocK directed by Com2-WT was 2.7-fold higher than chPylRS-IPYE obtained previously[7]. Additionally, the SCS efficiency of Com2-WT against BocK was nearly 17.9-fold higher than *Ma*PylRS, the PylRS enzyme lacking an NTD, although a previous study showed that *Ma*PylRS exhibited higher activity than wild-type *Mm*PylRS[10] (Supplementary Fig. 17a). Biochemical characterization of WT and Com2-WT using BocK confirmed that the k_{cat} of Com2-WT improved 2.6-fold, the K_m for BocK weakly increased, such that the catalytic efficiency ($k_{cat}/K_m{}^{BocK}$) of the evolved variant was enhanced by 2.2-fold compared to wild-type *Mm*PylRS (Supplementary Fig. 18). The kinetic parameters were also measured for tRNA, and the catalytic efficiency ($k_{cat}/K_m{}^{tRNA}$) of Com2-WT was improved by 1.8-fold compared to wild type (Supplementary Table 9). Interestingly, K_m for tRNA was increased by 1.4-fold for Com2-WT than wild type. Similarly, the binding affinity of Com2-WT with tRNA was around 2.1-fold lower than that of WT, suggesting that the mutations on the TBD improved the enzyme activity by modifying the tRNA binding conformation instead of enhancing the binding affinity (Supplementary Fig. 19).

Com1 and Com2 also enhanced the activities of other CD mutations toward their corresponding ncAAs. The addition of Com2 increased the incorporation efficiency of N-epsilon-Acetyl-L-lysine (AcK) by 13.3-fold compared to the original CD mutations MLAF, and the fold change was 32.2-fold when the incorporation efficiency was normalized to cAA incorporation (Fig. 3b, Supplementary Fig. 14b). The extent of improvement was significantly higher than the IPYE variant tested previously[7]. The SDS-PAGE also revealed that the amount of sfGFP expressed in presence of Com2-MLAF was significantly higher than that in presence of MLAF (Supplementary Fig. 15). Mass-spectra analysis confirmed the correct incorporation of AcK in sfGFP, but a misincorporation of lysine was also observed (Supplementary Fig. 16). Additionally, addition of Com2 improved incorporation efficiency of GML mutant by 5.4-fold, 5.8-fold and 3.9-fold against 3BrY, 3ClY and 3IY, respectively, while the fold change reached 117.3-fold, 587.6-fold and 6.5-fold when the SCS efficiency was normalized to cAA incorporation (Fig. 3b, Supplementary Fig. 14b). We checked expression of sfGFP with 3IY incorporated, and found that Com2-GML indeed significantly increased amount of expressed

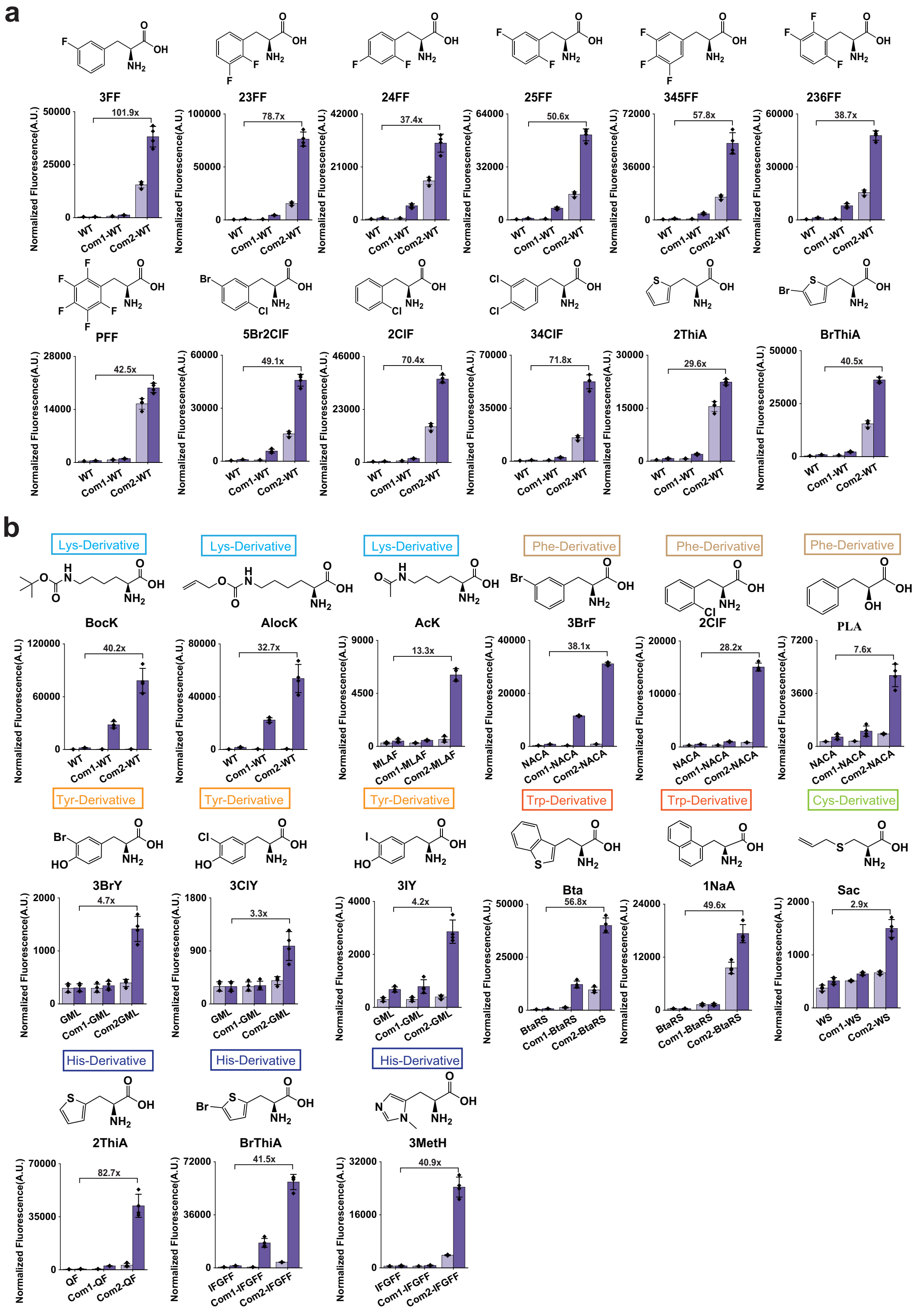

Fig. 3 | tRNA-binding domain mutations generally improved the incorporation efficiency of various ncAAs. a The SCS activity of IFRS, Com1-IFRS, and Com2-IFRS toward various substrates. **b** The SCS activity of combinatorial variants against various ncAAs. Light purple, absence of ncAAs in growth medium; Dark purple, presence of ncAAs in growth medium. Error bars represent ±standard deviation of the mean over 4 independent replicates. NcAAs include 3-fluoro-L-phenylalanine (3FF), 2,3-difluoro-L-phenylalanine (23FF), 2,4-difluoro-L-phenylalanine (24FF), 2,5-difluoro-L-phenylalanine (25FF), 3,4,5-trifluoro-L-phenylalanine (345FF), 2,3,6-trifluoro-L-phenylalanine (236FF), 2,3,4,5,6-pentafluoro-L-phenylalanine (PFF), 5-bromo-2-chloro-L-phenylalanine (5Br2ClF), 2-chloro-L-phenylalanine (2ClF), 3,4-dichloro-L-phenylalanine (34ClF), 3-(2-thienyl)-L-alanine (2ThiA), 2-(5-bromothienyl)-L-alanine (BrThiA), N6-(tert-butoxycarbonyl)-L-lysine (BocK), N6-((allyloxy)carbonyl)-L-lysine (AlocK), N-epsilon-Acetyl-L-lysine (AcK), 3-L-phenyllactic acid (PLA), 3-bromo-L-tyrosine (3BrY), 3-chloro-L-tyrosine (3ClY), 3-iodo-L-tyrosine (3IY), 3-benzothienyl-L-alanine (Bta), 3-(1-naphthyl)-L-alanine (1NaA), S-allyl-L-cysteine (Sac), 3-methyl-L-histidine (3MeH). Source data are provided as a Source Data file.

protein compared to GML. Mass-spectra analysis also confirmed the correct incorporation of 3IY (Supplementary Fig. 16).

Com1 and Com2 mutations were also constructed in BtaRS to explore their effect on the incorporation of the Trp derivatives[28]. The addition of Com2 increased the incorporation efficiency of 3-benzothienyl-L-alanine (Bta) and 3-(1-naphthyl)-L-alanine (1NaA) by 56.8-fold and 63.3-fold, respectively (Fig. 3b). SDS-PAGE revealed the significantly improved amount of sfGFP with Bta incorporated enabled by Com2-BtaRS compared to BtaRS. Kinetic parameters measurement using Bta revealed that the catalytic efficiency (k_{cat}/K_m) of Com2-BtaRS for tRNA was 4-fold higher than that of BtaRS (Supplementary Table 9). Mass-spectra analysis confirmed the correct incorporation of Bta and no misincorporation of cAAs was observed (Supplementary Table 11). However, misincorporation of cAAs was indeed observed when 1Na was incorporated by Com2-BtaRS (Fig. 3b). To explore the effect of Com1 and Com2 on the incorporation of Cys derivatives, Com1 and Com2 were combined with CD mutation WS. Com1 did not exhibit improved effect on fluorescence of sfGFP, while the addition of Com2 improved fluorescence intensity of sfGFP with Sac incorporated by 2.8-fold compared to WS (Fig. 3b), and the enhanced amount of sfGFP expressed was confirmed on SDS-PAGE (Supplementary Fig. 15). WS also showed a certain degree of misincorporation of cAAs, which was also observed for Com2-WS (Fig. 3b).

Com1 and Com2 were combined with two CD mutations, including QF and IFGFF, to explore their effect on incorporating His derivatives, including 3-(2-thienyl)-L-alanine (2ThiA), 2-(5-bromothienyl)-L-alanine (BrThiA), and 3-methyl-L-histidine (3MeH). In the results, the addition of Com2 increased the incorporation efficiency of 2ThiA, BrThiA, and 3MetH by 93.7-fold, 41.5-fold, and 40.9-fold, respectively, compared to their CD variants. The fold change reached to 223.1-fold, 61.4-fold and 201.5-fold, respectively, when the amber codon suppression activity was normalized to cAA incorporation (Fig. 3b, Supplementary Fig. 14b). SDS-PAGE revealed that Com2-QF and Com2-IFGFF indeed dramatically enhanced the amount of purified sfGFP incorporated with 2ThiA and 3MetH, respectively, compared to CD mutations alone (Supplementary Fig. 15). Kinetic parameters measurement using 3MetH confirmed that the catalytic efficiency of Com2-IFGFF for tRNA was improved by 4.3-fold compared to IFGFF (Supplementary Table 9). Moreover, according to mass-spectra analysis, no misincorporation of cAAs was observed for GFP-2MetH and GFP-2ThiA (Supplementary Fig. 16).

To explore if the variants obtained were useful in improving the expression of other proteins with ncAAs incorporated, we tested the expression of myoglobin containing 3MetH (Fig. 4a). 3MetH has been used as a heme ligand to enhance the activity of myoglobin or as a catalytic residue of an artificial esterase possessing a non-canonical organocatalytic mechanism[29]. Here, we used *Mm*PylRS variant Com2-IFGFF to incorporate 3MetH into the position His93 of myoglobin as a ligand of heme. The protein expression was explored by carrying out protein purification from the same amount of cells in the presence of Com2-IFGFF and IFGFF. It was found that the concentration of 3MetH-containing myoglobin was 28.3 mg/L for Com2-IFGFF, 6.3-fold higher than 4.5 mg/L of IFGFF (Fig. 4b). We then measured the activities of Mb-3MetH against guaiacol using the purified protein without dilution. The myoglobin catalyzes the oxidation of guaiacol by hydrogen peroxide to generate a stable tetrameric product whose formation can be readily monitored by absorbance at 470 nm. The yield of product was significantly higher for Com2-IFGFF compared to IFGFF (Supplementary Fig. 20). The ΔOD_{470} reached 0.063 in the reaction system containing Com2-IFGFF after a 40-min reaction, 7.9-fold higher than that using IFGFF (Fig. 4b, Supplementary Fig. 20), confirming the higher expression of target protein aided by the Com2-IFGFF variant.

Suppression of multiple amber codons of PylRS variants

We then characterized the ability of Com1-IFRS and Com2-IFRS to suppress multiple amber codons in sfGFP, which is important for incorporating multiple unnatural amino acids into proteins. One to five consecutive amber codons were inserted second position of sfGFP (Fig. 4c). Both Com1-IFRS and Com2-IFRS exhibited higher fluorescence intensity than IFRS in all situations, while Com2-IFRS showed higher suppression ability than the Com1-IFRS, with 122.4-fold, 99.4-fold, 91.2-fold and 53.3-fold improvement compared to IFRS, for S2TAG × 2, S2TAG × 3, S2TAG × 4 and S2TAG × 5, respectively (Fig. 4d, Supplementary Fig. 21a). It was also found that Com1-IFRS and Com2-IFRS improved the incorporation efficiency against native amino acids compared to IFRS. We also tested the incorporation efficiency of multiple unnatural amino acids at different positions of sfGFP (Fig. 4e). When 3BrF was incorporated at the position of D36 of sfGFP, the fluorescence intensity was different from that of sfGFP with 3BrF at the second position, for both the wild-type IFRS and mutant Com1-IFRS and Com2-IFRS (Fig. 4f, Supplementary Fig. 21b). Similar site-dependent incorporation efficiency has previously been observed for other ncAAs[7]. Despite this, Com2-IFRS still showed higher amber codon suppression efficiency than Com1-IFRS and IFRS, with 3.8-fold, 7.9-fold, 27.3-fold, 4.7-fold and 5.2-fold improvement compared to IFRS, for 1TAG, 2TAG, 3TAG, 4TAG and 5TAG, respectively.

MD simulations to explore the molecular change of PylRS variants

The whole 3D structure of *Mm*PylRS was not yet available as the full-length protein is insoluble. Hence, AlphaFold3 was used to predict the structures of *Mm*PylRS (WT), Com1-WT, and Com2-WT in complex with $tRNA^{Pyl}$ (Supplementary Fig. 22). The predicted *Mm*PylRS structure aligned well with the separate NTD structure and CTD structure determined previously (Fig. 5a). 50-ns MD simulations were then conducted for these structures in complex with Pyl-AMP to understand how the mutations influenced the binding of tRNA and the enzyme activity (Supplementary Figs. 23–25). The root-mean square deviation (RMSD) values revealed that the trajectories were well equilibrated at the last 10 ns, which were used for further analysis (Supplementary Fig. 25a). The reaction distance between the 3′-OH of tRNA A76 and the carboxyl carbon atom of amino acid Pyl was first analyzed. It was generally shorter for Com2-WT than wild type and Com1-WT (Fig. 5b, Supplementary Fig. 26). A total of 1000 snapshots were analyzed, and the number of snapshots with a distance shorter than 4 Å was 462, which is 21-fold and 10-fold higher than that of wild type and Com1-WT, respectively (Fig. 5c). These indicated that Com2 mutations mediated the binding of $tRNA^{Pyl}$ to make the aminoacylation reaction happen more easily. Interestingly, in the reaction conformations, we observed new hydrogen bonds formed between Pyl-AMP and tRNA in the two variants compared to WT. Com1-WT showed a new hydrogen bond formed between the main chain -NH_2 of Pyl and 2′-OH of tRNA A76, while Com2-WT exhibited a hydrogen bond formed between the main chain -NH_2 of Pyl and the 3′-OH of tRNA A76. No such hydrogen bonds were found in wild type (Supplementary Fig. 27). These new hydrogen bonds will contribute to the interaction between Pyl-AMP and tRNA, and hence accelerate the reaction.

The hydrogen bonds formed between tRNA and the protein were also analyzed. In last 10-ns MD simulations, the number of hydrogen bonds in PylRS TBD with occupancy over 60% was 22, 19, and 22 for WT, Com1-WT, and Com2-WT, respectively (Supplementary Fig. 28). Specifically, both Com1-WT and Com2-WT formed new hydrogen bonds including LYS3-A58, ARG19-A46, ARG52-G52, ARG193-A5, ARG193-C13 and ARG193-U12, while Com2 formed several extra hydrogen bonds such as ARG55-C45, ARG55-A46, Arg58-A20 (Fig. 5d). Additionally, several hydrogen bonds were disrupted in the variants, such as ASN49-G47, ARG55-A46, ARG55-G21, ARG58-A58, R66-G21 and so forth. Specifically, in the conserved Motif 2 loop that is responsible for tRNA recognition, two hydrogen bonds Lys336-C71 and Lys336-C72 were disrupted, and a new hydrogen bond Asp334-C71 was formed in both the Com1 and Com2 variants (Supplementary Fig. 29).

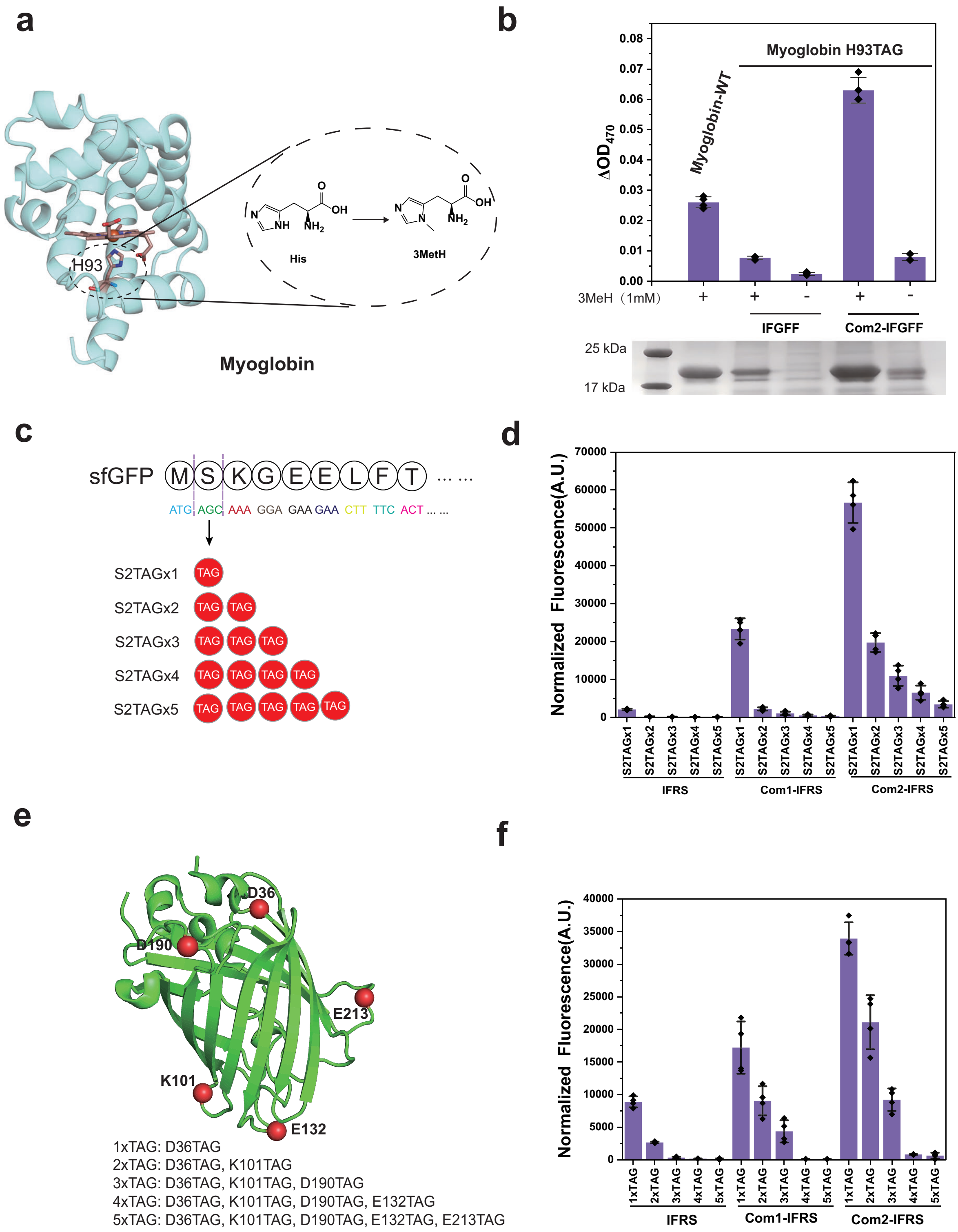

Fig. 4 | Further characterization of PylRS variants. **a** Introduction of 3MetH at His93 position of myoglobin as a ligand of heme. **b** Product yield of reaction catalyzed by myoglobin-3MetH after 40-min reaction, and expression of myoglobin-3MetH detected on SDS-PAGE. The incorporation of 3MetH was enabled by IFGFF and Com2-IFGFF. +, with 3MetH added; −, no 3MetH added. Myoglobin-WT indicates no ncAA incorporation in myoglobin. **c** Suppression of multiple amber codons by PylRS variants, with multiple consecutive amber codons inserted at the second position of sfGFP. **d** Fluorescence intensity of sfGFP with multiple 3BrF inserted at the second position, enabled by IFRS, Com1-IFRS, and Com2-IFRS. **e** The positions where multiple 3BrF are inserted in sfGFP. **f** Fluorescence intensity of sfGFP with multiple 3BrF inserted at different positions, enabled by IFRS, Com1-IFRS, and Com2-IFRS. Error bars represent ±standard deviation of the mean over four independent replicates. Source data are provided as a Source Data file.

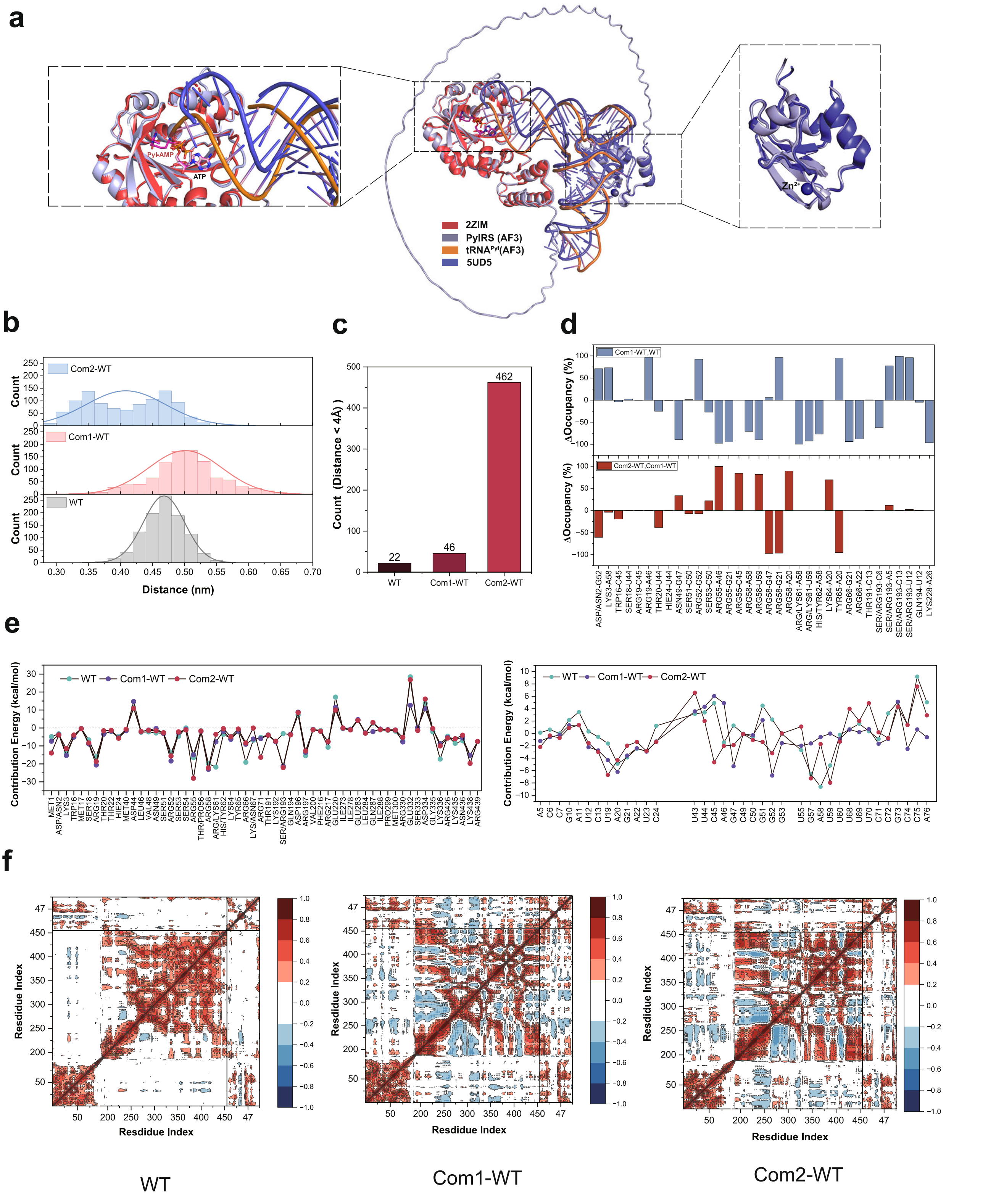

Fig. 5 | MD analysis for WT, Com1-WT, and Com2-WT. a The *Mm*PylRS structure predicted by Alphafold 3. **b** The reaction distance between the 3′-OH of tRNA A76 and the C_a atom of amino acid Pyl, calculated through analysis of 10 ns equilibrated trajectories. **c** The number of snapshots with a distance shorter than 4 Å. **d** Hydrogen bonds occupancy in WT, Com1-WT, Com2-WT, calculated through analysis of 10 ns equilibrated trajectories. **e** The binding free energy of amino acids within 4 Å of the tRNA as ligand and the tRNA bases. **f** Dynamics cross-correlation map for the C_α atom and tRNA P atom pairs within *Mm*PylRS and variants calculated with the last 150 ns MD trajectory. Protein contains 454 amino acids, and the tRNA contains 72 bases (Supplementary Fig. 34). The correlation coefficient (C_{ij}) was shown in different colors. C_{ij} with values from 0 to 1 represents positive correlations, whereas C_{ij} with values from −1 to 0 represents negative correlations.

Additionally, an extra H-bond GLU332-C74 was formed in Com1 variant but not in WT and Com2. This indicated that mutations reshaped the interactions between tRNA and protein, and the Com1 and Com2 improved the tRNA binding in a different way. As a result, the interaction energy between tRNA and protein was different for the wild type and variants.

The binding free energy between tRNA and different domains of the protein was analyzed. Com1-WT and Com2-WT exhibited lower binding free energy compared to the wild type, which was mainly attributed to the decreased binding free energy of the tRNA binding domain (Supplementary Fig. 30). Interestingly, the binding free energy of full-length Com1-WT was lower than Com2-WT, while that of the tRNA binding domain of Com1-WT was higher than Com2-WT. The binding free energy determined in MD simulations seemed to contradict with binding affinity and K_m values measured for WT and Com2-WT. This could be attributed to the different ncAA substrate used. In the MD simulations, the substrate was native substrate pyrrolysine, while the substrate used in biochemical characterization was BocK. Kinetic parameters measurement using different ncAA did reveal that Com2 influenced the K_m for tRNA in a different way in the presence of different ncAA (Supplementary Table 9).

Residue-level binding energy contribution analysis for both protein and tRNA was carried out. Several amino acids and tRNA bases were indeed found to impact the binding free energy. For example, GLU332 in Com1-WT exhibited reduced binding energy, while no significant changes were observed in WT or Com2-WT, which might be attributed to the newly formed H-bond Glu332-C74 in Com1-WT. S193R mutation formed new salt bridges with tRNA, including Arg193-A5, Arg193-C13, Arg193-U12, which led to a significant decrease in the binding free energy (Fig. 5e). Also, Arg55 of Com2-WT showed a significantly low binding energy due to the T56P mutation, although it is not the case for Com1-WT. As for the analysis of tRNA binding energy, it was found that the mutations significantly increased the binding free energy for the last three bases C74, C75, and A76, which might facilitate the aminoacylation reaction (Fig. 5e).

Dynamics cross-correlation matrices (DCCMs) were also computed for the WT and two variants to understand how the mutations impact protein dynamics. Since the MD simulation systems are large and complex, a robust coupling analysis of dynamic cross-coupling correlation might need simulations of a longer timescale than 50 ns. We hence carried out 200-ns simulations for WT, Com1-WT, and Com2-WT (Supplementary Figs. 31–33, Supplementary Note 1), and the last 150 ns trajectories were used to compute DCCM of the protein C_α atom pairs and tRNA P atom pairs. Generally, the variants showed more dynamics cross-correlations between residue pairs compared to the wild type, while Com1-WT and Com2-WT exhibited similar C_{ij} values in most of the regions (Fig. 5f). Specifically, it was found that the dynamics correlation between NTD and CTD was more significant in the Com1-WT and Com2-WT, compared to WT. Although NTD and CTD are distant from each other, they are connected together with the tRNA. The mutations on tRNA binding domain hence impact the dynamics of CTD through modification of interactions with tRNA. The increased correlated dynamical network would help to maintain more reaction conformations, thereby enhancing the enzyme activity of Com1-WT and Com2-WT.

Discussion

This study utilized machine learning to explore the combinatorial mutation space of *Mm*PylRS TBD and identified great variants including Com1-IFRS and Com2-IFRS (Supplementary Table 12), which modified the tRNA binding, enhanced the aminoacylation rate by up to 5.6-fold against 3BrF, and subsequently significantly improved SCS efficiency by up to 101.9-fold improvement for 3FF. PylRS has been engineered for improved activity through N-terminal mutations. However, all previous studies applied directed evolution strategies for the PylRS N-terminal engineering, and the methods used included error-prone PCR to construct a library coupled with screening based on GFP fluorescence or white/blue colony[11], and a PACE[7]. Although directed evolution is a powerful strategy for enzyme engineering, its success relies on iterative cycles of library construction and screening, and it can be trapped in local fitness optima due to taking one mutation step at a time. In our study, we applied deep learning models capable of zero-shot prediction of high-fitness variants to explore mutation target sites in the whole TBD, and a supervised model, FFT-PLSR, to explore the sequence space once the mutation target sites were identified. Thanks to the ML models we used, the variants obtained were significantly more active than the variants obtained previously.

Due to epistatic interactions between mutations, combining multiple mutations does not always result in positive effects. Investigating the combinatorial effects of mutants involves two main tasks: (1) pairwise combinations of specified mutations, and (2) exhaustive combinations of 20 amino acids at specified mutation sites. Using the FFT-PLSR model, we demonstrated that, by employing a training dataset composed of 38 single, double, and triple mutants, it was possible to identify multi-point combinatorial mutants with significantly improved activity within the sequence space of 12 single-point mutations and their pairwise combinations, totaling 4,096 mutants, thereby providing a solution to Task 1. We also attempted to apply the ML model to tackle task 2 by restricting the mutant combinatorial space and focusing solely on combinations of single-point improved mutations. We found that using multiple AA indexes to encode protein sequences yielded better results than a single-index model, with the model achieving a fit of 0.729 for the test data set. The advantage of the FFT-PLSR model lies in its effective utilization of experimental mutant activity data, guiding the construction of the next set of combinatorial mutants. With the Com1-IFRS as the parent sequence, only 20 variants predicted by the ML model were tested. Among them, the variant Com2-IFRS showed a 2.8-fold increase in activity. In the future, the activity data of these variants could be fed back and used to update the ML model, which would then be employed to predict the next round of variants. Additionally, the construction and characterization of PylRS variants could be carried out by an automatic biofoundry, ensuring high reproducibility and efficient data collection. This would accelerate the protein engineering efficiency of PylRS, as successfully demonstrated with the *Methanocaldococcus jannaschii* tyrosyl-tRNA synthetase[30].

Zero-shot ML models learn general patterns in proteins and can predict high-fitness protein variants without requiring any prior knowledge other than protein sequence or structure. These models can help to design an initial variant library or explore additional potential mutations when directed evolution reaches a local minimum, opening additional evolutionary pathways. In our study, sequence-based zero-shot model ESM-1V did not perform well, with mutations predicted primarily located in the linker region 90–185 aa, and only a positive mutant I176S was obtained, showing a minor improvement in activity. The ESM model also suggested the low-activity C42N mutation, while C42, C69, C72, and H24 are critical Zn-binding residues, highlighting the model's limitation in considering structural constraints (Supplementary Fig. 35). Previously, ESM models were successfully applied to guide the affinity maturation of antibodies, and optimization of an uracil-N-glycosylase variant activity that enables programmable T-to-G and T-to-C base editing[31]. Several ESM models are currently available, with each trained on different protein sequence datasets with varying numbers of parameters[21,32,33]. We also tested the effect of ESM2_t33_650M_UR50D and ESM2_t36_3B_UR50D on the prediction of high-fitness variants of IFRS, which predicted seven single variants, and six of them have been predicted by ESM-1v, and none of the six variants tested showed improved activity (Supplementary Table 13). The poor performance of ESM models on engineering IFRS might indicate that although the protein language models, such as ESM trained on vast datasets of natural proteins, allowed zero-shot optimization of specific proteins, it remains a challenge to generally select promising variants for various enzymes possessing remarkable diversity in terms of their classes and catalytic mechanisms. Instead of being used for zero-shot prediction of high-fitness variants, the protein large language models (PLMs) such as ESM could be used to encode protein sequences for building ML models. EVOLVEpro, a few-shot active learning framework that combines PLMs and regression models, has been developed to rapidly improve protein activity[34]. The single-point variant data of IFRS predicted by the zero-shot ML models could serve as the input of EVOLVEpro, enabling further prediction of high-fitness single-point variants. Through multiple rounds of this iterative process, the model can generate variants with significantly improved activity.

Structure-based model MutCompute predicted mutations to optimize the protein structure, and six improved variants were obtained using this method. Interestingly, MutCompute made predictions based on the crystal structure of wild-type PylRS, and mutations beneficial to the wild-type PylRS do not necessarily work when transferred to PylRS mutants. However, experimental results showed that six mutations predicted by MutCompute still enhanced the activity of the Com1-IFRS. Actually, MutCompute has been used to improve activity and stability of several enzymes, including PET-degrading enzymes[35], DNA polymerase[36], and haloalkane dehalogenase[37]. ProRefiner, developed for inverse protein folding, also predicted two improved Com1-IFRS variants. A similar model, ProteinMPNN, has recently been used to generate myoglobin and tobacco etch virus protease designs with improved expression, elevated melting temperatures, and enhanced function[38]. However, these designs contain multiple mutations and only have 41% to 85% sequence identity to the parent sequences. In contrast, the ProRefiner has been validated for designing single-point mutants with improved activity for transposon-associated transposase, which aligns closely with our research task[23].

The evolved PylRS TBD mutants have demonstrated exceptional effectiveness in increasing the yield of ncAA-containing proteins. The resulting TBD combinatorial mutants improved IFRS activity across its substrate spectrum, and when combined with diverse CD mutants, they enhanced the incorporation efficiency of six types of ncAAs.

*Ma*PylRS, the enzyme lacking an NTD, has been shown to be slightly more active than the wild-type *Mm*PylRS[10]. However, the incorporation efficiency of BocK directed by the Com2 variant of *Mm*PylRS was 17.9-fold higher than that of *Ma*PylRS, indicating the important role of the NTD of *Mm*PylRS in catalysis. Since *Ma*PylRS lacks the NTD, the strategy of engineering the NTD to enhance enzyme activity might not be applicable to *Ma*PylRS. Previously, Lin et al. achieved efficient SCS by fusing the *Mm*PylRS TBD region with the CD of various cAA RS, including histidine, phenylalanine, and alanine[39]. We transplanted Com2 mutations to chHisRS, which contained an NTD from MbPylRS-IPYE, MmPylRS, and a CTD from HisRS. The newly formed variant Com2-HisRS exhibited a 6.8-fold improvement in SCS efficiency characterized by fluorescence intensity of sfGFP2TAG, compared to the chHisRS (Supplementary Fig. 17b). The TBD mutations obtained in this study are hence promising to increase the activities of diverse chimeric aaRS to enhance the incorporation efficiency of more types of ncAAs into proteins. Moreover, these aaRS obtained are expected to have a marked impact on the suppression of multiple TAG stop codons, and hence make the genetic code expansion technology more useful.

We carried out the aminoacylation reaction in vitro to characterize the activity and catalytic efficiency of PylRS variants. However, the extent of improvement in catalytic efficiency (k_{cat}/K_m) of the Com2 variant relative to the wild type was not as high as the ncAA incorporation efficiency determined by sfGFP2TAG fluorescence intensity. This might be because the mutations improved the PylRS expression, which hence enhanced the suppression of the amber codon. There is a mutation D2N in Com1-IFRS and Com2-IFRS, which showed a 3.6-fold improvement in sfGFP2TAG expression yield compared to IFRS. Based on the N-terminal rule that governs the rate of protein degradation, the N-terminal amino acid of a protein determines its half-life in vivo[40]. The first methionine of *Mm*PylRS variant could be removed after translation, and the Asn2 became the N-terminal residue, which might enhance the half-life of PylRS, and hence affect the protein expression of sfGFP.

Additionally, the activity of PylRS was quantified based on the production of AMP. However, we observed a strong AMP production background during the measurement of kinetic parameters for tRNA (Supplementary Fig. 36a). We tested the stability of ATP in the reaction buffer and found that ATP was stable in the reaction buffer before adding the enzyme, and hydrolysis only happen in presence of PylRS (Supplementary Fig. 36b). Further characterization of WT and Com2-WT revealed that the ATP hydrolysis led to the production of AMP, and this could occur even in absence of tRNA and ncAA (Supplementary Fig. 36c). And, interestingly, Com2-WT exhibited improved ATP hydrolysis capability compared to WT. This might cause some error when using the production AMP to quantify the enzyme activity, as part of the AMP determined might be from ATP hydrolysis instead of the aminoacylation reaction.

The wild-type PylRS exhibits poor solubility due to its hydrophobic NTD, which hinders crystallization of the full-length enzyme. The latest protein-tRNA structure prediction tool, AlphaFold3, was used to model the structure, allowing us to illustrate the mechanism of improved mutations. To better represent the binding conformation, Zn^{2+} and ATP were included in the system. Analysis of the MD simulation results suggested that the enhanced activity of Com2-IFRS likely stems from the mutations reshaping the binding interactions between protein and tRNA, which reduced the reaction distance between tRNA and Pyl-AMP, thereby enhancing aminoacylation efficiency. Additionally, through DCCM analysis, we found that Com2-WT exhibited a significantly enhanced overall dynamics network, implying that tRNA binding strengthens both intra- and inter-domain couplings. To our knowledge, this is the first study working on MD simulations of the predicted *Mm*PylRS structure containing both N-terminal and C-terminal domains, paving the way for computational design of more effective PylRS variants.

Methods

Reagents

Primers and genes were synthesized by Tsingke Biotechnology Co., Ltd. (Beijing, China). The high-fidelity DNA polymerase was from Vazyme Biotech Co., Ltd (Nanjing, China), while the *Dpn* I enzyme was obtained from Takara (Kusatsu, Japan). The T7 RNA polymerase was from ABclonal Technology (Wuhan, China). The Easy RNA Cleanup Kit was from Zhejiang Easy-Do Biotechnology Co., Ltd (Hangzhou, China). The AMP-Glo^{TM} assay was from Promega (Beijing, China). The Ni NTA Beads 6FF were from LABLEAD (Beijing, China). The pBK-PylRS plasmid and the pMyo4TAG-PylT plasmid are a kind gift from Dr. Jason Chin, Medical Research Council Laboratory of Molecular Biology. All ncAAs used in these studies were purchased from Aladdin (Shanghai, China), Bidepharm (Shanghai, China), and Macklin (Shanghai, China). The concentrations of tRNA, DNA, and protein were measured using NanoDrop One (Thermo Fisher Scientific Inc., USA). Disposable fiber-optic streptavidin-coated tips were purchased from ForteBio Inc. (USA).

PCR-based methods for the construction of PylRS mutants

The gene sequences encoding each of the *Mm*PylRS mutants were codon-optimized for expression in *E. coli* and were cloned between the *Nde*I site and *Pst*I sites of the pBK-PylS plasmid[41]. The pBK-*Mm*PylRS plasmid was chosen as a template for mutant construction with the plasmid PCR approach. The primer sequences used for PylRS variant construction are supplied in the Supplementary Data. The polymerase chain reaction (PCR) mix used was a high-fidelity DNA polymerase Mix (P525, Vazyme). The PCR conditions are 95 °C 3 min, (95 °C 30 s, 60 °C 30 s, 72 °C 90 s) × 30 cycles, 72 °C 10 min. After PCR, 2 μL *Dpn* I was added to a 50 μL PCR reaction mixture, and the digestion was carried out at 37 °C for 2 h and 75 °C for 15 min. After *Dpn* I digestion, 10 μL of PCR products were directly transformed into chemically competent *E. coli* BL21(DE3) for the following experiments.

Superfolder GFP (sfGFP) reporter assay

The sfGFP expression plasmid, named pGFP-PylT, was constructed by replacing the myoglobin gene in pMyo4TAG-PylT (a kind gift from Dr. Jason Chin, Medical Research Council Laboratory of Molecular Biology) with the sfGFP gene. The plasmids pGFP2TAG-PylT, pGFPS2XTAG-PylT, and pGFPNTAG-PylT, containing multiple amber codons, were constructed by multiple rounds of PCR using the pGFP-PylT plasmid as a template. The plasmids encoding *Mm*PylRS mutants were co-transformed with pGFP2TAG-PylT, pGFPS2XTAG-PylT, or pGFPNTAG-PylT into *E. coli* DH10B cells for assessment of PylRS activity by the expression of sfGFP. After incubation in lysogeny broth (LB) medium at 37 °C for 60 min, the co-transformed cells were spread onto LB agar containing 50 μg/mL kanamycin and 12.5 μg/mL tetracycline. After cultivation for 12 h at 37 °C, individual colonies were inoculated into 5 mL LB with required antibiotics at 37 °C and grown to an OD_{600} of 0.4–0.6. The cells were harvested by centrifugation (7104 × *g*, 10 min), washed, and resuspended in antibiotics-supplemented minimal medium. The sfGFP expression was induced by arabinose with a final concentration of 0.2%. 200 μL aliquots of induced cells were transferred into the 96-well plates in the presence or absence of the corresponding ncAAs in each well. After 12 h induction at 37 °C, the fluorescence was quantified using a BioTek Synergy H1 microplate reader (excitation/emission: 485/515 nm). Signals were background-corrected and normalized to cell density (OD_{600}) measured on the same instrument.

Protein expression and purification for sfGFP

For sfGFP expression and purification, the DH10B cells containing pBK-PylRS and pGFP2TAG-PylT were grown overnight in 5 mL LB with required antibiotics at 37 °C. 2 mL of cells were then inoculated into 200 mL of fresh LB medium supplemented with the same antibiotics, and grown to an OD_{600} of 0.3. sfGFP expression was induced by adding L-arabinose to a final concentration of 0.2%, followed by incubation at 30 °C and 220 rpm for 20 h, either with or without the corresponding ncAAs. Cells were harvested by centrifugation at 4000 × *g*, 10 min, 10 °C. The pellets were resuspended in ice-cold phosphate-buffered saline (PBS) buffer containing 20 mM imidazole (pH 8.0) and then sonicated. After centrifugation at 12,000 × *g*, 4 °C for 20 min, the supernatant was loaded through Ni-NTA beads equilibrated with PBS buffer containing 20 mM imidazole, pH 8.0, and then washed with PBS buffer containing 50 mM imidazole, pH 8.0. The proteins were eluted with PBS buffer containing 500 mM imidazole, pH 8.0. The purified proteins were concentrated using an Amicon Ultra 10,000 MWCO (Millipore) with PBS buffer pH 8.0, and stored at −80 °C.

Protein expression and purification for myoglobin

For myoglobin expression and purification, the co-transformed DH10B cells containing pBK-PylRS and pMyoglobin-93TAG were cultured overnight and diluted at a ratio of 1:100 into 200 mL of fresh LB medium supplemented with the required antibiotics and 1 mM 3-methyl-L-histidine. Cells were grown until OD_{600} reached 0.3. L-arabinose was added with a final concentration of 0.2% to induce myoglobin expression (37 °C, 220 rpm, 12 h) with or without the addition of 3MeH. Cells were harvested by centrifugation at 4000 × *g* for 10 min at 4 °C. The resulting cell pellets were suspended in ice-cold Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 20 mM imidazole, pH 7.6) and then sonicated. The suspension was centrifuged at 12,000 × *g* for 20 min at 4 °C. For 6xHis tag fusion proteins, the resulting supernatant was purified via Ni^{2+}-affinity chromatography on chelating Sepharose equilibrated with Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 20 mM imidazole, pH 7.6), and washed with 10 volumes of Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 50 mM imidazole, pH 7.6). The proteins were eluted with Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 250 mM imidazole, pH 8.0). The protein was concentrated using an Amicon Ultra 10,000 MWCO (Millipore) with Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, pH 7.6), and stored at −80 °C.

Myoglobin activity assay

The purified myoglobin was standardized to the same volume to ensure that the activity differences were due to protein concentration. Guaiacol was used as the reaction substrate at a final concentration of 2.5 mM[42]. The reaction mixture had a total volume of 200 μL, consisting of 10 μL of purified myoglobin, 5 μL of 0.1 M guaiacol, and 185 μL of 10 mM H_2O_2-PBS solution. Product formation was monitored spectrophotometrically at 470 nm. The OD changes of the reaction mixture per minute were recorded using kinetic scanning by a BioTek Synergy H1.

AARS variant expression and purification for aminoacylation assay

The genes of PylRS and its variants were cloned into pET28a and transformed into BL21(DE3) cells for expression. Cells were grown in LB medium containing 50 μg/mL kanamycin and 50 μM $ZnCl_2$ at 37 °C until the absorbance at 600 nm (OD_{600}) reached 0.6. The protein was then induced by the addition of isopropyl-β-D-l-thiogalactopyranoside (IPTG) to a final concentration of 100 μM and shifted to 18 °C for ~18 h before harvesting. The cells were harvested and resuspended in Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 20 mM imidazole, pH 7.6). After sonication, the suspension was centrifuged at 12,000 × *g* for 20 min at 4 °C. The resulting supernatant was purified via Ni^{2+}-affinity chromatography on chelating Sepharose equilibrated with Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 20 mM imidazole, pH 7.6) and washed with 10 volumes of Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 50 mM imidazole, pH 7.6). The proteins were eluted with Tris-HCl buffer (50 mM Tris-HCl, 300 mM NaCl, 250 mM imidazole, pH 7.6). The protein was concentrated using an Amicon Ultra 10,000 MWCO (Millipore), stored at −80 °C.

Preparation of $tRNA^{Pyl}$

The *M.mazei* $tRNA^{Pyl}$ was transcribed using T7 RNA polymerase. The DNA oligonucleotides used to construct double-stranded DNA templates for $tRNA^{Pyl}$ include 5′-primer **TAATACGACTCACTATA**GGAAACCTGATCATGTAGATCGAATG, middle template **CACTATA**GGAAACCTGATCATGTAGATCGAATGGACTCTAAATCCGTTCAGCCGGG TTA and 3′-primer TGGCGGAAACCCCGGGAATCTAACCCGGCTGAACGGATTTAGAG (reverse), in which the T7 promoter sequence is shown in bold. In vitro transcription was performed for 8 h at 37 °C in a solution containing T7 RNA polymerase, NTPs (ATP, UTP, CTP, GTP), and double-stranded DNA transcription templates obtained by PCR. Transcribed tRNAs were purified with the Easy RNA Cleanup Kit according to the manufacturer's instructions. Purified tRNA was dissolved in 10 mM Tris-HCl, 10 mM $MgCl_2$ (pH 7.6) and stored at −80 °C.

Aminoacylation assay

The transcribed $tRNA^{Pyl}$ was first refolded by heating at 85 °C for 2 min, then at 37 °C for 15 min. A 10 μL reaction mixture was prepared by mixing 5 µL 2× assay buffer (50 mM Tris pH 7.6, 40 mM KCl, 4 mM DL-dithiothreitol, 20 mM $MgCl_2$, 0.2 mg/ml bovine serum albumin in RNAse-free water), 1 µL RNAse-free water, 1 µL 20 mM ncAA, 1 µL tRNA, 1 µL synthetases, and 1 µL 1 mM ATP. For WT and Com2-WT, synthetases were diluted to 100 nM in 1× assay buffer. For IFRS, Com1-IFRS, Com2-IFRS, BtaRS, Com2-BtaRS, IFGFF, and Com2-IFGFF systems, synthetases were diluted to 200 nM in 1× assay buffer. The reaction was carried out at 37 °C for 30 min. Various concentrations of tRNA (0.5-32 µM) were used to determine kinetic parameters for the corresponding synthetases. Various concentrations of BocK (0–10 mM) were also used to determine kinetic parameters for WT and Com2-WT with 5 µM tRNA in the reaction system. The AMP generated by the reaction was then measured with the AMP-Glo™ assay[43] by following the instructions from the manufacturer.

Binding affinity measurement

The $tRNA^{Pyl}$ with 5′ end RNA biotinylation was synthesized by Tsingke Biotech. Wild-type PylRS (WT) and its Com2-WT variant were expressed and purified as above. Before measurement, the $tRNA^{Pyl}$ was refolded by heating at 85 °C for 2 min, followed by incubation at 37 °C for 15 min. Octet platform (ForteBio Inc. USA) was used to monitor and quantify the binding affinity between $tRNA^{Pyl}$ and either WT or Com2-WT. The procedure consisted of five sequential steps: (1) baseline; (2) loading; (3) washing; (4) association; (5) dissociation. In a 96-well microtiter plate, the following solutions were prepared (a total of 200 µL per well) respectively: baseline solution (PBST buffer), loading solution (PBST buffer with 200 nM tRNA-Probe), washing solution (i.e., baseline solution), association solution (PBST buffer with various synthetase concentrations), and dissociation solution (i.e., baseline solution). A disposable fiber-optic streptavidin-coated sensor tip was first dipped into the baseline solution for 60 s with gentle automated shaking, then loaded in the tRNA-Probe solution for 300 s to allow coupling of the biotinylated $tRNA^{Pyl}$. The probe-saturated tip was then washed for 120 s to remove any nonspecifically adsorbed tRNA-Probe. During the association phase, the sensor tip was immersed in the solution containing target synthetase for 300 s, allowing it to bind with the immobilized $tRNA^{Pyl}$ probe in a sandwich-like format involving the tethered probe, pendant tRNA structure, and target protein. Finally, the binding tip was in the dissociation solution for 300 s. The resulting binding curves were analyzed using Octet Analysis Studio. Data from control wells (without synthetase in the association step) were subtracted as background, and KD values were determined based on the known concentrations of the synthetases.

FFT-PLSR model training

For each model, the process from training to prediction is divided into three stages: data processing, encoding index searching, and comprehensive model training.

Data processing. The amino acid sequences of mutants are transformed into 566 numerical sequences using 566 different AAindex representations[44,45]. Prior to applying FFT, mean standardization is performed, where each element in the numerical sequence is adjusted by subtracting the sequence's mean value. Subsequently, FFT is applied to each numerical sequence, as represented by Eq. 1[46]:

$$f_j = \sum_{k=0}^{N-1} x_k e^{\frac{-2i\pi}{N} j \cdot k} \tag{1}$$

where f_j is the protein output spectrum of complex numbers, j is an index of the Fourier Transform. The numerical sequence includes N value(s) denoted as x_k, with $0 \le k \le N{-}1$ and $N \ge 1$; k is the frequency in the spectrum; i defines the imaginary number such that $i^2 = -1$.

The output of the FFT is a complex sequence, with each element represented as $a + bi$, where a is the real component, b is the imaginary part, and i is the imaginary unit. From these two components, both a real and imaginary protein spectrum can be derived, along with an absolute spectrum (or power) spectrum[46]. The absolute spectrum was the spectrum of choice for the encoding strategy studied. Due to the symmetry of the transformed sequence, the absolute spectrum of the first half is applied. The absolute spectrum is represented as:

$$AS(m) = \left| f_j \right|, m = 1, 2, 3, \cdots, M \tag{2}$$

Where f_j is the output of the FFT, M is the number of protein sequences.

After FFT processing, a spectral form of the protein called the protein spectrum is generated. The amplitudes of the spectrum were then normalized to be between 0 and 1 by Min-max scaling. These normalized protein spectra, along with protein activity data, formed the ML model training set.

Encoding index searching. This stage aims to score the encoding method for each AAindex. To enhance the model's fit, a GridSearchCV approach is employed to optimize the hyperparameter n_components of PLSR, with a search range of 2 to 10. Depending on the dataset size, either LOOCV or k-fold Cross-Validation (k-fold CV) is chosen. The evaluation metrics include the cross-validated Mean Squared Error (cvMSE) and the coefficient of determination (R^2). Here, cvMSE is derived from the average of mean squared errors during cross-validation, which facilitates the construction and selection of the most robust models, while R^2 quantifies the correlation between the predicted and observed values based on the optimal hyperparameter model.

$$\text{cvMSE} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n} \tag{3}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

Where y_i is the measured activity of the ith sequence, $\hat{y}_i$ is the predicted activity of the ith sequence, $\bar{y}$ is the average of measured activities, and n is the number of sequences.

Incorporating the aforementioned process with different AAindex encoding methods allows us to retrain the model and score each of the 566 indices. We then select the highest-scoring index to be used as the final index for model construction.

Model training. Utilizing the optimal index selected in the previous step, we encode the data and train the model using the entire dataset. This trained model is subsequently employed for further exploration of the prediction space.

Expansion of the FFT-PLSR model. An Extended Sequence (Ext_SEQ) method has been proposed to enhance the FFT-PLSR model, demonstrating improved model fitting capabilities. In this study, we applied this method to models constructed using single and double mutants data from Com1. The core of this approach lies in the combination of sequences encoded by multiple indices. The best index, index1, is identified as the one that yields the lowest cross-validated Root Mean Square Error (cvMSE) and is selected as the first index for constructing the Ext_SEQ. The protein sequence is encoded using this index, generating a corresponding protein spectrum. In the second iteration, another index, index2, is chosen to construct an Ext_SEQ composed of two elemental sequences. The index2 is selected based on a second ranking that excludes the previously used index, index1. This process is repeated iteratively in each subsequent iteration to identify the optimal index for modeling and to expand the Ext_SEQ to include three or more indices. Ultimately, this concatenation process is performed based on purely statistical criteria.

Molecular dynamics simulations

The full-length structure of the PylRS dimer was predicted by AlphaFold3, including PylRSs, tRNAs, zinc ions, and ATPs. The structure of Pyl-AMP was obtained from the crystal structure of the CTD of PylRS (PDB: 2ZIM). Upon alignment, it was found that the position of AMP in Pyl-AMP closely overlaps with the position of ATP in the predicted structure. Therefore, Pyl-AMP was directly substituted for ATP. All simulations were performed using GROMACS 2022.02[47] using an Amber14sb/parmbsc1 force field[48]. The force field parameters of the substrate (Pyl-AMP) were constructed using ACPYPE[49]. Periodic boundary conditions were applied in all simulations. The initial structure was solvated in a cubic simulation box with a layer of water at least 10.0 Å from the protein surface. Sufficient counter ions (Na^+) were added to neutralize the system. The protonation states of titratable residues (histidine, glutamic acid, and aspartic acid) were assigned based on the default setting of the program *pdb2gmx* in GROMACS in combination with careful visual inspection of local hydrogen-bonded networks. *Pdb2gmx* took the protonation states of amino acids free in solvent at pH 7 as the default. The Lys and Arg were protonated, while the Asp and Glu were unprotonated. For His, the proton could be either on δ position, on ε position, or on both, and the selections were done automatically based on optimal hydrogen bonding conformations. Histidines 63 and 392 in both chain A and chain B were protonated at the δ position. Histidines 28, 29, 45, 62, 338, 369, and 432 at both chain A and chain B were protonated at the ε position. Cysteines 42, 69, and 72 in both chain A and chain B were deprotonated due to their binding to Zn^{2+}.

The simulations were carried out for three systems at 300 K, including WT, Com1-WT, and Com2-WT. At the beginning, the entire system was minimized using the steepest descent method with no positional restraints. The minimization process was configured to stop when the maximum force acting on any atom in the system fell below a threshold of 1000 kJ mol^{-1} nm^{-1}. An initial step size of 0.01 nm was set for the minimization process. The maximum number of minimization steps allowed was set to 5000. No other constraints were applied during the minimization. In the pre-equilibrium stage, the system was gradually heated to 300 K over 1 ns in the NVT ensemble, followed by 2 ns in the NPT ensemble at 1 atm. The potential energy curve and RMSD of the protein were analyzed to confirm the proper equilibration and stabilization of the system. During the production run, the NPT ensemble was employed for 50 ns, at 300 K and 1 atm. The RMSD of protein backbone atoms was calculated using the initial structure of the equilibrium run as the reference. Due to the irregular motion of the linker region affecting the RMSD calculation, as well as the fact that the linker area does not influence catalysis, the linker region was excluded from the RMSD calculation. The reaction distance, hydrogen bonds occupancy, and binding free energy were analyzed for the single chain, maintaining correct conformations of tRNA and substrate, using the last 10 ns equilibrated trajectories. The NPT ensemble was further employed for another 200 ns production simulations, and the dynamics cross-correlation map (DCCM) was calculated from the trajectory spanning from 50 to 200 ns, with data collected at 150 ps intervals.

The velocity-rescaling thermostat[50] with a time constant equal to 0.1 ps was employed throughout the simulations to keep the temperature constant. To maintain the pressure, the Parrinello-Rahman pressure coupling[51,52] was utilized in the equilibrium and production run, with the pressure time constant and isothermal compressibility set to 2 ps and 4.5×10^{-5} bar^{-1}, respectively. A time step of 2 fs for integration of the equations of motion was used throughout the simulation. A cutoff of 10 Å was used for nonbonded interactions. The particle mesh Ewald algorithm[53] was used to calculate long-range electrostatic interactions.

Mass spectrometry

Liquid chromatography-electrospray ionization mass spectrometry (LC-ESI-MS) was applied to detect the incorporation of ncAAs into sfGFP as described previously[30]. LC-ESI-MS analysis was performed on an Agilent 1290 Infinity II LC system coupled with a 6545 Q-TOF mass spectrometer (Agilent, UK). Protein samples (5 μL, 0.4 μg/μL) were injected onto a PLRP-S column (50 mm × 2.1 mm, 1000 Å, 5 μm) at 30 °C. A binary gradient using mobile phase A (5% MeCN, 0.1% formic acid) and B (95% MeCN, 0.1% formic acid) was applied at 0.3 mL/min. The column was equilibrated at 15% B for 1.9 min, held for 1 min, then ramped to 90% B over 16 min, followed by a rapid return to 15% B in 0.1 min. The Q-TOF scanned *m/z* 100–3100 using positive ESI with the following settings: capillary voltage 4000 V, nozzle 500 V, fragmentor 175 V, skimmer 65 V, and octopole RF peak 750 V. Nitrogen was used as nebulizer gas (45 psi, 5 L/min). Spectra were processed using MassHunter Bioconfirm (vB.10.00) and deconvolved via maximum entropy.

Zero-shot prediction of high-fitness variants

For ESM prediction, the publicly available ESM-1v scripts were used to retrieve "wt-marginals" for each of the five ESM-1v and ESM2 models. Mutations exhibiting enhanced probability scores relative to wild-type residues were systematically screened across all models. For Mutcompute prediction, the target protein's PDB ID was submitted to the web server (https://mutcompute.com). Results were retrieved via automated email delivery. Notably, the tool exclusively supports pre-existing crystallographic structures from the PDB database and cannot process custom-predicted structural models. For ProRefiner prediction, the package was deployed locally via its GitHub repository (https://github.com/veghen/ProRefiner). AlphaFold2-predicted variants' structures served as input to calculate mutant amino acid probabilities. Mutants with scores exceeding wild-type thresholds were prioritized for further analysis.

Analysis of epistasis between mutations

Equation (5) was used to quantify the epistasis between mutations.

$$\Delta NF_{ij} = \Delta NF_{exp} - (\Delta NF_i + \Delta NF_j) \tag{5}$$

Where NF is Normalized Fluorescence intensity of the sfGFP with ncAA incorporated at position 2 per OD_{600}, resulting from aminoacylation by PylRS and its mutants, ΔNF_{exp} is the difference in NF between the double variant and the wild type experimentally obtained, and ΔNF_i and ΔNF_j are the differences in the NF between single variants and the wild type. In any case, additivity occurs when $\Delta NF_{ij} = 0$. Positive magnitude epistasis (+ME) or negative magnitude epistasis (−ME) occurs when $\Delta NF_{ij} > 0$ or $\Delta NF_{ij} < 0$, respectively. Positive sign epistasis (+SE) or negative sign epistasis (−SE) occurs when $\Delta NF_{ij} > 0$ if $\Delta NF_i < 0$ or $\Delta NF_j < 0$, or $\Delta NF_{ij} < 0$ if $\Delta NF_i > 0$ or $\Delta NF_j > 0$, respectively. Positive reciprocal sign epistasis (+ RSE) or negative reciprocal sign epistasis (-RSE) occurs when $\Delta NF_{ij} > 0$ if $\Delta NF_i < 0$ and $\Delta NF_j < 0$, or $\Delta NF_{ij} < 0$ if $\Delta NF_i > 0$ and $\Delta NF_j > 0$, respectively[54].

Statistics and reproducibility

Data were analyzed using Prism (GraphPad) and Origin 2025. Data are presented as mean ± standard deviation with error bars. All experiments were repeated independently with similar results at least three times. No data were excluded from the analyses. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Authors declare that all data supporting the findings of this study are available within the paper and its supplementary information files. The PDB structures used in this work include 5UD5, 4TQD, and 2ZIM. The LC-MS data underlying Supplementary Fig. 16 have been deposited to the ProteomeXchange Consortium via the PRIDE[55] partner repository with the dataset identifier PXD065336 [https://www.ebi.ac.uk/pride/archive/projects/PXD058768]. The enzyme data has been provided in two JSON format files prepared with the EnzymeML tool[56], which are available at Github [https://github.com/zjuhaoran/FPFORCOM] and at Zenodo[57]. Source data are provided with this paper.

Code availability

The source code employed for generating descriptors and training ML models in this research is available at https://github.com/zjuhaoran/FPFORCOM. The source code has also been deposited to Zenodo[57].

References

1. Shandell, M. A., Tan, Z. & Cornish, V. W. Genetic code expansion: a brief history and perspective. *Biochemistry* **60**, 3455–3469 (2021).
2. Wan, W., Tharp, J. M. & Liu, W. R. Pyrrolysyl-tRNA synthetase: an ordinary enzyme but an outstanding genetic code expansion tool. *Biochim. Biophys. Acta* **1844**, 1059–1070 (2014).
3. Koch, N. G. & Budisa, N. Evolution of pyrrolysyl-tRNA synthetase: from methanogenesis to genetic code expansion. *Chem. Rev.* **124**, 9580–9608 (2024).
4. Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Genetically encoding N(epsilon)-acetyllysine in recombinant proteins. *Nat. Chem. Biol.* **4**, 232–234 (2008).
5. Tawfik, D. S. & Gruic-Sovulj, I. How evolution shapes enzyme selectivity—lessons from aminoacyl-tRNA synthetases and other amino acid utilizing enzymes. *FEBS J.* **287**, 1284–1305 (2020).
6. Suzuki, T. et al. Crystal structures reveal an elusive functional domain of pyrrolysyl-tRNA synthetase. *Nat. Chem. Biol.* **13**, 1261–1266 (2017).
7. Bryson, D. I. et al. Continuous directed evolution of aminoacyl-tRNA synthetases. *Nat. Chem. Biol.* **13**, 1253–1260 (2017).
8. Sharma, V. et al. Evolving the N-terminal domain of pyrrolysyl-tRNA synthetase for improved incorporation of noncanonical amino acids. *ChemBioChem* **19**, 26–30 (2018).
9. Liu, K. et al. An evolved pyrrolysyl-tRNA synthetase with polysubstrate specificity expands the toolbox for engineering enzymes with incorporation of noncanonical amino acids. *Bioresour. Bioprocess.* **10**, 92 (2023).
10. Willis, J. C. W. & Chin, J. W. Mutually orthogonal pyrrolysyl-tRNA synthetase/tRNA pairs. *Nat. Chem.* **10**, 831–837 (2018).
11. Dunkelmann, D. L., Willis, J. C. W., Beattie, A. T. & Chin, J. W. Engineered triply orthogonal pyrrolysyl–tRNA synthetase/tRNA pairs enable the genetic encoding of three distinct non-canonical amino acids. *Nat. Chem.* **12**, 535–544 (2020).
12. Taylor, C. J. et al. Engineering mutually orthogonal PylRS/tRNA pairs for dual encoding of functional histidine analogues. *Protein Sci.* **32**, e4640 (2023).
13. Yu, H. & Dalby, P. A. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proc. Natl. Acad. Sci. USA* **115**, E11043–E11052 (2018).
14. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25**, 1260–1272 (2016).
15. Johnston, K. E. et al. A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proc. Natl. Acad. Sci. USA* **121**, e2400439121 (2024).
16. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. USA*. **116**, 8852–8858 (2019).
17. Cadet, F. et al. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci. Rep.* **8**, 16757 (2018).
18. Li, G., Jia, L., Wang, K., Sun, T. & Huang, J. Prediction of thermostability of enzymes based on the amino acid index (AAindex) database and machine learning. *Molecules* **28**, 8097 (2023).
19. Owens, A. E., Grasso, K. T., Ziegler, C. A. & Fasan, R. Two-tier screening platform for directed evolution of aminoacyl-tRNA synthetases with enhanced stop codon suppression efficiency. *ChemBioChem* **18**, 1109–1116 (2017).
20. Jiang, H.-K. et al. Linker and N-terminal domain engineering of pyrrolysyl-tRNA synthetase for substrate range shifting and activity enhancement. *Front. Bioeng. Biotechnol.* **8**, 235 (2020).
21. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* **34**, 29287–29303 (2021).
22. Shroff, R. et al. Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synth. Biol.* **9**, 2927–2935 (2020).
23. Zhou, X. et al. ProRefiner: an entropy-based refining strategy for inverse protein folding with global graph attention. *Nat. Commun.* **14**, 7434 (2023).
24. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
25. van der Meer, J.-Y., Biewenga, L. & Poelarends, G. J. The generation and exploitation of protein mutability landscapes for enzyme engineering. *ChemBioChem* **17**, 1792–1799 (2016).
26. Hecht, M., Bromberg, Y. & Rost, B. News from the protein mutability landscape. *J. Mol. Biol.* **425**, 3937–3948 (2013).
27. Guo, L.-T. et al. Polyspecific pyrrolysyl-tRNA synthetases from directed evolution. *Proc. Natl. Acad. Sci. USA* **111**, 16724–16729 (2014).
28. Englert, M. et al. Probing the active site tryptophan of Staphylococcus aureus thioredoxin with an analog. *Nucleic Acids Res.* **43**, 11061–11067 (2015).
29. Burke, A. J. et al. Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature* **570**, 219–223 (2019).
30. Zhang, Q. et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat. Commun.* **16**, 1553 (2025).
31. He, Y. et al. Protein language models-assisted optimization of a uracil-N-glycosylase variant enables programmable T-to-G and T-to-C base editing. *Mol. Cell* **84**, 1257–1270.e6 (2024).
32. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
33. Verkuil, R. et al. Language models generalize beyond natural proteins. Preprint at *BioRxiv* https://doi.org/10.1101/2022.12.21.521521 (2022).
34. Jiang, K. et al. Rapid in silico directed evolution by a protein language model with EVOLVEpro. *Science* **387**, eadr6006 (2025).
35. Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
36. Paik, I. et al. Improved bst DNA polymerase variants derived via a machine learning approach. *Biochemistry* **62**, 410–418 (2023).
37. Kunka, A. et al. Advancing enzyme's stability and catalytic efficiency through synergy of force-field calculations, evolutionary analysis, and machine learning. *ACS Catal.* **13**, 12506–12518 (2023).
38. Sumida, K. H. et al. Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
39. Ding, W. et al. Chimeric design of pyrrolysyl-tRNA synthetase/tRNA pairs and canonical synthetase/tRNA pairs for genetic code expansion. *Nat. Commun.* **11**, 3154 (2020).
40. Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. The N-end rule in bacteria. *Science* **254**, 1374–1377 (1991).

41. Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Genetically encoding Nε-acetyllysine in recombinant proteins. *Nat. Chem. Biol.* **4**, 232–234 (2008).
42. Pott, M. et al. A noncanonical proximal heme ligand affords an efficient peroxidase in a globin fold. *J. Am. Chem. Soc.* **140**, 1535–1543 (2018).
43. Mondal, S., Hsiao, K. & Goueli, S. A. Utility of adenosine monophosphate detection system for monitoring the activities of diverse enzyme reactions. *ASSAY Drug Dev. Technol.* **15**, 330–341 (2017).
44. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **28,** 374 (2000).
45. Kawashima, S. et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202–D205 (2007).
46. Mckenna, A. & Dubey, S. Machine learning based predictive model for the analysis of sequence activity relationships using protein spectra and protein descriptors. *J. Biomed. Inform.* **128**, 104016 (2022).
47. Abraham, M. J. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
48. Ivani, I. et al. Parmbsc1: a refined force field for DNA simulations. *Nat Methods* **13**, 55–58 (2016).
49. Sousa Da Silva, A. W. & Vranken, W. F. ACPYPE—AnteChamber PYthon parser interfacE. *BMC Res. Notes* **5**, 367 (2012).
50. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
51. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
52. Nosé, S. & Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **50**, 1055–1076 (1983).
53. Herce, H. D., Garcia, A. E. & Darden, T. The electrostatic surface term:(I) periodic systems. *J. Chem. Phys.* **126**, 124106 (2007).
54. Acevedo-Rocha, C. G. et al. Pervasive cooperative mutational effects on multiple catalytic enzyme traits emerge via long-range conformational dynamics. *Nat. Commun.* **12**, 1621 (2021).
55. Perez-Riverol, Y. et al. The PRIDE database at 20 years: 2025 update. *Nucleic Acids Res.* **53**, D543–D553 (2025).
56. Lauterbach, S. et al. EnzymeML: Seamless data flow and modeling of enzymatic data. *Nat. Methods* **20**, 400–402 (2023).
57. Zhang, Q. et al. Machine learning-guided evolution of pyrrolysyl-tRNA synthetase for improved incorporation efficiency of diverse noncanonical amino acids. zjuhaoran/FPFORCOM: https://doi.org/10.5281/zenodo.15690090 (2025).

Acknowledgements

This work was supported by the Key Research and Development Program of China (Grant No. 2022YFA0913000), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2025C01097), and the National Natural Science Foundation of China (Grant Nos. 22108245 and 22378351) to H.Y. We thank the AI + High Performance Computing Center of ZJU-ICI. We would like to thank iBioFoundry and the Core Facility at the Institute for Intelligent Bio/Chem Manufacturing, ZJU-Hangzhou Global Scientific and Technological Innovation Center.

Author contributions

H.Y. and Q.Z. designed research; Q.Z., Y.N., Y.L., W.C., J.C., H.D., K.L., S.Y., J.C., and J.W. performed studies; B.C. and L.J. provided guidance on molecular dynamics simulation; L.Y., J.W., G.X., J.L., and H.Y. supervised the studies; H.Y. and Q.Z. analyzed data; H.Y., L.J., and Q.Z. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-61952-2.

Correspondence and requests for materials should be addressed to Haoran Yu.

Peer review information *Nature Communications* thanks Carlos Acevedo-Rocha and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.