Materials and Methods

Materials

Escherichia coli BL21(DE3) (provided by instructor's lab) was used as recipient and pET32a vector containing an ampicillin resistance gene worked as the selection marker.

PolyM were prepared by acid hydrolysis of alginate. Briefly, the mixture containing alginate (2% (w/v)) and 1 M HCl was incubated at 90 ℃ for 6 h. The resulting precipitate was collected and dissolved in the $NaHCO_3$ solution (8% (m/v)). Then, the pH of the mixture was adjusted to 1.0 to precipitate polyM.

For LB liquid medium, 10 g tryptone, 10 g NaCl, 5 g yeast extract were saluted into 1000 mL ddH_2O and then the medium was high-pressure steam sterilized at 121 ℃ for 30 min. LB solid medium: 10 g tryptone, 10 g NaCl, 5 g yeast extract, 15 g agar were saluted into 1000 mL ddH_2O and then the medium was high-pressure steam sterilized at 121 ℃ for 30 min. The tryptone and the yeast extract were purchased from Beijing Aobox Bio-Technology Co. Ltd. The agar was purchased from BBI ®.

Preparation of competent cells

A single colony of *Escherichia coli* BL21(DE3) was picked and inoculated into 5 mL of LB liquid medium, followed by overnight incubation at 37 ℃ with shaking at 150 rpm. An aliquot of 2 mL from the overnight culture was transferred into 200 mL of fresh LB medium and incubated at 37 ℃ with agitation at 180 rpm until the optical density at 600 nm (OD_{600}) reached 0.4~0.5. The culture was immediately chilled on ice for 20 min in 50 mL centrifuge tubes and was centrifuged at 4 ℃, 3,000rpm for 5 min, then the supernatant was discarded. The pellet was resolubilized in 40 mL of cold $CaCl_2$ and centrifuged at 4 ℃, 2,500 rpm for 5 min, then the supernatant was discarded. The pellet was resolubilized in 20 mL of cold $CaCl_2$ and incubated on ice for 30 min. Then it was centrifuged at 4°C, 2,500 rpm for 5 min and the supernatant was discarded. Finally, the pellet resolubilized in 4 mL of cold $CaCl_2$ and aliquoted into EP tubes or PCR tubes to 70 μL per tube, store at -80 ℃.

Extraction of plasmid and vector construction

50 μL of glycerol bacteria (provided by our instructor's lab) was inoculated into 5

mL of LB liquid medium containing a one-thousandth concentration of ampicillin (Sangon biotech) and incubated overnight at 37 ℃ and 180rpm on a constant temperature shaker (MQL-61R, Shanghai Minquan Instrument Co., Ltd). The plasmid was extracted with a plasmid extraction kit, followed by the protocol from the manufacturer (E.Z.N.A. Plasmid DNA Mini Kit I, V spin, Omega Bio-Tek). The concentration and quality of the extracted plasmids were detected by Nano Drop (Thermo Scientific Nanodrop One).

PCR and Transformation

To amplify the site-directed mutants, the primers were about 27 nucleotides in length with mutation site located in the center of the premiers. Degenerate base NNK or MNN were used to generate mutation. All primers used were listed in Table2. Then they were amplified by PCR. The reaction mixture of PCR was 50 μL and was listed in Table3. The PCR program consisted of following parts: pre-denaturation at 95 ℃ for 5 min, 18 cycles of reaction: denaturation at 95 ℃ for 30 s, annealing at (Tm-5) ℃ for 1 min and extension at 72 ℃ for a certain time (1kb for 1 min), finally terminal extension at 72 ℃ for 5 min. The product containing a mutation was obtained and were verified via plasmid sequence performed by Sangon Biotech (Shanghai) Co., Ltd. For each sample, 1 μL of *DPN* I (purchased from Takara) was added and was incubated in a 37°C water bath for 1 hour. Then the mixture with 153 μL ethanol, 17 μL NaAc and the sample was frozen at -20 ℃ for 30 min to precipitate the mutated plasmid. The plasmid was collected by centrifuging at 4 ℃ and 12000 rpm for 10 min.

All plasmid collected was mixed with about 70 μL competent *E. coil* BL21(DE3) and was incubated on ice for 30 minutes. The solution was placed in a 42°C water bath for 60 seconds to open the plasma membrane and then cooled down on ice for 2 min. With the plasmid transferred into the cell, the bacteria were mixed with 500 μL of antibiotic-free LB medium and incubated on a shaker at 150 rpm, 37°C for 40 minutes to 1 hour. All the bacteria were then spread on the LB solid medium to get the single colony. Each of the single colony were picked up to 5 mL LB liquid medium for further cultivation.

Table 2 Site-mutagenic primers used in PCR

PRIMER NAME	SEQUENCE (FROM 5' TO 3')
E144X-F	CACGATAAACGTNNKCAGGGCGGCAAA
E144X-R	TTTGCCGCCCTGMNNACGTTTATCGTG
R159X-F	CTGGCGACCGTTNNKCTGAACAACAAA
R159X-R	TTTGTTGTTCAGMNNAACGGTCGCCAG

PRIMER NAME	SEQUENCE (FROM 5' TO 3')
F170K172-F	GCGGGCCGTNNKAAANNKATCGTTCTG
F170K172-R	CAGAACGATMNNTTTMNNACGGCCCGC
K172E-F	GGCCGTTTCAAAGAAATCGTTCTGGGT
K172E-R	ACCCAGAACGATTTCTTTGAAACGGCC
K172F-F	GGCCGTTTCAAATTTATCGTTCTGGGT
K172F-R	ACCCAGAACGATAAATTTGAAACGGCC
K172MV-F	GGCCGTTTCAAARTGATCGTTCTGGGT
K172MV-R	ACCCAGAACGATCAYTTTGAAACGGCC
K172RI-F	GGCCGTTTCAAAAKAATCGTTCTGGGT
K172RI-R	ACCCAGAACGATTMTTTTGAAACGGCC
K172HL-F	GGCCGTTTCAAACWTATCGTTCTGGGT
K172HL-R	ACCCAGAACGATAWGTTTGAAACGGCC
K172N-F	GGCCGTTTCAAAAACATCGTTCTGGGT
K172N-R	ACCCAGAACGATGTTTTTGAAACGGCC
R143X-F	AGCTGGCACGATAAANNKGAACAGGGCGGCAAA
R143X-R	TTTGCCGCCCTGTTCMNNTTTATCGTGCCAGCT
R143HQNK-F	TGGCACGATAAAMANGAACAGGGCGGC
R143HQNK-R	GCCGCCCTGTTCNTKTTTATCGTGCCA
R143IT-F	TGGCACGATAAAAYHGAACAGGGCGGC
R143IT-R	GCCGCCCTGTTCDRTTTTATCGTGCCA
R143EFV-F	TGGCACGATAAAKWYGAACAGGGCGGC
R143EFV-R	GCCGCCCTGTTCRWMTTTATCGTGCCA
R143C-F	TGGCACGATAAATGTGAACAGGGCGGC
R143C-R	GCCGCCCTGTTCACATTTATCGTGCCA

Table 3 PCR Reaction Mixture

Reagent Name	Dosage(μL)
PrimeSTAR HS DNA Polymerase	0.5
Primers (containing forward primer and reverse primer)	1.3*2
dNTP	4
5x PS Buffer	10
Plasmid	50 ng/μL
ddH_2O	Make up to 50μL

Induction and Sonication Lysis

The *E. coil* containing expression vector were incubated in 5 ml LB liquid medium containing a one-thousandth concentration of ampicillin on a shaker at 37°C with 220 rpm for 5 hours to get seed liquid. The entire seed liquid was then transferred into 300 mL LB liquid medium containing ampicillin at the same concentration. The culture was incubated at 37°C with shaking at 180 rpm until OD_{600} reaches 0.6. The cells were

induced by 120 μL isopropyl-β-D-thiogalactopyranoside (IPTG, the concentration of IPTG should be 0.2 mM) to induce expression at 16 ℃ with shaking at 180 rpm for 18 hours. Induced cells were harvested by centrifugation at 4 ℃ 3000 rpm for 30 min and 4 ℃ 8000 rpm for 15 min for removing the medium left. The sediment was suspended in Tris-HCl buffer (20 mM, pH 8.0) containing 200 mM NaCl and lysed by sonication. The sonication program was set as 2 s sonication and stopped for 4 s, lasting for 15 min. The product was centrifugated at 4 ℃ 12000 rpm for 30 min. The supernatant was collected as the unpurified enzymes.

SDS-PAGE

The expression level of the enzyme was detected by SDS-PAGE. The supernatant was mixed with a quarter of Loading Buffer in volume and then heated at 95 ℃ for 10 min. The protein with the buffer was electrophoresed and colored.

TLC Analysis

The reaction system and duration were designed based on SDS-PAGE results to allow the enzyme fully catalyzing the substrate polyM sufficiently at room temperature. The reaction was stopped by 100°C metal bath for 10 minutes, and then we centrifuged the sample at 12,000 rpm for 10 minutes to obtain supernatant, which is the reaction products. Primary assessment of enzyme activity, product distribution and relative content was performed by thin-layer chromatography (TLC). The reaction solution was putted on the TLC board (purchased from Merck KGaA) then developed in acidic developing agent (n-butanol: formic acid: water = 4:6:1, in volume ratio). The result was showed by color developer (2g diphenylamine, 2 mL aniline, 100mL acetone, 10 mL phosphoric acid and 1 mL concentrated hydrochloric acid).

HPLC Analysis

Quantitative determination of product distribution and absolute content will be performed by high-performance liquid chromatography (HPLC).

Samples were prepared by heating at 100 °C for 10 min followed by centrifugation at 12,000 rpm for 10 min. The supernatant was collected and directly subjected to HPLC analysis without dilution. A 40 μL aliquot of each sample was injected for analysis. Due to slight inconsistencies in sample loading during TLC preparation, we increased the injection volume for certain samples; the exact values are provided in the raw data available Supplementary Materials 3.

HPLC analysis was carried out on a Thermo Fisher Ultimate 3000 UHPLC system equipped with a UV detector set at 230 nm (channel 1). Separation was achieved on a YMC-Pack ODS-A C18 column (150 mm × 4.6 mm i.d., 5 μm, 12 nm pore size) maintained at 35 °C.

The mobile phases consisted of solvent B (acetonitrile) and solvent C (5 mM tetrabutylammonium bromide, aqueous). The flow rate was set to 1.0 mL/min, and the following gradient program was applied:

1. 0–5 min: 20% B, 80% C
2. 5–18 min: 20% B, 80% C
3. 18–25 min: 55% B, 45% C
4. 25–30 min: 90% B, 10% C
5. 30–35 min: 90% B, 10% C
6. 35–42 min: 20% B, 80% C

Peak separation and integration were performed to determine the product distribution.

Data Collection and Preprocessing

In this study, single-site and double-site mutation data were obtained from experimental results and stored in the files 172_and_159_single_point_mutation_data.xlsx and 159-172_double_point_combination_mutation_data.xlsx. Each file contains mutation names, mutation ratios (Ratio), and peak areas (Area). The data were first cleaned and preprocessed according to the following steps:

Single-Site Mutation Data: Data from the 172_and_159_single_point_mutation_data.xlsx file were used to extract mutation names, ratios, and peak areas. For each mutation, the ratio and peak area were compared with the wild-type mutation to calculate the effect size.

Double-Site Mutation Data: Data from the 159-172_double_point_combination_mutation_data.xlsx file were used to extract ratio and peak area values for the double-site mutation combinations. These data were used for synergy effect analysis in conjunction with the single-site mutations.

Feature Engineering and Synergy Modeling

To analyze the synergy between single and double-point mutations, we constructed a feature space based on the physical-chemical properties and structural information of the mutations. The feature engineering process included the following steps:

Single-point Mutation Effect Extraction: The effect of each single-point mutation relative to the wild type was calculated for both the ratio and peak area. These effects were used as baselines for subsequent double-point mutation effect predictions.

Double-point Mutation Synergy Modeling: Based on the effects of single-point mutations, the expected effect of double-point mutations was calculated by summing the individual effects of each single-point mutation. The actual effect of the double-point mutation was compared with the expected effect to calculate the synergy, which was defined as the difference between the actual and expected effects. The synergy effect formula is as follows:

$$Synergy\ Effect = Actual\ Effect - Expected\ Effect$$

Advanced Feature Construction: For each mutation (single or double), a 28-dimensional feature vector was extracted based on the physical-chemical properties of amino acids (e.g., hydrophobicity, volume, polarity). This feature vector included:

Changes in physical-chemical properties (e.g., hydrophobicity, volume, polarity)

due to amino acid substitutions.

Interaction features between mutation sites (e.g., product terms, difference terms).

Multi-dimensional combination features to capture interactions between different mutations.

Model Development and Evaluation

To predict mutation effects and synergy effects, multiple non-linear regression models were developed and evaluated:

Model Selection:

Support Vector Regression (SVR): An SVR model using an RBF kernel was employed to capture non-linear relationships in high-dimensional feature spaces.

Multi-layer Perceptron (MLP): A neural network-based regression model using ReLU activation was used to model complex non-linear data.

Gaussian Process Regression (GPR): A GPR model using RBF and white noise kernels was employed to estimate uncertainty in predictions.

Kernel Ridge Regression: A combination of ridge regression and kernel methods, suitable for handling high-dimensional non-linear relationships.

Data Standardization: Prior to training, all input features were standardized, ensuring that each feature had a mean of 0 and a variance of 1. This step prevented differences in feature scales from affecting model training.

Cross-validation: All models were evaluated using 5-fold cross-validation, with performance metrics including R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The model with the best performance was selected for final predictions.

Model Selection: The best model was chosen based on the highest performance in cross-validation, ensuring the model's robustness and reliability.

Mutation Pattern Recognition and Clustering

To further understand the underlying patterns of mutations, Principal Component Analysis (PCA) and K-means clustering were used: PCA for Dimensionality Reduction: The 28-dimensional feature space was reduced to 2 dimensions using PCA, making it easier to visualize and analyze mutation patterns. The variance explained by each principal component was calculated to assess the amount of information retained during dimensionality reduction. K-means Clustering: After PCA, the K-means clustering algorithm was used to group mutations into three clusters based on their similarities in

ratios and peak areas. Each cluster represented a different mutation pattern, and mutations within the same cluster shared similar characteristics.

Optimal Mutation Combination Prediction

Using non-linear regression models and synergy analysis, the optimal double-site mutation combinations were predicted. The prediction process included the following steps:

Mutation Space Search: All possible double-site mutation combinations were enumerated, and a feature vector was calculated for each combination.

Model Prediction: The trained models were used to predict the ratio and peak area for each mutation combination.

Composite Scoring: The predicted ratios and peak areas were normalized, and a weighted composite score (e.g., 0.6 for ratio and 0.4 for peak area) was computed for each mutation combination.

Optimal Combination Selection: The top 10 mutation combinations with the highest composite scores were selected, and their synergy effects and physical-chemical properties were analyzed.

Results Reporting and Visualization

Finally, a detailed results report was generated, including:

Best Model Selection: The best model for ratio and peak area predictions was selected, and its performance metrics were reported.

Clustering Analysis: The results of PCA and K-means clustering were presented, highlighting different mutation patterns.

Optimal Mutation Combination Recommendation: The top 10 optimal mutation combinations were recommended for experimental validation.

Additionally, the results were exported to an Excel file for further analysis and experimental verification.

Data analysis and availability

HPLC data processing

HPLC data processing and peak fitting were carried out using Origin Pro 2024b V2 (Origin Lab). The baseline was manually defined, and peak detection was performed using the second-derivative method. Integration was conducted within the retention time range of 15–22 min (±1 min) to calculate the total peak area and the proportion of DP3–DP6 products. The original data are available at https://doi.org/10.5281/zenodo.17238377.

Data organization and plotting were performed in R version 4.4.3 using the tidyverse and ggplot2 packages.

Dry experiment project

The project code is available on GitHub at：https://github.com/sunanqispecial/idec