nature portfolio

Peer Review File

Integrating Protein Language Models and Automatic Biofoundry for Enhanced Protein Evolution

Corresponding Author: Professor Haoran Yu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Title: Integrating Protein Language Models and Automatic Biofoundry for Enhanced Protein Evolution

Concerns: Zhang, Chen, Qin, et al.

The authors report an end-to-end protein engineering workflow which fully automates the design-build-test-learn cycle of synthetic biology. Central to their technology is a computational platform which leverages a protein language model (PLM). The model, which they call PLMeAE, initiates a protein evolution campaign by making zero-shot predictions to give an initial set of informative variants for testing. In successive rounds, the model combines data taken from testing of this initial set of mutants to train a fitness predictor which combines a multi-layer perceptron with the aforementioned PLM. The authors provide two implementations of the model: 1) when there is no prior knowledge of a protein scaffold then PLMeAE both predicts sites which can be simultaneously mutated and provides initial mutations at these positions (Module I) and 2) when the protein engineer knows sites which are relevant to the desired activity then PLMeAE will predict initial multi-mutants at these positions (Module II). After testing of an initial set of predicted variants, PLMeAE updates the model and predicts new multi-mutants based on the collected data.

Along with their machine learning implementation the authors describe an automated 'biofoundry' platform whose only input is sequences predicted by PLMeAE. The platform clones predicted sequences and is equipped to test protein fitness for functions which can be screened by fluorescence or colorimetric assays. In this report, the authors validate their model on the GB1 landscape before experimentally testing it through the evolution of a tRNA synthetase (pCNF-RS) for the incorporation of noncanonical amino acids (ncAA). In 3 rounds of evolution on previously reported sites in pCNF-RS the synthetase activity is improved approximately 2 fold. After using the model to predict previously untested positions in the sequence, the function is further improved to 2.4 fold from parent.

The content of the manuscript is well thought out and the findings will be of significant interest to the protein engineering community. Particularly, the insights on the application of PLMs for the discovery of mutations relevant to a desired function for an understudied scaffold could be of great value. However, it is my opinion that the manuscript does not properly provide benchmarking experiments such that the reader can assess the benefits of using such a platform. Additionally, I find that the language structure and grammar throughout the article are not of high enough quality for publication in Nature Communications in its current state. Several sentences in the manuscript are challenging to understand. A response to these comments and corrections to grammar and structure in the manuscript will be necessary before acceptance.

Major Points:

1. In the application of PLMeAE to the evolution of pCNF-RS, the authors report improvements from parent and function statistics of tested libraries, with mutagenesis on positions 283, 284, 285, and 286 resulting in incremental improvements in activity up to a maximum of 2.1x parent over 3 rounds (Figure 5b). However, it is not made evident how challenging a task it would be to arrive at these improvements in activity simply through traditional directed evolution. The authors cite Gan et al. regarding previous engineering of pCNF-RS. A discussion of the extent of engineering described in that study would help to provide context for the effectiveness of the PLM. Furthermore, all of the variants screened throughout this campaign were

multi-mutants predicted by PLMeAE. While the model does predict improved variants in every round, there is no comparison to a randomly constructed multi-site library at the same positions. This random library will help benchmark the efficacy of the model for predicting improved variants.
2. The authors describe Module I for discovering new sites to mutate and Module II for predicted mutations and known positions. One way to test the effectiveness of Module I could be to apply it to proteins previously engineered in the literature and see how likely it is to suggest positions which proved essential to those campaigns. In the application of Module I to pCNF-RS, it is found that previously untested sites were predicted. However, the overall boost in activity found in the top mutant in this library relative to the best variant from round 3 (Figure 5c) is only 1.2x. Is this a significant enough improvement to suggest that generating mutants through Module I is more effective than screening of a random mutant library generated through error-prone PCR?

Minor Points:

1. Figure 3c seems to indicate that the authors considered one-, two-, three-, and four-point mutants in PLM selection. Firstly, this is an incorrect usage of the term 'point mutant'. This term specifically refers to a single base mutation at the DNA or RNA level. The authors should instead use 'single, double, triple, and quadruple mutants'. Secondly, as a result of this figure it is not clear until much later in the manuscript whether PLMeAE is a tool for predicting single mutant or multi-mutant variants. This point should be clarified much earlier, for example in the model description in lines 119-135 or in the introduction.
2. In lines 419-432 the authors describe the improvements in time spent evolving variants. A further discussion of how the platform could be applied to enzymatic functions which require non-colorimetric assays would help to strengthen the manuscript. This is briefly mentioned in line 473, but further elaboration on how the biofoundry could integrate with other screens would be appreciated.
3. In lines 454-463, the authors correctly assess that, for a non-natural function, a PLM-based predictor may not contain the necessary information to provide improved variants. The authors should further elaborate the capability of PLMeAE to navigate a fitness landscape for a protein function which is distant from that of the native protein. To what extent would the supervised machine learning model be able to predict variants which are strongly disfavored by the PLM element?
4. In lines 593-595, the authors state that mutagenesis primers were designed online and that sequences are provided in the SI. However, only primers for PCR external to the pCNF-RS sequence and mutagenesis of sfGFP. Some example primers which were generated through the online tool should be shared.
5. Sequences for all plasmids and sequences used in this work should be provided in the supplementary information or cited.
6. A cartoon representation of the functional screen used for assessing pCNF-RS variant fitness would add clarity to this work. The current description in the main text (lines 284-384) is thorough; however, addition of a figure describing the biochemistry of the screen in the SI would be useful.
7. The image qualities of figures 1, 2, and 4 are not high enough in the provided documents. These should be enhanced before publication.

Some statements in the manuscript seem poorly justified:

1. In lines 237-238, it is stated that site-directed mutagenesis is the most widely used approach in protein engineering, without a citation. I don't believe this is true. It seems that the authors do not consider protein engineering methods which leverage random mutagenesis.
2. In lines 247-250, the transformation process in the automated biofoundry is described. It is not clear to me that there is an assembly step for the PCR products that are being transformed. The E. coli strain used in this work, BL21(DE3), is not capable of circularizing linear DNA through homologous end joining. A description of how the generated PCR products are assembled into a plasmid is required in the main text.
3. In lines 430-432, the authors state that they estimate an equivalent directed evolution campaign would take 3-6 months to complete. Since one of the strongest points of this workflow is its speed, the authors should clarify the assumptions they made in arriving at this 3-6 month figure to justify their assertion.
4. In line 379-380, the authors posit that some mutations in M-R3 and M-R4 influenced substrate specificity. From the way these sentences are written it is not clear if the authors are referring specifically to mutations in M-R3, M-R4, or to mutations in both. It should be clearly stated which mutations are believed to be impacting specificity.

(Remarks on code availability)
A README file is provided, and I was able to download and run the code. The code seems appropriately written, however commenting in the provided Python scripts is is scarce. Annonatation of the code would allow for a better understanding of the steps being performed by the model.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript by Zhang and co-workers describes an autonomous system integrating protein language models with automatic biofoundary. The developed system was applied to change the substrate specificity of tRNA synthetase in four rounds. The manuscript is interesting and suitable for Nature Communications. This work represents an important step towards developing fully automated protein engineering systems.
Comments:
26 Avoid abbreviations in the Abstract.
35 Add improvement in activity to the Abstract.
37 Novel proteins … I do not think these are novel proteins
46 directed evolution is not emerging, it is well-established
49 discuss also focused directed evolution
51 costly … small-but-smart libraries are preferably used at these days
53 other software for protein design: https://loschmidt.chemi.muni.cz/portal/
102 please avoid superlative writing … transformative approach
116 within half a month … can you discuss the robustness of your system … how many cycles can be performed in a row?
151 please add A, B, C to the figure for clarity
181 How were 4 residues selected?
201 provide references
209 improvement among three cycles is hardly visible
221 VERY IMPORTANT: The data obtained by individual cycles should be reported systematically, e.g., in the Table. Not only maximum values should be reported, but also all mutants with improved activities. This is essential to provide this data.
266 standard error 5% is impressive, congratulations
Figure 4 providing timing would make this figure more useful, consider also adding throughput
315 it is again important to report if these 7 were activating or neutral, report these data in a Table
324 the great effect … superlative writing
327 provide sequence similarity/diversity values in a table
395 the fonts are too small in Figure 6
419 not sure the system is intelligent
426 report also the total number of mutants
431 under half a month … can you be more precise?
488 I would appreciate some comments on the generalizability of the system, on its pros and cons, on future development

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Revisions are appropriate and address my earlier comments.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors addressed all my comments thoroughly and properly. I am very pleased with this new manuscript version and support its publishing by Nature Communications. Sincere congratulations!

(Remarks on code availability)

Reviewer #1:

Title: Integrating Protein Language Models and Automatic Biofoundry for Enhanced Protein Evolution

Concerns: Zhang, Chen, Qin, et al.

The authors report an end-to-end protein engineering workflow which fully automates the design-build-test-learn cycle of synthetic biology. Central to their technology is a computational platform which leverages a protein language model (PLM). The model, which they call PLMeAE, initiates a protein evolution campaign by making zero-shot predictions to give an initial set of informative variants for testing. In successive rounds, the model combines data taken from testing of this initial set of mutants to train a fitness predictor which combines a multi-layer perceptron with the aforementioned PLM. The authors provide two implementations of the model: 1) when there is no prior knowledge of a protein scaffold then PLMeAE both predicts sites which can be simultaneously mutated and provides initial mutations at these positions (Module I) and 2) when the protein engineer knows sites which are relevant to the desired activity then PLMeAE will predict initial multi-mutants at these positions (Module II). After testing of an initial set of predicted variants, PLMeAE updates the model and predicts new multi-mutants based on the collected data.

Along with their machine learning implementation the authors describe an automated 'biofoundry' platform whose only input is sequences predicted by PLMeAE. The platform clones predicted sequences and is equipped to test protein fitness for functions which can be screened by fluorescence or colorimetric assays. In this report, the authors validate their model on the GB1 landscape before experimentally testing it through the evolution of a tRNA synthetase (pCNF-RS) for the incorporation of noncanonical amino acids (ncAA). In 3 rounds of evolution on previously reported sites in pCNF-RS the synthetase activity is improved approximately 2 fold. After using the model to predict previously untested positions in the sequence, the function is further improved to 2.4 fold from parent.

The content of the manuscript is well thought out and the findings will be of significant interest to the protein engineering community. Particularly, the insights on the application of PLMs for the discovery of mutations relevant to a desired function for an understudied scaffold could be of great value. However, it is my opinion that the manuscript does not properly provide benchmarking experiments such that the reader can assess the benefits of using such a platform. Additionally, I find that the language structure and grammar throughout the article are not of high enough quality for publication in Nature Communications in its current state. Several sentences in the manuscript are challenging to understand. A response to these comments and corrections to grammar and structure in the manuscript will be necessary before acceptance.

RESPONSE： We thank the reviewer for the constructive feedback. We have made general improvements throughout the manuscript following your comments. The questions and suggestions have been addressed point-by-point below. Revisions to the main text are highlighted in red font.

Major Points:

1. In the application of PLMeAE to the evolution of pCNF-RS, the authors report improvements from parent and function statistics of tested libraries, with mutagenesis on positions 283, 284, 285, and 286 resulting in incremental improvements in activity up to a maximum of 2.1x parent over 3 rounds (Figure 5b). However, it is not made evident how challenging a task it would be to arrive at these improvements in activity simply through traditional directed evolution. The authors cite Gan et al. regarding previous engineering of pCNF-RS. A discussion of the extent of engineering described in that study would help to provide context for the effectiveness of the PLM. Furthermore, all of the variants screened throughout this campaign were multi-mutants predicted by PLMeAE. While the model does predict improved variants in every round, there is no comparison to a randomly constructed multi-site library at the same positions. This random library will help benchmark the efficacy of the model for predicting improved variants.

RESPONSE： Many thanks for your comments and suggestions. We have given a discussion about previous engineering of pAzFRS to illustrate how challenging it was to engineer the enzyme. Additionally, we constructed a random mutation library targeting on positions 283, 284, 285, and 286 and randomly picked 90 variants for testing enzyme activity. We compared the effect of random selection and PLMeAE, and gave a detailed discussion in the main text. The added sentences are as follows.

"A previous study has modified four positions H283, P284, M285, and D286 of the MjTyrRS mutant pAzFRS, aiming to improve the efficiency of introducing *p*-azidophenylalanine (*p*AzF). Since modifying substrate binding pocket of MjTyrRS could lead to mutants exhibiting misincorporation of nature amino acids, a positive-negative selection procedure was carried out to obtain improved variants in the mutation library containing a theory of 160 000 mutations. In the positive selection, functional aminoacyl-tRNA synthetase (aaRS) constructs are tested *in vivo* for their capability to suppress an in-frame amber codon in a chloramphenicol acetyl transferase (*cat*) reporter, which provides resistance to chloramphenicol. In the negative selection, synthetases that incorporate a natural amino acid are selected against using the toxin protein barnase. The plasmid expressing aaRS mutation library was first transformed into *E. coli* cells containing plasmid expressing *cat* gene for positive screening. After that, the plasmid expressing aaRS mutants was isolated from surviving colonies and separated from positive screening plasmid, and then transformed into *E. coli* cells containing plasmid expressing barnase for negative screening. This process is complex, time-consuming and laborious."

"To further illustrate the effect of PLMeAE in engineering pCNF-RS, we constructed a random mutation library targeting on residues H283, P284, M285, and D286 using primers containing four consecutive NNK codons. We then randomly picked 192 colonies into two 96-well plates for sequencing and testing for enzyme activity. Ninety variants with different sequences were finally obtained, among which 82 showed lower activity, 6 showed similar activity and 2 showed higher activity compared to the wild type. The two improved variants H283P/P284A/M285C/D286L and H283A/M285L/D286G improved the enzyme activity by 1.26-fold and 1.20-fold, respectively, which was comparable with M-R1, while significantly lower than that of M-R2, M-R3 and M-R4. The positive rate of random selection was around 2.2%, similar with that that of zero-shot predictions using PLM, but was significantly lower than 50% of second-round and 62.5% of third-round PLMeAE. This demonstrates the superior effectiveness of PLMeAE over random selection. Although the positive rate of random selection is similar to zero-shot predictions of PLM, it has been shown that the training data based on random sampling does not perform well in building supervised ML models"

2. The authors describe Module I for discovering new sites to mutate and Module II for predicted mutations and known positions. One way to test the effectiveness of Module I could be to apply it to proteins previously engineered in the literature and see how likely it is to suggest positions which proved essential to those campaigns. In the application of Module I to pCNF-RS, it is found that previously untested sites were predicted. However, the overall boost in activity found in the top mutant in this library relative to the best variant from round 3 (Figure 5c) is only 1.2x. Is this a significant enough improvement to suggest that generating mutants through Module I is more effective than screening of a random mutant library generated through error-prone PCR?

RESPONSE： Thanks for your comments and suggestions. It was true that the activity of M-R4 obtained through Module I was only 1.2-fold higher than M-R3. This was partially due to the activity measurement method used, which was based on fluorescence intensity of sfGFP reporter. The improvement was indeed significant as the fluorescence of sfGFP2TAG in presence of M-R4 nearly reached the level of wild-type sfGFP. To further illustrate the effectiveness of Module I in identifying potential mutants, we tested the PLM model in two datasets, the SUMO-conjugating enzyme UBC9 dataset[36] and the Ubiquitin dataset[37], which contain activity data of almost all possible single variants of the two proteins. ESM-2 was used to predict top 96 single variants based on the two protein sequences, while other 96 variants were also randomly selected for comparison. As for UBC9 variants, the maximum fitness data obtained by EMS-2 prediction was 2.35, much higher than 1.72 achieved by random selection. Moreover, the mean fitness value of all variants predicted by ESM-2 was 0.53, higher than 0.43 of those randomly selected (Figure S1). Similarly, for ubiquitin, the mean fitness of all variants predicted by ESM-2 was 0.75, significantly higher than the 0.44 of those randomly selected, although the random selection achieved a

maximum fitness value similar to that of ESM-2. We also found that high-fitness variants predicted by ESM-2 provided hot spots for further engineering. For example, E18C and D32A of ubiquitin predicted by ESM-2 ranked 44th and 87th in the dataset, respectively. If these two amino acids were selected for further engineering, D32K and E18M would be obtained, which ranked 13th and 19th in the dataset, respectively, indicating the potential of EMS-2 in identifying positions essential for protein engineering. We have added above discussion in the main text, which is highlighted in red.

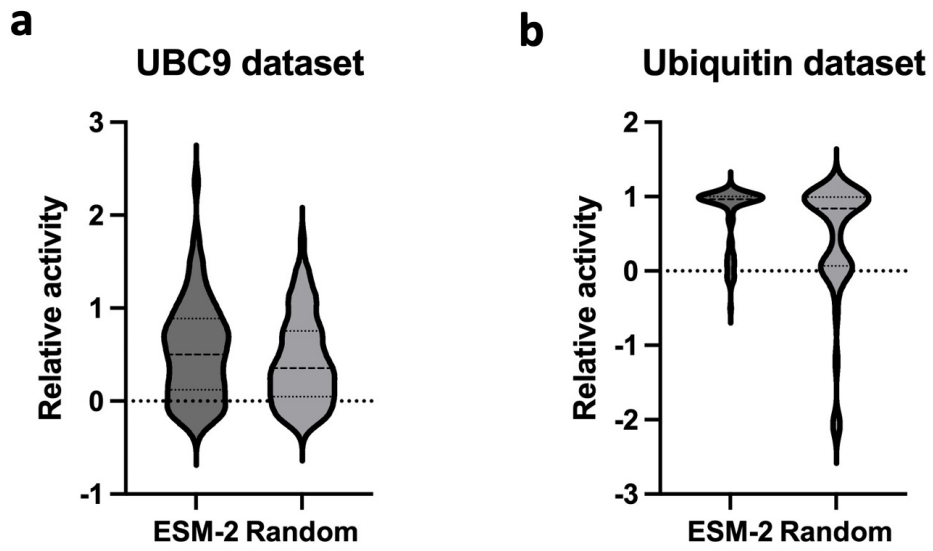

Figure S1 The performance of ESM-2 in predicting top 96 singe variants of SUMO-conjugating enzyme UBC9 and ubiquitin. Random selection was also carried out for comparison. UBC9 dataset contains activity data of 2563 single variants accounting for 85% of all possible single variants. Ubiquitin dataset contains activity data of E1 enzyme against 1528 single variants of ubiquitin. Since the datasets did not cover all possible singe variants, the label of some variants predicted were missing. For the UBC9 dataset, the activity data of 84 variants predicted and 84 variants randomly selected are shown. For ubiquitin dataset, the activity data of 96 variants predicted and 91 variants randomly selected are shown.

Minor Points:

1. Figure 3c seems to indicate that the authors considered one-, two-, three-, and four-point mutants in PLM selection. Firstly, this is an incorrect usage of the term ‘point mutant’. This term specifically refers to a single base mutation at the DNA or RNA level. The authors should instead use ‘single, double, triple, and quadruple mutants’. Secondly, as a result of this figure it is not clear until much later in the manuscript whether PLMeAE is a tool for predicting single mutant or multi-mutant variants. This

point should be clarified much earlier, for example in the model description in lines 119-135 or in the introduction.

RESPONSE： Thanks for your comments. We have corrected the legend of Figure 3c, with “single, double, triple, and quadruple mutants” to replace “one-point, two point, three-point and four-point mutation”. We also clarified that PLM would be used to predict singe mutant or multi-mutant variants in the beginning of “Results” section. The revised sentence is as follows:

“In specific, two zero-shot tasks were solved by the PLMs, depending on the availability of mutation target sites. Firstly, with no prior information about the target protein, PLMs were used to predict high-fitness single mutants in a zero-shot setting. Secondly, when the mutation sites were already identified based on previous experiments or through physical modelling techniques such as docking, molecular dynamics simulations, the PLMS were used to predict zero-shot high-fitness multi-mutant variants at the given target sites.”

2. In lines 419-432 the authors describe the improvements in time spent evolving variants. A further discussion of how the platform could be applied to enzymatic functions which require non-colorimetric assays would help to strengthen the manuscript. This is briefly mentioned in line 473, but further elaboration on how the biofoundry could integrate with other screens would be appreciated.

RESPONSE： Thanks for your suggestion. We have added a paragraph to discuss how the platform could be applied to engineer enzymes which require non-colorimetric assays.

“The PLMeAE platform has the potential to expand applications to engineer enzymes with activity measured by liquid chromatography (LC), gas chromatography (GC) and mass spectrometry (MS). Various automation set up has been developed to accelerate chromatographic and mass spectrometric analysis. For example, a number of liquid chromatography systems are equipped with 96-well plate autosamplers, which are easily operated by robotic arms and scheduling software. Additionally, employing a low-pressure GC configuration combined with low thermal mass (LTM) ovens allows for a cycle time less than 1 min for detecting samples such as fatty acid methyl esters. Various instruments, such as liquid handler robots, microfluidics, and automated solvent extraction equipment, have been used to automate sample preparation for chromatographic analysis. Recent advances also promote the use of high-throughput MS by enabling direct analyte introduction into mass spectrometers using laser, microfluidics, and acoustics. An automated matrix-assisted laser desorption/ionization time of flight (MALDI-ToF) MS has been developed to directedly analyze bacterial and yeast colonies at a rate of ~2 s per sample. A robotic platform was also developed to integrate a liquid handler, wash station, plate dryer, sample stacker, and mass spectrometer, enabling fully automated MALDI MS analysis and

allowing for the analysis of over a million samples in a week. Integrating these automated, high-throughput assays into the BioFoundry would significantly expand the application of PLMeAE in engineering other enzymes".

3. In lines 454-463, the authors correctly assess that, for a non-natural function, a PLM-based predictor may not contain the necessary information to provide improved variants. The authors should further elaborate the capability of PLMeAE to navigate a fitness landscape for a protein function which is distant from that of the native protein. To what extent would the supervised machine learning model be able to predict variants which are strongly disfavored by the PLM element?

RESPONSE： This is a good question. As for the Module II of the PLMeAE system, the function of PLM model is to predict zero-shot high-fitness multi-mutant variants, which is then used as dataset for training a supervised machine learning model. It is true that PLM might predict the variants with a low accuracy if the fitness landscape is distant from that of the native protein, leading to low or even no positive variants in the training dataset. However, since we use an Information Transport Complexity (ITC)-based sampling strategy, the variants predicted by the PLM have a high diversity in sequence, which is helpful for supervised ML models to learn the fitness landscape. Additionally, since only several amino acids are selected as mutation targets, the sequence space is limited. Hence, it is still highly likely to obtain improved variants using PLMeAE, although more rounds of evolution may be needed for such a challenging setting. We have added above discussion in the revised manuscript.

"As for the more challenging task that the target fitness landscape is distant from that of the native protein, PLM would predict the variants with a very low accuracy, leading to no positive variants in the training dataset. However, since we use an ITC-based sampling strategy, the variants predicted by the PLM have a high diversity in sequence, which is helpful for supervised ML model to learn the fitness landscape. Additionally, since only several amino acids are selected as mutation targets, the sequence space is limited. It is still highly possible to obtain improved variants using PLMeAE, although more rounds of evolution may be needed for such a challenging task."

4. In lines 593-595, the authors state that mutagenesis primers were designed online and that sequences are provided in the SI. However, only primers for PCR external to the pCNF-RS sequence and mutagenesis of sfGFP. Some example primers which were generated through the online tool should be shared.

RESPONSE： Thank you for your suggestion. All primer sequences for constructing pCNF-RS variants were provided in Table S9 in the supplementary information.

5. Sequences for all plasmids and sequences used in this work should be provided in the supplementary information or cited.

RESPONSE： Thanks for your suggestion. All protein sequences used in this work are listed in supplementary information Table S5. The plasmids such as pULTRA-CNF, pQR711 used in this work have been cited.

6. A cartoon representation of the functional screen used for assessing pCNF-RS variant fitness would add clarity to this work. The current description in the main text (lines 284-384) is thorough; however, addition of a figure describing the biochemistry of the screen in the SI would be useful.

RESPONSE： Thanks for your suggestions, a cartoon describing the biochemistry of the screen was provided in Figure S6.

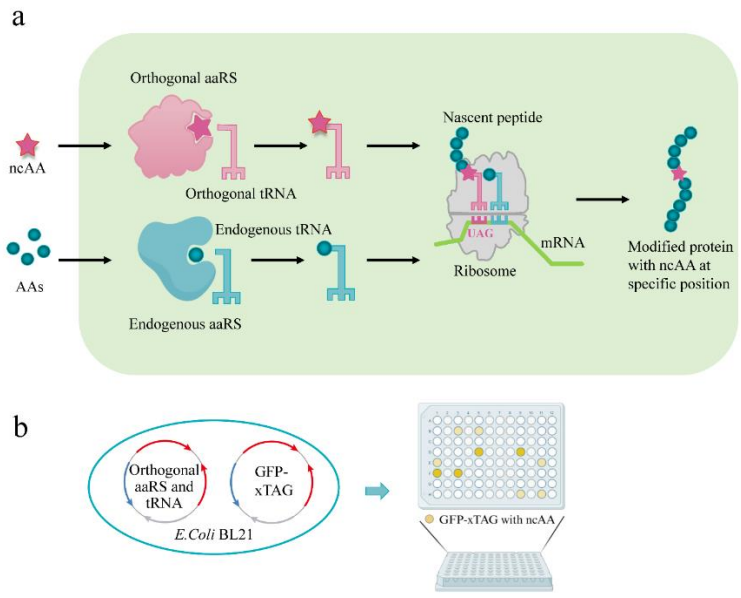

Figure S6 Measurement of pCNF-RS amber codon suppressing efficiency based on fluorescence intensity of sfGFP containing an amber codon. a, A diagram illustrating how a ncAA is incorporated into a protein using orthogonal aaRS/tRNA pair. b, Suppression of the sfGFP-UAG2 gene by ncAA investigated by measuring their fluorescence intensities. The enzyme activity is presented by fluorescence intensity. Created in BioRender. Yu, H. (2024) https://BioRender.com/m58k006

7. The image qualities of figures 1, 2, and 4 are not high enough in the provided documents. These should be enhanced before publication.

RESPONSE： Thanks. Qualities of these figures have been improved.

Some statements in the manuscript seem poorly justified:
1. In lines 237-238, it is stated that site-directed mutagenesis is the most widely used approach in protein engineering, without a citation. I don't believe this is true. It seems that the authors do not consider protein engineering methods which leverage random mutagenesis.

RESPONSE： Thanks for your comments. We have corrected the sentence as follows.

"Site-directed mutagenesis is a commonly used approach in protein engineering[37]"

2. In lines 247-250, the transformation process in the automated biofoundry is described. It is not clear to me that there is an assembly step for the PCR products that are being transformed. The E. coli strain used in this work, BL21(DE3), is not capable of circularizing linear DNA through homologous end joining. A description of how the generated PCR products are assembled into a plasmid is required in the main text.

RESPONSE： Thanks for your comments. We apologize for the misunderstanding**.** Actually, there is no assembly step in the process for the site-directed mutagenesis. Instead, a QuikChange method was used to carry out the site-directed mutagenesis with the entire plasmid as template of the PCR. We have given a description about this process in the text as follows.

"The QuikChange method uses 30-35 base megaprimers with the intended mutation-bearing bases in the middle to amplify the entire recombinant plasmid. The methylated parent plasmids are removed using *Dpn*I treatment before transformation."

3. In lines 430-432, the authors state that they estimate an equivalent directed evolution campaign would take 3-6 months to complete. Since one of the strongest points of this workflow is its speed, the authors should clarify the assumptions they made in arriving at this 3-6 month figure to justify their assertion.

RESPONSE： Many thanks for your suggestions. We have corrected this sentence and clarify the assumptions. The added sentences are as follows.

"Previously, a combinatorial mutation library was constructed at H283, P284, M285, D286 of pAzFRS and screened using a strategy of positive and negative selections[43]. After two rounds of positive-negative selections using *cat*/barnase system, they obtained improved variants against pAzF. The whole process took around 5 days for preparations, 7 days for one round positive selection, 7 days for one round negative selection and 5 days for final screening based on fluorescence, 38 days in total for two rounds of selections (Figure S5). However, experimental failure could happen at steps such as mutation library construction, cells transformation, plasmid Midiprep and

isolation. Additionally, more rounds of positive-negative screening might be needed to obtain improved variants. This would result in the whole process taking much longer in real situation."

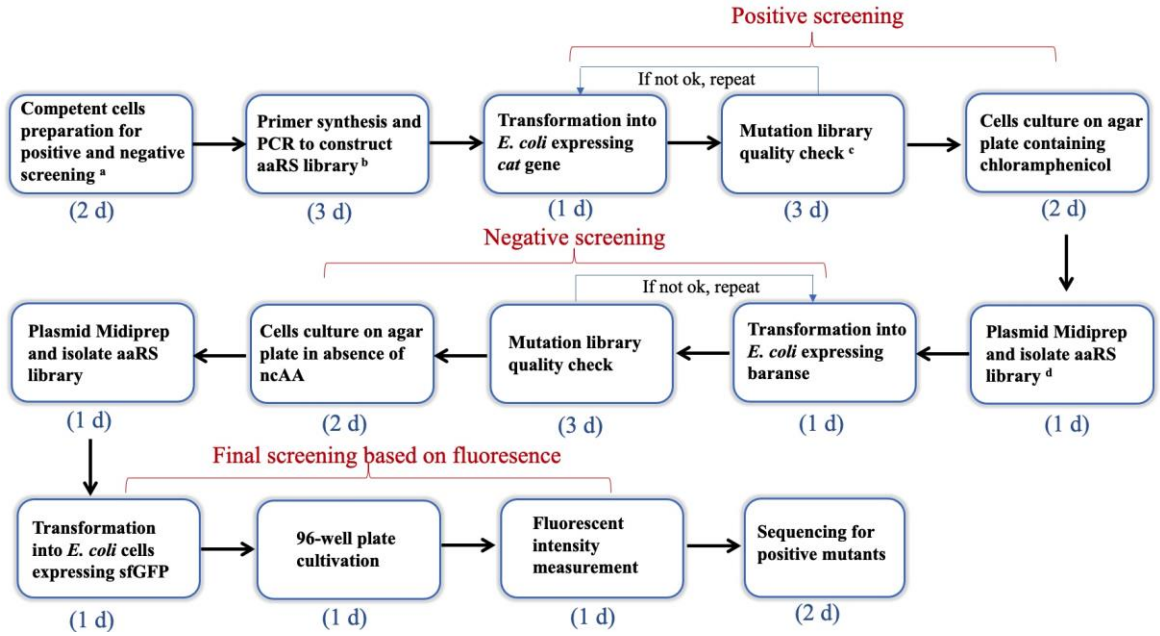

Figure S5 *Mj*TyrRS mutation library positive-negative screening flowchart and time used in each step.

[a] Two electroporation competent cells should be prepared, with one containing plasmid expressing chloramphenicol acetyl transferase (*cat*) with an in-frame amber codon for positive screening. Another one contains plasmid expressing toxin barnase with an in-frame amber codon for negative screening.

[b] Primer synthesis and deliver requires around 2 days, while PCR reaction takes around 1 day. Primer for constructing library contains 4 consecutive NNK at positions H283, P284, M285, and D286.

[c] Mutation library quality was evaluated by counting the number of colonies and sequencing to check the sequence diversity. Since the mutation library was constructed with NNK codon degeneracy (32 codons for 20 aa), the number of colonies should be larger than around 3 times of 32×32×32×32 to cover the whole library.

[d] Plasmid expressing aaRS library was isolated from plasmid expressing cat by purifying the DNA from agarose gel. Failure was often happening at this process as the plasmid having different forms could be mixed and difficult to separate.

4. In line 379-380, the authors posit that some mutations in M-R3 and M-R4 influenced substrate specificity. From the way these sentences are written it is not clear if the authors are referring specifically to mutations in M-R3, M-R4, or to mutations in both. It should be clearly stated which mutations are believed to be impacting specificity.

RESPONSE： Thanks for your comments. Since both of the mutations are located at the tRNA binding domain, they might impact activity through modifying the interactions with tRNA. Hence, both of the mutations might have influenced the

enzyme substrate specificity. We have modified the sentence to make it clear as follows.

"Interestingly, although M-R4 showed higher incorporation efficiency toward *p*AcF compared to M-R3, M-R3 displayed similar activity toward ncAA4, ncAA5, ncAA7, ncAA9 and ncAA10 relative with M-R4, indicating that both mutations might have influenced the enzyme substrate specificity through modifying the interactions with tRNA."

Reviewer #1 (Remarks on code availability):
A README file is provided, and I was able to download and run the code. The code seems appropriately written, however commenting in the provided Python scripts is scarce. Annotation of the code would allow for a better understanding of the steps being performed by the model.
RESPONSE：
Thank you for the comments on our codes. We have followed your suggestions and added more annotations to improve the readability.

Reviewer #2:

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.
RESPONSE： Thank you very much for the kind involvement in the review process.

Reviewer #3:

The manuscript by Zhang and co-workers describes an autonomous system integrating protein language models with automatic biofoundary. The developed system was applied to change the substrate specificity of tRNA synthetase in four rounds. The manuscript is interesting and suitable for Nature Communications. This work represents an important step towards developing fully automated protein engineering systems.

RESPONSE： We thank the reviewer for the positive feedback to our study. We have made general improvements throughout the manuscript following your comments. The questions and suggestions have been addressed point-by-point below. Revisions to the main text are highlighted in red font.

1. 26 Avoid abbreviations in the Abstract.

RESPONSE： Thanks for your suggestions. The abbreviations have been deleted in the Abstract.

2. 35 Add improvement in activity to the Abstract.

RESPONSE： Thanks for your suggestions. We have rewritten the sentence to make it clear.

"The mutants obtained improved enzyme activity by 2.4-fold, leading to a 12.1-fold improvement in production of ncAA-containing protein."

3. 37 Novel proteins … I do not think these are novel proteins

RESPONSE： We have removed the "novel" expression from the sentence.

4. 46 directed evolution is not emerging, it is well-established

RESPONSE： Thanks. We have corrected this sentence as follows.

"Various strategies have been developed for protein engineering, with directed evolution being a well-established and powerful approach."

5. 49 discuss also focused directed evolution

RESPONSE： Thanks for your suggestion. We have added a discussion about focused directed evolution. The added content is as follows:

Recently, targeted mutagenesis guided by structure or sequence information has become a popular way to produce so-called small-but-smart libraries. With an increased ratio of advantageous to deleterious mutations, such libraries increase the efficiency of directed evolution. For example, iterative saturation mutagenesis (ISM) and CASTing methods have been effectively used for enzyme evolution, requiring screening of only moderate-sized libraries including 5,000 to 20,000 variants. However, the effect of targeted mutagenesis is highly dependent on the selected target site, which requires a deep understanding of the relationship between protein structure and function.

6. 51 costly … small-but-smart libraries are preferably used at these days

RESPONSE： Thanks for pointing this out. We have removed the word “costly”, and given an introduction about small-but-smart libraries used for direted evolution.

7. 53 other software for protein design: *https://loschmidt.chemi.muni.cz/portal/*

RESPONSE： Thanks for your suggestion. We have cited the corresponding software suite for protein design in the revised manuscript.

8. 102 please avoid superlative writing … transformative approach

RESPONSE： Thanks. We have revised this sentence as follows.

In this study, we propose a protein engineering strategy that integrates the predictive power of PLMs with the operational efficiencies of an automated biofoundry.

9. 116 within half a month … can you discuss the robustness of your system … how many cycles can be performed in a row?

RESPONSE： Thanks for your suggestion. We have added a discussion as follows.

“We used the *Methanocaldococcus jannaschii p*-cyanophenylalanine tRNA synthetase (pCNF-RS) as a model enzyme for validating the process. In each round, 96 variants were designed by PLMs or a supervised ML model, and then constructed and tested by an automated biofoundry. Four rounds were performed within half a month with the enzyme activity progressively improving, reaching its peak in the fourth round. The PLMeAE system exhibited superior performance compared to random selection and

the traditional directed evolution strategy, and it has the potential to accelerate the engineering of other proteins."

10. 151 please add A, B, C to the figure for clarity

RESPONSE： Thanks. We have added a, b, c in the Figure 2 for clarity.

11. 181 How were 4 residues selected?

RESPONSE： We apologize for the misunderstanding. The sentence has been revised as follows.

"For example, suppose there are four types of amino acids sampled at each masked position, there are a total of 4845 (C(20,4)) subsets (Figure 3a)"

12. 201 provide references

RESPONSE： Thanks. We have added the references.

13. 209 improvement among three cycles is hardly visible

RESPONSE： Thanks for pointing that out. The improvement among the last three rounds was not that significant, although the maximum fitness value was increasing as evolution went, from 5.50 in the second round, to 5.73 in the third, and 6.20 in the fourth round. This was probably due to that the initial supervised ML model predictions had already achieved substantial gains, leaving fewer beneficial mutations available in the remaining sequence space. Since we used an Information Transport Complexity (ITC)-based sampling strategy in the initial round, the data for training the supervised ML model was informative, which leads to the most significant improvement in initial supervised ML model prediction. We have added a sentence to make it clear in the revised manuscript.

"Subsequent rounds also showed incremental improvements: the maximum fitness reached 5.50 in the second round, 5.73 in the third, and peaked at 6.20 in the fourth round, thereby enhancing the original wild-type protein's fitness by 520% (Figure 3d). Despite this, the improvement over the last three rounds was not significant, possibly because the initial supervised ML model predictions had already achieved substantial gains, leaving fewer beneficial mutations available in the remaining sequence space. There are 149 361 data in the GB1 dataset, and the fitness value of 6.20 obtained in fourth round ranked 21st in the dataset. Although the largest fitness value of 8.76 in

the dataset was not obtained yet, this method showed the potential to be used in protein engineering task."

14. 221 VERY IMPORTANT: The data obtained by individual cycles should be reported systematically, e.g., in the Table. Not only maximum values should be reported, but also all mutants with improved activities. This is essential to provide this data.

RESPONSE： Yes, we have provided a separate Table S2 named "GB1 four round PLMeAE data" to report fitness values of all mutants in the four PLMeAE rounds in the supplementary information.

15. 266 standard error 5% is impressive, congratulations

RESPONSE： Thanks. In the experiments, 6 replicates have been carried out for investigating the reproducibility of the automatic platform for measuring enzyme activity.

16. Figure 4 providing timing would make this figure more useful, consider also adding throughput

RESPONSE： Thanks for your suggestions, we added corresponding timing and throughput in figure 4a.

17. 315 it is again important to report if these 7 were activating or neutral, report these data in a Table

RESPONSE： Thanks. We have rewritten the sentences to make it clear. We also provided a Table S3 in supporting information to report all variants activity data from four PLMeAE rounds.

In the first round, the best mutant H283S/P284A/M285A/D286Q (M-R1) showed a 1.3-fold improvement in fitness, while 89 of 96 variants showed lower activity compared to the wild type (Figure 5c & Table S3). The other six mutants include H283T/P284A/M285Q/D286E, which shows a 1.2-fold improvement, and H283K/P284S/M285L/D286A, which shows a 1.1-fold improvement in fitness, along with four variants exhibiting activity similar to that of the wild type.

18. 324 the great effect … superlative writing

RESPONSE： Thanks for pointing this out. We have replaced the word “great” with “positive”.

19. 327 provide sequence similarity/diversity values in a table

RESPONSE： Thanks. The sequence diversity values are provided in Table S4.

Table S4 Entropy data to quantify the diversity of sequence at each site[a]

	Entropy score			
	283	284	285	286
Round 1	1.36295	1.38168	1.38523	1.38569
Round 2	1.88836	1.77672	1.79252	1.28716
Round 3	2.32648	2.21875	2.46325	2.36703

[a]BioEdit was used to calculate the entropy data, which is based on the frequency distribution of different amino acids at each site.

20. 395 the fonts are too small in Figure 6

RESPONSE： Thanks. We have modified the figure.

21. 419 not sure the system is intelligent

RESPONSE： Thanks for pointing this out. We have removed the word “intelligent” in the revised manuscript.

22. 426 report also the total number of mutants

RESPONSE： Thanks for your suggestion. We have added a sentence to report the total number of mutants constructed in the process of evolution, as follows.

“A total of 384 mutants were constructed and tested during the four rounds evolution”.

23. 431 under half a month … can you be more precise?

RESPONSE： Thanks for your suggestion. We have corrected this sentence to make it more precise, as follows.

"In the case of engineering pCNF-RS, a single round of experimental testing takes around 59 h on our biofoundry including around 24 h of primers shipping delay, while ML model training and new variants prediction take less than 1 h. The four design-build-test-learn cycles only took 240 hours, around 10 days."

24. 488 I would appreciate some comments on the generalizability of the system, on its pros and cons, on future development

RESPONSE： Thanks for your suggestions. We have added discussion about the generalizability of the system, on its pros and cons and on future development.

"The PLMeAE platform has the potential to expand applications to engineer enzymes with activity measured by liquid chromatography (LC), gas chromatography (GC) and mass spectrometry (MS). Various automation set up has been developed to accelerate chromatographic and mass spectrometric analysis. For example, a number of liquid chromatography systems are equipped with 96-well plate autosamplers, which are easily operated by robotic arms and scheduling software. Additionally, employing a low-pressure GC configuration combined with low thermal mass (LTM) ovens allows for a cycle time less than 1 min for detecting samples such as fatty acid methyl esters. Various instruments, such as liquid handler robots, microfluidics, and automated solvent extraction equipment, have been used to automate sample preparation for chromatographic analysis. Recent advances also promote the use of high-throughput MS by enabling direct analyte introduction into mass spectrometers using laser, microfluidics, and acoustics. An automated matrix-assisted laser desorption/ionization time of flight (MALDI-ToF) MS has been developed to directedly analyze bacterial and yeast colonies at a rate of ~2 s per sample. A robotic platform was also developed to integrate a liquid handler, wash station, plate dryer, sample stacker, and mass spectrometer, enabling fully automated MALDI MS analysis and allowing for the analysis of over a million samples in a week. Integrating these automated, high-throughput assays into the BioFoundry would significantly expand the application of PLMeAE in engineering other enzymes.

Despite the potential of PLMeAE in protein engineering, developing a new PLMeAE system from scratch is challenging due to the specialized expertise required at the intersection of synthetic biology, computer science, and laboratory automation and robotics. Additionally, establishing new biofoundries is costly. To overcome these challenges, it is crucial to foster collaboration among researchers from diverse disciplines and to train the next generation of synthetic biology researchers with development in artificial intelligence and laboratory automation. In the future, researchers will be able to efficiently perform protein engineering with minimal human intervention."

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Revisions are appropriate and address my earlier comments.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

The authors addressed all my comments thoroughly and properly. I am very pleased with this new manuscript version and support its publishing by Nature Communications. Sincere congratulations!

Response: We thank the reviewers very much for the insightful feedback and constructive comments on our manuscript