Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sequence Redesign of Glycosyltransferases for Enhanced Heterologous Expression and Glycosylation Efficiency in Escherichia coli

Yifei Zhang

yifeizhang@mail.buct.edu.cn

Beijing University of Chemical Technology https://orcid.org/0000-0002-0014-611X

Shanshan Zhang

Beijing University of Chemical Technology

Yilei Han

Tsinghua University

Yujing Ding

Beijing University of Chemical Technology

Xiaofan Lin

Beijing University of Chemical Technology

Zhipeng Yuan

Beijing University of Chemical Technology

Xinxiao Sun

Beijing University of Chemical Technology https://orcid.org/0000-0001-5590-6473

Qipeng Yuan

Beijing University of Chemical Technology https://orcid.org/0000-0002-7899-1487

Zheng Liu

Tsinghua University

Article

Keywords:

Posted Date: March 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8083040/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Sequence Redesign of Glycosyltransferases for Enhanced Heterologous Expression and Glycosylation Efficiency in *Escherichia coli*

Shanshan Zhang[1,2], Yilei Han[3*], Yujing Ding[1,2], Xiaofan Lin[1,2], Zhipeng Yuan[1,2], Xinxiao Sun[1,4*], Qipeng Yuan[1,4], Zheng Liu[3], Yifei Zhang[1,2*]

1. State Key Laboratory of Chemical Resources Engineering, Beijing University of Chemical Technology, Beijing 100029, China
2. Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China
3. Department of Chemical Engineering, Tsinghua University, Beijing 100084, China
4. College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China.

[*]Corresponding authors:

Yilei Han: hanyl@mail.tsinghua.edu.cn;

Xinxiao Sun: sunxx@mail.buct.edu.cn;

Yifei Zhang: yifeizhang@mail.buct.edu.cn

Abstract

Heterologous expression of enzymes from higher organisms remains a major bottleneck in the construction of microbial cell factory. To address the problem, we developed a sequence redesign strategy guided by essential structural features and evolutionary conservation. This strategy integrates ProteinMPNN with several scorers to comprehensively redesign enzymes that inherently exhibit poor heterologous expression. We applied this approach to two glycosyltransferases, TOGT and UGT84A56, for esculetin glycosylation. The redesigned enzymes featured alterations in more than 20% of their sequences, resulting in significant improvements in soluble expression in *Escherichia coli*. Furthermore, both *in vitro* and *in vivo* assays demonstrated notable enhancements in the catalytic activity of the redesigned enzymes. Specifically, the best-performing TOGT variant exhibited a 6.4-fold increase in soluble expression and an 11.8-fold improvement in catalytic efficiency, enabling the production of cichoriin at a titer of 299 mg/L in whole-cell catalysis. Similarly, the top-performing UGT84A56 variant achieved 402 mg/L aesculin, representing a 13.9-fold improvement over the native enzyme. Fed-batch fermentation using these redesigned enzymes achieved record-high product titers of 2.11 g/L for cichoriin and 4.05 g/L for aesculin, respectively. This study demonstrates a powerful strategy for enzyme redesign, paving the way for efficient microbial cell factories to synthesize high-value natural products.

Main

The construction of microbial cell factories represents a transformative approach for the sustainable production of high-value natural products, providing a solution to the growing demand for environmentally friendly and cost-effective chemical synthesis[1,2]. Central to this strategy is the integration and expression of heterologous genes encoding foreign enzymes within host cells[3,4]. For example, the reconstitution of plant biosynthetic pathways has enabled the synthesis of alkaloids[5,6], flavonoids[7,8], and terpenoids[9] in *E. coli*, offering an alternative to traditional plant-based extraction methods. However, the expression of heterologous enzymes is often hindered by factors such as codon usage bias, protein misfolding and aggregation, host cell toxicity, and the lack of necessary post-translational modification systems[10,11,12]. Conventional strategies to enhance enzyme expression include optimizing expression conditions, adjusting promoter strength, changing expression vectors and/or hosts, optimizing codon, co-expressing molecular chaperones, and fusing with solubility tags[13,14,15,16]. However, the effectiveness of these approaches is often protein-dependent and, in many cases, compromises enzyme activity[17,18,19,20]. Thus, alternative strategies are needed to reliably improve the expression of heterologous enzymes.

Cichoriin (7-O-β-$_D$-glucoside of esculetin) and aesculin (6-O-β-$_D$-glucoside of esculetin) are both glycosylated derivatives of esculetin (6,7-dihydroxycoumarin), a bioactive coumarin compound found in Fraxini cortex, a traditional oriental medicine used for over 1000 years. Recent investigations have shown that esculetin can reduce inflammation by inhibiting the NF-κB (nuclear factor-kappa B) and MAPK signaling pathways, suppress tumor growth by regulating the cell cycle, and exert strong antioxidant effects by scavenging reactive oxygen species (ROS)[21,22]. However, its low solubility and bioavailability hinder its clinical development[23]. In contrast, cichoriin and aesculin (commonly extracted from *Cichorium intybus*[24] and *Fraxinus chinensis*[25], respectively) possess therapeutic activities comparable to esculetin, but with enhanced solubility and stability due to glycosylation (Fig. 1a), making them promising alternatives with improved pharmacokinetic properties[26,27]. The regioselective glycosylation of esculetin is mediated by distinct enzymes (Fig. 1b). Sun *et al.* identified enzymes involved in aesculin biosynthesis in horse chestnut (*Aesculus chinensis*) and

demonstrated its *de novo* production from glucose in *E. coli*, in which a glycosyltransferase, UGT84A56, was employed to specifically catalyze the 6-O-glucosylation of esculetin[28]. Subsequently, Xie *et al.* enhanced esculetin supply by engineering cinnamyl-CoA ortho-hydroxylase from *Arabidopsis thaliana* (*At*F6'H) and utilized *Nicotiana tabacum*-derived TOGT for its 7-O-glycosylation, thereby producing cichoriin[29]. Using this strategy, they achieved titers of 73.9 mg/L for cichoriin and 254.8 mg/L for aesculin. Notably, the accumulation of residual esculetin suggests that glycosylation remains limited by the activity of glycosyltransferases. Therefore, enzyme engineering is required to enhance the glycosylation efficiency.

Recent advances in machine learning have dramatically revolutionized protein engineering, moving beyond limited point mutations toward the comprehensive redesign of enzyme scaffolds and sequences[30,31]. Notably, deep learning models have shown success in generating candidate amino acid sequences tailored to given protein backbones or desired functions[32,33]. For example, ProteinMPNN-based protocols have been successfully used to redesign myoglobin and tobacco etch virus protease, resulting in enhanced thermostability and activity[34]. However, ProteinMPNN tends to generate protein designs with greater structural rigidity, where gains in folding and stability often come at the expense of catalytic activity[35,36,37]. Achieving the right balance among structural stability, expressibility, and enzymatic activity requires careful identification and preservation of functionally and evolutionarily important residues from the original sequence in the redesign process.

In this study, we present a structure- and evolution-guided design strategy, leveraging ProteinMPNN to redesign the sequences of TOGT and UGT84A56, to improve their heterologous expression in *E. coli* for efficient esculetin glycosylation. Ten experimental designs and a total of twelve control designs were proposed for each enzyme and evaluated for their expression level, activity, and product titers. Whole-cell biotransformation experiments confirmed enhanced product titers in all 10 designs (100%) of TOGT and 2 designs (20%) of UGT84A56. For TOGT, the top-performing design, $sdT_{exp}5$, achieved a remarkable 6.4-fold increase in soluble expression and an 11.8-fold boost in catalytic efficiency, yielding a cichoriin titer of 299 mg/L under feeding experiments with *E. coli* BL21(DE3)—33.2 times higher than

that of the native TOGT. Likewise, the best design of UGT84A56, sdU$_{exp}$6, exhibited a 2.2-fold increase in soluble expression and a 10.8-fold improvement in catalytic efficiency, resulting in an aesculin titer of 402 mg/L, 13.9 times greater than that of the native UGT84A56. Comparative experiments demonstrated that this approach outperformed other commonly used solubility-enhancement strategies, such as molecular chaperone co-expression and fusion to solubility tags. Molecular dynamics simulations revealed that the redesigned enzymes adopted more compact structures, improved conformational stability near the active site, and increased surface hydrophilicity, providing mechanistic insights into the observed performance enhancements. Application of these designs in engineered *E. coli* BW1 strains for fed-batch fermentation yielded record-high titers of 2110 mg/L cichoriin and 4050 mg/L aesculin. These findings highlight the potential of comprehensive sequence engineering to enhance heterologous protein expression, paving the way for the development of highly productive microbial cell factories for the synthesis of complex natural products.

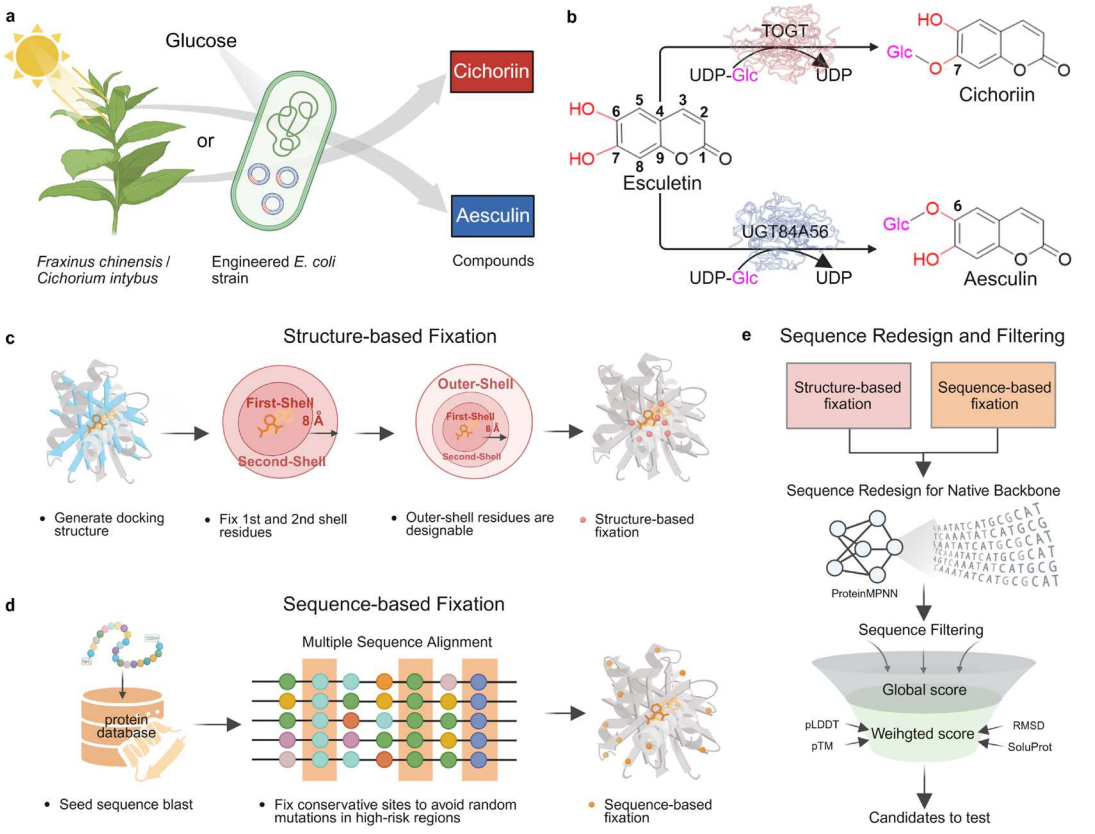


Fig. 1. Design strategy for optimizing protein expression and activity using ProteinMPNN. a, Schematic representation of natural and engineered biosynthetic approaches for producing cichoriin and

aesculin. **b,** Enzymatic regioselective glycosylation of esculetin by TOGT and UGT84A56 to synthesize cichoriin and aesculin, respectively. **c-d,** Schematics of structure-based and sequence-based fixation strategies. Fixed sites include residues within the first-shell and second-shell (8 Å radius) of the sugar donor and acceptor, as well as conserved residues. **e,** Workflow of the enzyme redesign. Step 1: Structure-based and sequence-based fixation analysis to prepare input parameters for ProteinMPNN; Step 2: ProteinMPNN-based generation of new enzyme sequences based on the native backbone and the fixed sites; Step 3: Sequence filtering through a two-tier evaluation based on global score and weighted score, resulting in candidates for experimental testing. UDP, uridine diphosphate; UDP-Glc, uridine diphosphate glucose; Glc, glucose; pLDDT, predicted local distance difference test; pTM, predicted template modeling score; RMSD, root mean square deviation; SoluProt, solubility prediction.

Results

Redesign the sequences of glycosyltransferases with ProteinMPNN

We developed a strategy to identify essential residues to fix during sequence redesign (Fig. 1c and 1d), aiming to enhance the soluble expression of glycosyltransferases while preserving their catalytic activity. Two sets of residues were fixed: (1) residues within 8 Å of the substrate binding pocket as suggested by molecular docking, containing the first-shell and second-shell residues to preserve enzyme-substrate interactions and active-site geometry; and (2) evolutionarily conserved residues identified through multiple sequence alignment (MSA) using ConSurf[38], to minimize disruption of structural and functional integrity. The remaining residues were alterable with inverse folding algorithms and were designated as designable sites. In this way, the redesign primarily focused on solvent-exposed regions that were spatially distant from the active site and could potentially affect solubility. The fixed residues and the native protein backbone were input into ProteinMPNN to generate new enzyme sequences, which were further screened to obtain new enzyme designs for experimental validation (Fig. 1e). During the sequence generation stage, a large number of sequences, each assigned a global score by ProteinMPNN were produced. Among these, the top sequences with the lowest global scores[39], indicative of higher overall sequence quality, were selected for further evaluation. The structures of these designs were predicted using AlphaFold2[40] and subsequently assessed through a series of structure-based metrics, including backbone RMSD, predicted solubility

(assessed using SoluProt[41]), pLDDT and pTM values. These metrics were normalized and then aggregated using specified weighting factors to calculate a weighted score. After re-ranking these sequence designs, the top designs with the highest weighted scores were selected for experimental validation. Using this workflow, we redesigned the native sequences of TOGT and UGT84A56 to improve their solubility and further enhance their activity.

Redesign of TOGT for enhanced soluble expression and activity in *E. coli*

The enzyme redesign strategy was firstly applied to the glycosyltransferase TOGT (Fig. 2a). Molecular docking identified 14 residues in the first shell, 69 in the second shell, and 393 in the outer shell. Conservation analysis further revealed that approximately half of the residues in the outer shell (190 sites) were designable, accounting for roughly 40% of the full-length sequence (Fig. S1). After determining the residues to fix, a two-tiered filtering pipeline was applied to select the top 10 sequence designs with the highest weighted scores from 20 candidates (Table S1). These candidates were preselected from 100 ProteinMPNN-generated sequences with the lowest global scores. At each step of the design and evaluation process, control groups were established to systematically assess the contributions of shell fixation and filtering criteria to sequence designs.

The yielded TOGT designs and the native enzyme were expressed using *E. coli* BL21(DE3) cells, and the soluble protein concentration and specific activities of the cell lysate were measured, as shown in Fig. 2b. All of the 10 designs exhibited markedly improved soluble expression and specific activity compared to the native TOGT. SDS-PAGE analysis confirmed that all enzymes generated by ProteinMPNN were expressed predominantly as soluble proteins, whereas the native TOGT was primarily formed insoluble inclusion bodies (Figs. S2 and S3). Kinetic characterization was further conducted with the purified enzymes (Table 1, Figs. S4 and S5). Most designs maintained catalytic activity comparable to the native enzyme, with certain designs, such as $sdT_{exp}8$, displaying a substantially higher turnover number (k_{cat}). Additionally, all designs showed reduced K_m values, suggesting improved substrate affinity. As a result, the catalytic efficiencies (k_{cat}/K_m) of the designed enzymes were enhanced by 1.8- to 11.8-fold relative to native TOGT. The melting temperatures (T_m) of these designs increased significantly

(by 9.4 to 23 °C) compared to native TOGT (Fig. S6), showing the inherent bias of ProteinMPNN-guided design toward enhanced thermostability. Ablation experiments revealed that while ProteinMPNN alone improved soluble expression, the effect was further amplified when more functional and conserved residues were fixed (Figs. 2b and S7a). Furthermore, the specific activity progressively improved with the introduction of global score evaluation and weighted score evaluation, highlighting the critical contributions of each component in the design process (Figs. 2b and S7b).

Whole-cell fermentation experiments were conducted using the designed enzymes with esculetin feeding, and the cichoriin titer were measured as an indicator of the *in vivo* activity of the enzyme designs. As shown in Fig. 2c, the in vivo activity showed a strong correlation with the total enzyme activity of the cell lysate measured *in vitro*. As expected, the designs generated by the workflow exhibited the highest catalytic performance compared to all control sets (Fig. S8). *E. coli* BL21(DE3) strains expressing the designs produced 236–299 mg/L cichoriin, corresponding to conversion rates of 51–65%, representing a 26.2- to 33.2-fold improvement compared to the native counterpart. The significantly enhanced expression levels, combined with the improved *in vitro* and *in vivo* activities, collectively demonstrated the effectiveness and robustness of our enzyme design strategy.

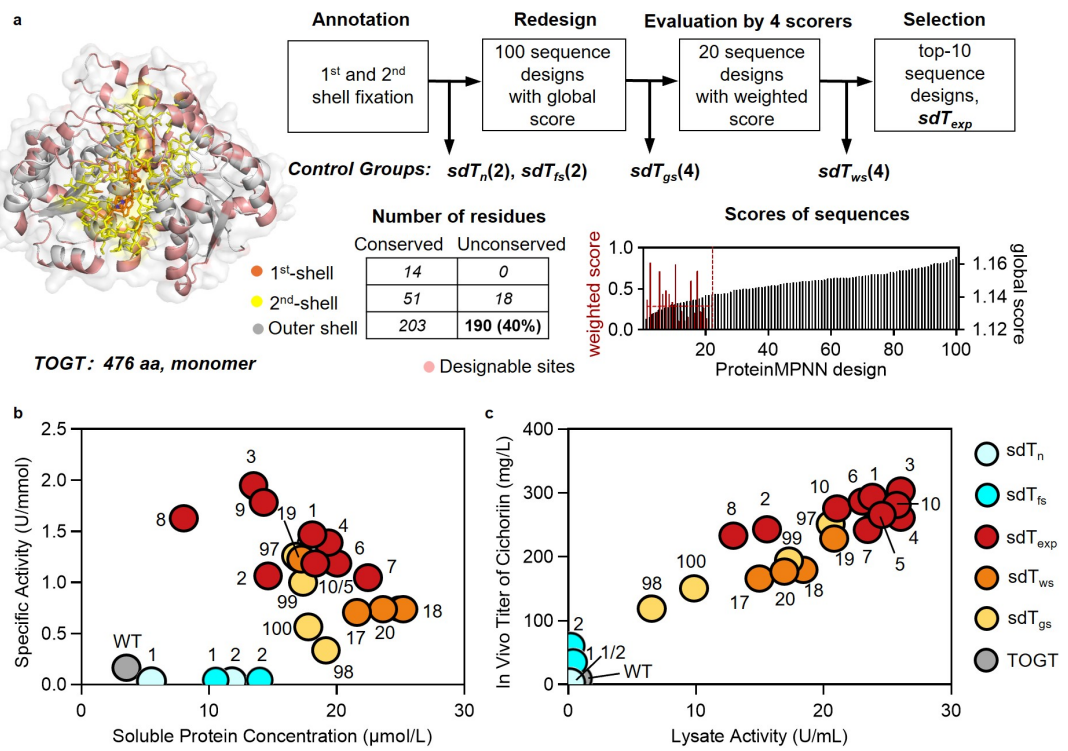


Fig. 2. Redesign of TOGT and characterizations of the designs. a, Workflow for TOGT redesign, including annotation of the fixed/designable sites, sequence redesign, sequence filtering, and experimental validation. The structure of TOGT is shown, with the first-shell sites colored in orange, the second-shell sites in yellow, the outer-shell sites in grey, and designable sites in pink. The confusion matrix categorizes and quantifies the number of residues based on structure-based and sequence-based fixation analysis. The bar plot illustrates the global scores of 100 redesigns generated by ProteinMPNN (in black) and the weighted scores of the top 20 redesigns (in dark red). The vertical and horizontal dashed lines mark the thresholds for global score and weighted score, respectively. **b,** Soluble protein expression levels and specific activity of TOGT (WT) and its designs expressed in *E. coli* BL21(DE3). The designs, ranked by weighted score from highest to lowest, were labeled as sdT_{exp}1-10. Four additional groups were designed and tested as control: two designs with no residue fixation (sdT_n1-2), two designs fixing only the first shell residues (sdT_{fs}1-2), and two control groups based on the global scores (sdT_{gs}97-100, the last four of the ProteinMPNN-generated 100 sequences) and weighted scores (sdT_{ws}17-20, the last four of the top-20 sequences). **c,** Correlation between the cell lysate activity and whole-cell transformation of esculetin to cichoriin.

Table 1. Kinetic parameters for TOGT and the sdT_{exp} designs

Designs	K_m (μM)	k_{cat} (min^{-1})	k_{cat}/K_m (μM^{-1} min^{-1})	fold improvement
TOGT	93.6±19.8	4.2±0.5	0.04	N.A.
$sdT_{exp}1$	10.1±1.4	3.6±0.1	0.36	9.0
$sdT_{exp}2$	16.5±3.3	1.8±0.1	0.11	2.8
$sdT_{exp}3$	6.4±1.0	3.0±0.1	0.47	11.8
$sdT_{exp}4$	38.8±12.1	4.7±0.2	0.12	3.0
$sdT_{exp}5$	76.0±14.5	5.5±0.2	0.07	1.8
$sdT_{exp}6$	21.1±2.2	4.1±0.1	0.19	4.8
$sdT_{exp}7$	73.4±18.7	5.8±0.3	0.08	2.0
$sdT_{exp}8$	82.9±23.9	18.5±0.1	0.22	5.5
$sdT_{exp}9$	21.7±5.6	4.6±0.2	0.21	5.3
$sdT_{exp}10$	68.1±8.5	4.8±0.1	0.07	1.8

Redesign of UGT84A56 for enhanced soluble expression and activity in *E. coli*

The design strategy was further extended to UGT84A56 from *Aesculus chinensis*, a glycosyltransferase with 6-O-glycosylation selectivity that also suffers from poor soluble expression in *E. coli*. Among the 472 residues in UGT84A56, 212 sites were identified as designable, accounting for 45% of the full-length sequence (Figs. 3a and S9). Like TOGT, the first and second shells of UGT84A56 also contained several non-conserved residues that required fixation during the design process, highlighting the complementary of sequence- and structure-based strategies.

Experimental validation of 10 selected designs from the workflow showed that 8 designs achieved improved expression levels, while 3 designs, $sdU_{exp}4$, $sdU_{exp}6$, and $sdU_{exp}8$, exhibited specific activities that were higher or comparable to the native UGT84A56 (Figs. 3b and S10-11). Enzymatic kinetic analysis of the purified enzymes revealed that the enhanced activity was attributed to both reduced K_m and increased k_{cat} values (Table 2, Fig. S12). Among these, the top-performing design, $sdU_{exp}6$, demonstrated a catalytic efficiency (k_{cat}/K_m) that was 10.8-fold greater than the native enzyme. Additionally, these designs also exhibited enhanced

thermostability (Fig. S13). Ablation experiments further indicated that most sequences generated by ProteinMPNN exhibited increased protein expression, although the magnitude of improvement was less pronounced than that observed for TOGT (Figs. 3b, S10 and S14). Two designs from the experimental group, $sdU_{exp}2$ and $sdU_{exp}9$, failed to achieve soluble expression unexpectedly. Regarding specific activity, two control designs, $sdU_{ws}20$ and $sdU_{gs}98$, exhibited levels comparable to the native enzyme (Fig. S15).

In vivo whole-cell biotransformation experiments were conducted using the high-activity designed enzymes alongside the control groups (Figs. 3c and S16). For most cases, the *in vivo* productions of aesculin were consistent with the *in vitro* catalytic activity observed in cell lysates, except for native UGT84A56 and $sdU_{exp}8$, presumably due to their low cellular stability. Notably, $sdU_{exp}4$ and $sdU_{exp}6$ achieved the highest aesculin titers of 313 mg/L and 402 mg/L, respectively, representing a 10.8- and 13.9-fold increase compared to the native enzyme (Fig. S16). In contrast, designs generated without fixing the first- and second-shell residues showed significantly reduced activity. These findings mirrored the significant increases in product titers observed in TOGT designs and validated again the effectiveness of our enzyme redesign strategy.

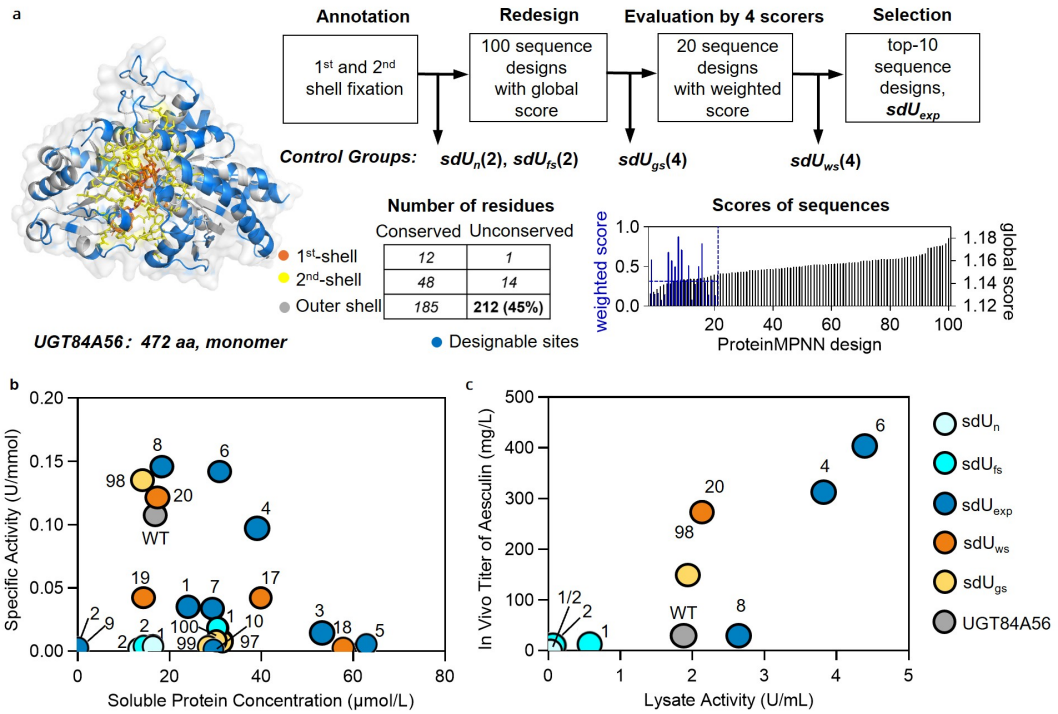


Fig. 3. Redesign of UGT84A56 and characterizations of the designs. a, Workflow for UGT84A56 redesign, including annotation of the fixed/designable sites, sequence redesign, sequence filtering, and experimental validation. The structure of UGT84A56 is shown, with the first-shell sites colored in orange, the second-shell sites in yellow, the outer-shell sites in grey, and designable sites in blue. The confusion matrix categorizes and quantifies the number of residues based on structure-based and sequence-based fixation analysis. The bar plot illustrates the global scores of 100 redesigns generated by ProteinMPNN and the weighted scores of the top 20 redesigns. The vertical and horizontal dashed lines mark the thresholds for global score and weighted score, respectively. **b,** Soluble protein expression levels and specific activity of UGT84A56 (WT) and its designs expressed in *E. coli* BL21(DE3). The designs, ranked by weighted score from highest to lowest, are labeled as sdU_{exp}1-10. Four additional groups were designed and tested as control: two designs with no residue fixation (sdU_n1-2), two designs fixing only the first shell residues (sdU_{fs}1-2), and two control groups based on the global scores (sdU_{gs}97-100, the last four of the ProteinMPNN-generated 100 sequences) and weighted scores (sdU_{ws}17-20, the last four of the top-20 sequences). **c,** Correlation between the cell lysate activity and whole-cell transformation of esculetin to aesculin.

Table 2. Kinetic parameters for UGT84A56 and the sdU_{exp} designs

Designs	K_m (μM)	k_{cat} (min^{-1})	k_{cat}/K_m (μM^{-1} min^{-1})	fold improvement
UGT84A56	203.2±29.5	0.33±0.03	0.0016	N.A.
$sdU_{exp}4$	59.4±10.6	0.38±0.02	0.0064	4.0
$sdU_{exp}6$	72.8±13.6	1.25±0.07	0.0172	10.8
$sdU_{exp}8$	163.2±23.5	0.38±0.02	0.0023	1.4

Comparison with conventional expression-enhancing strategies

The ProteinMPNN-redesign-based solubility-enhancement strategy was compared to commonly used approaches, including co-expression with molecular chaperones and fusion to solubility tags[42], by assessing product titers in whole-cell biotransformation experiments (Fig. 4). These experiments were conducted in an alternative *E. coli* chassis strain, BW25113 (Fig. S17).

For the TOGT enzyme, all 10 designs exhibited higher cichoriin production and enhanced enzyme expression, as indicated by increased OD_{600} values (Fig. 4a). In contrast, designs generated without residue fixation or with only first-shell residues fixed achieved higher enzyme expression but showed limited product titers. Among the four tested chaperones[15,43]—ibpAB, RpoH (σ^{32}), DnaKJ, GroSL and—only GroSL co-expression led to an increase in yield. However, chaperone-based approaches constrained cell growth, as evidenced by reduced OD_{600} values, possibly due to the metabolic burdens they imposed on cells. Four solubility tags, including CysQ, GST, SUMO and MBP, were tested[20,44]. Among them, SUMO or MBP tags led to modest improvements in soluble expression and cichoriin titers. However, their effects remained substantially lower than those achieved with enzyme redesigns, possibly due to the introduction of solubility tags impairing enzymatic activity or disrupting proper structural folding[17,45] (Fig. S18).

For UGT84A56, comparisons among different strategies revealed a similar pattern (Fig. 4b). Among the three enzymes that demonstrated superior *in vitro* activities, $sdU_{exp}4$ and $sdU_{exp}6$ again exhibited enhanced expression and catalytic activity, whereas the designs lacking full residue fixation showed even lower yields compared to the native enzyme. The utilization of

chaperones ibpAB, DnaKJ, and GroSL, as well as the tags SUMO and MBP, were found to enhance the aesculin titer. However, these improvements remained significantly lower than those achieved through enzyme-based redesign. These comparative experiments underscore the reliability and efficiency of the proposed enzyme redesign strategy as an alternative to traditional methods for designing and identifying enzymes that exhibit both high soluble expression and catalytic performance.

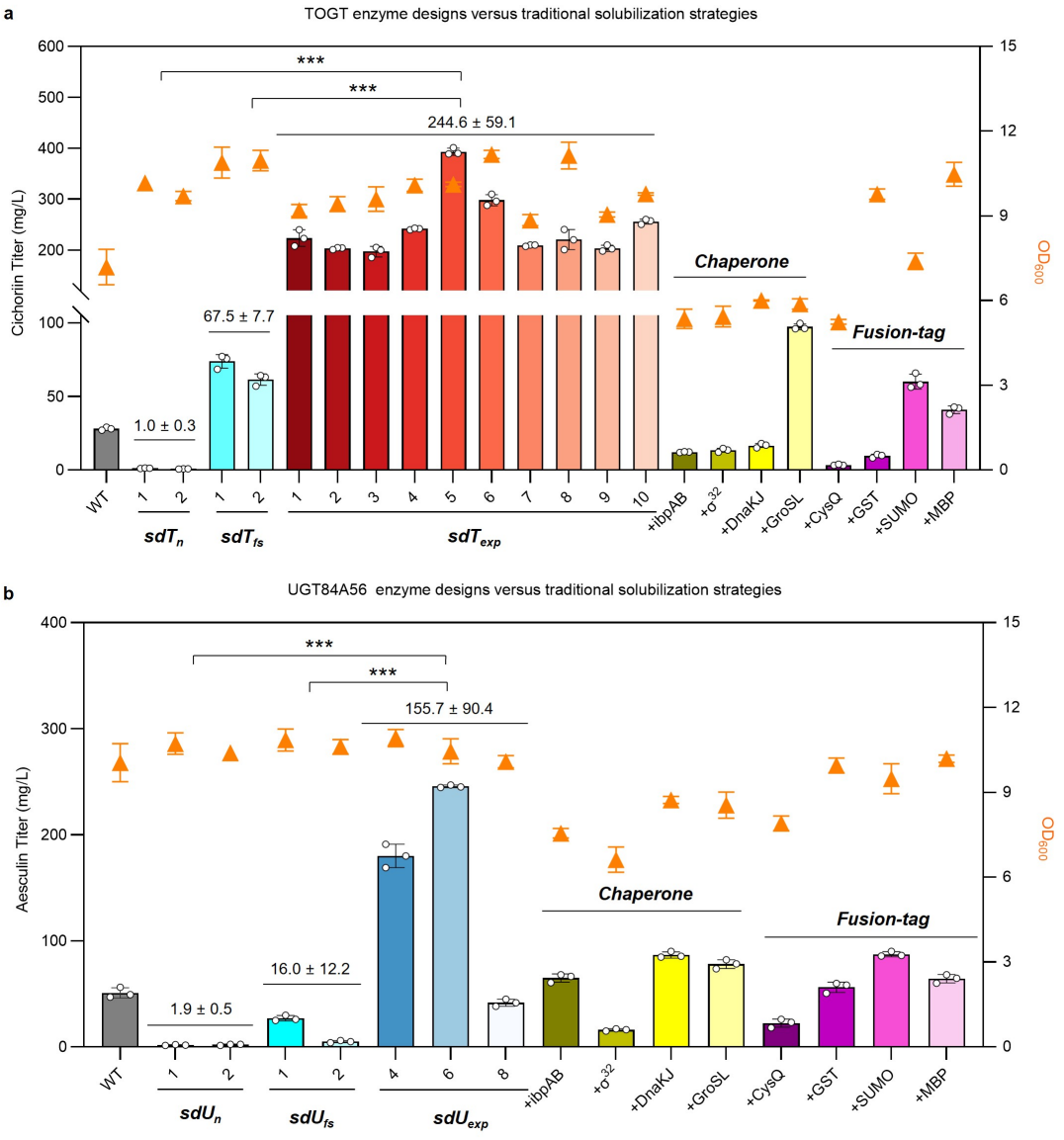


Fig. 4. Comparison of whole-cell biotransformation titers between protein redesign strategies and other expression-enhancing strategies. a, Cichoriin production by host strain BW25113 expressing native TOGT, TOGT redesigns (sdT_{exp}1-10, sdT_{n}1-2, sdT_{fs}1-2), TOGT co-expressed with molecular chaperones, or fused with solubility tags. Statistical comparisons between the sdT_{exp} designs and other groups were performed using Welch's *t*-test. **b,** Aesculin production by host strain BW25113 expressing

native UGT84A56, UGT84A56 redesigns (sdU_n1-2, sdU_{fs}1-2, sdU_{exp}4, sdU_{exp}6 and sdU_{exp}8), UGT84A56 co-expressed with various molecular chaperones, or fused with different solubility tags. Statistical comparisons between sdU_{exp} designs and other groups were performed using Welch's *t*-test. Data are presented as the mean ± standard deviation from three replicates. Statistical significance is indicated as follows: ***, $P \leqslant 0.001$; **, $P \leqslant 0.01$; *, $P \leqslant 0.05$.

Mechanism analysis of performance-enhanced designs

Molecular simulations (MD) and structural analysis were conducted for the best-performing designs to elucidate the molecular mechanisms underlying the enhanced catalytic efficiency and solubility. The designed enzyme, sdT_{exp}5, shares 76% sequence identity with TOGT, while exhibiting consistent fold and secondary structures (Fig. 5a). Molecular docking results revealed more favorable substrate positioning in the active site, as shown in Fig. 5b, as evidenced by a shorter distance between the Nε atom of the catalytic His^{16} and the O7 atom of esculetin, as well as a reduced distance from the O7 atom to the anomeric carbon C1 of the UDP-glucose moiety. These factors are conductive to the catalytic efficiency of the O-glycosylation mechanism[46] of UDP-glycosyltransferases (Fig. S19). MD simulations were performed for 100 ns in the absence of substrate binding. As shown in Fig. 5c, RMSF analysis indicated a decrease in conformational flexibility in regions proximal to the substrate-binding pocket, while increased flexibility was observed in distal regions, suggesting stable substrate positioning and a more flexible substrate channel (Fig. S20). Hydrophobicity analysis revealed notable increases in surface hydrophilicity, which may contribute to the improved enzyme solubility (Fig. 5d). At the global structural level, the redesigned enzyme exhibits several favorable features compared to TOGT, including a reduced surface-to-volume ratio (Table S3), a smaller solvent-accessible surface area (SASA) (Fig. 5e), a lower RMSD value (Fig. S20), a reduced radius of gyration (R_g) (Fig. S21), and a similar global hydrophobicity score (Figs. 5f and S22). Collectively, these features indicate a more compact structure and improved organization of the active site, which facilitate substrate binding and contribute to enhanced overall structural stability.

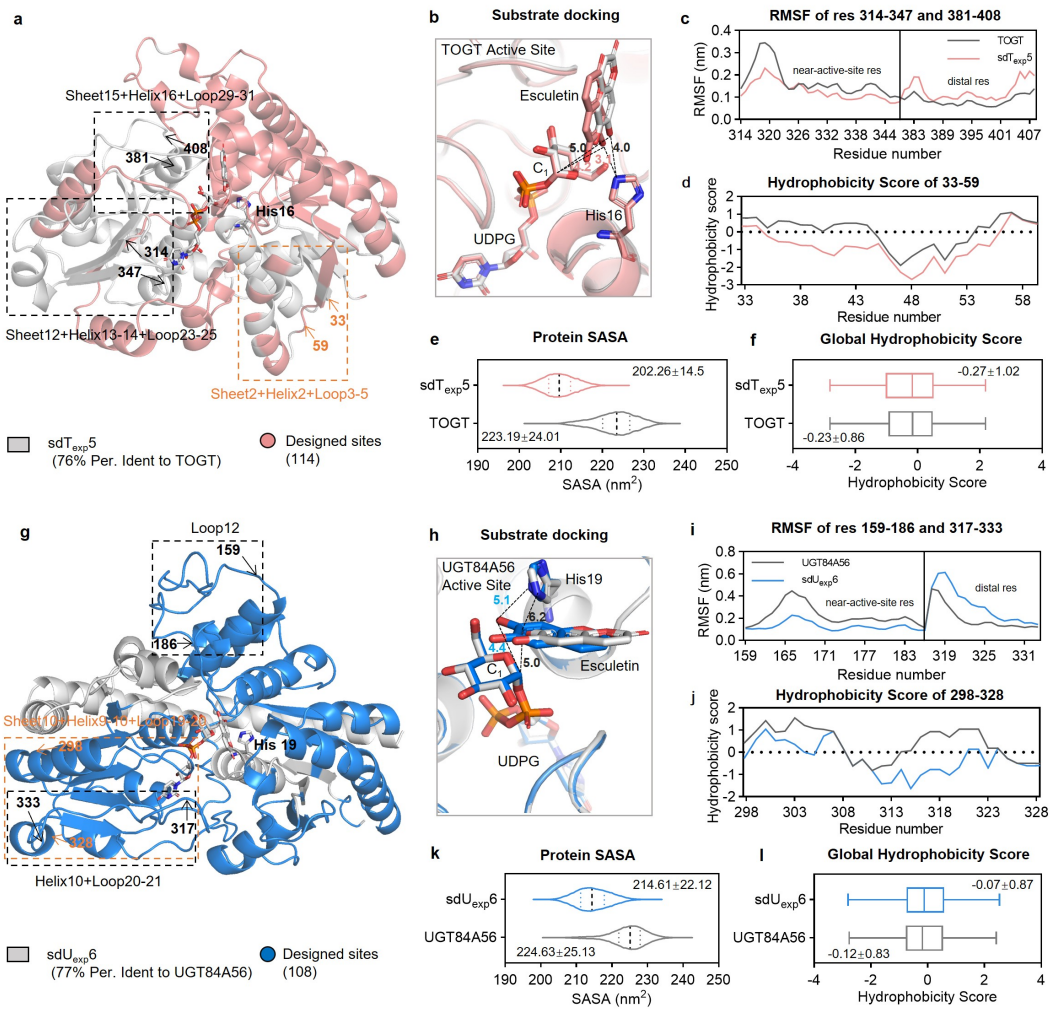


Fig. 5. Structural characteristics analysis of TOGT, UGT84A56 and their redesigned enzymes. a, Structures of sdT$_{exp}$5, with the fixed sites colored as grey and the designed as pink. Regions with significant differences from the original TOGT were labeled with their local secondary structures and the start-end residue numbers. **b,** Substrate docking for TOGT and sdT$_{exp}$5. The TOGT structure was colored as grey, while the sdT$_{exp}$5 structure was colored as pink. Dashed lines indicate the distances between the Nε atom of His16 and the O7 atom of esculetin, as well as between the O7 atom of esculetin and the C1 atom of UDP-Glc. **c,** RMSF values of residues 314-347 and 381-408, corresponding to Sheet 12, Helix 13-14 and Loop 23-25 near the substrate pocket and Sheet 15, Helix 16 and Loop 29-31 in distal regions, respectively. **d,** Hydrophobicity scores of residues 33-59, corresponding to Sheet 2, Helix 2 and Loop 3-5. **e,** Protein SASA analysis of 100 ns simulation trajectory of TOGT and sdT$_{exp}$5, summarizing results from 1000 frames, with mean and quartiles marked. **f,** Global hydrophilicity analysis of the equilibrated structures of TOGT and sdT$_{exp}$5, summarizing data across all residues. Error bars represent the standard variance. **g,** Structures of sdU$_{exp}$6, with the fixed sites colored as grey and the designed as blue. Regions

with significant differences from the original UGT84A56 were labeled with their local secondary structures and the start-end residue numbers. **h,** Substrate docking for UGT84A56 and $sdU_{exp}6$. The UGT84A56 structure was colored as grey, while the $sdU_{exp}6$ structure was colored as blue. Dashed lines indicate the distances between the Nε atom of His^{16} and the O6 atom of esculetin, as well as between the O6 atom of esculetin and the C1 atom of UDP-Glc. **i,** RMSF values of residues 159-186 and 317-333, corresponding to Loop 12 near the substrate pocket and Helix 10 and Loop 20-21 in the distal regions, respectively. **j,** Hydrophobicity scores of residues 298-328, corresponding to Sheet 10, Helix 9-10 and Loop 19-20. **k,** Protein SASA analysis of 100 ns simulation trajectory of UGT84A56 and $sdU_{exp}6$, summarizing results from 1000 frames, with mean and quartiles marked. **l,** Global hydrophilicity analysis of the equilibrated structures of UGT84A56 and $sdU_{exp}6$, summarizing data across all residues. Error bars represent the standard variance.

Similarly, structural factors contributing to the catalytic activity enhancement of the redesigned enzyme $sdU_{exp}6$ compared to UGT84A56 were analyzed (Fig. 5g-l). Molecular docking with the substrate esculetin revealed shorter distances for deprotonation and nucleophilic attack, consistent with enhanced catalytic efficiency (Fig. 5h). RMSF analysis indicated reduced flexibility in the loop region (residues 159-183) near the active site and increased flexibility in the helix region (residues 317–333) in the distal regions (Figs. 5i and S20). Hydrophilicity analysis showed localized increases in hydrophilicity in the outer-shell regions of the enzyme (Fig. 5j). In addition, comparisons of global structural metrics revealed a lower solvent-accessible surface area (SASA) (Fig. 5k), a smaller surface-to-volume ratio (Table S4), a lower RMSD (Fig. S20), a reduced R_g (Fig. S21) and a similar overall hydrophilicity score (Fig. 5l and S23). These findings parallel the observations for TOGT and $sdT_{exp}5$, suggesting that the redesign enhances catalytic activity and expression through mechanisms involving dynamic conformational changes and structural compactness.

Microbial production of cichoriin and aesculin from glucose using *E. coli*

The genes encoding the redesigned enzymes, along with other pathway enzymes, were introduced into the *E. coli* BW1 chassis (Table S6), resulting in the engineered strains BW-$sdT_{exp}5$/$sdT_{exp}6$ and BW-$sdU_{exp}4$/$sdU_{exp}6$ for the biosynthesis of cichoriin and aesculin starting

from glucose, as illustrated in Fig. 6a. Small-scale fermentations in 50 mL shake flasks were conducted to determine the optimal growth conditions for these strains, with the resulting growth curves and product titers shown in Figs. 6b and 6c. During the first 24 hours, rapid cell growth was observed, with biomass peaking at 48 hours. Concurrently, cichoriin and aesculin titers steadily increased throughout the fermentation process. Ultimately, BW-sdT$_{exp}$5 and BW-sdT$_{exp}$6 produced 174 mg/L and 145 mg/L of cichoriin within 3 days, respectively, demonstrating significant improvements over the TOGT-deployed system. Similarly, BW-sdU$_{exp}$4 and BW-sdU$_{exp}$6 achieved aesculin titers of 399 mg/L and 441 mg/L within 3 days, respectively, showcasing enhanced production efficiency. Fed-batch fermentations in a 3-L bioreactor were performed using the high-yielding strains BW-sdT$_{exp}$5 and BW-sdU$_{exp}$6, with results presented in Fig. 6d and 6e. After 120 hours, BW-sdT$_{exp}$5 achieved a maximum OD$_{600}$ of 49 and a cichoriin yield of 2110 mg/L, while BW-sdU$_{exp}$6 achieved an OD$_{600}$ of 65 and an aesculin yield of 4050 mg/L—both representing the highest fermentation yields reported to date for cichoriin and aesculin production. Intermediate product sampling was conducted during fermentation, revealing that esculetin levels became nearly undetectable, indicating that the glycosyltransferases are no longer the rate-limiting enzymes. Future engineering efforts should therefore focus on the synthesis of the precursor esculetin.

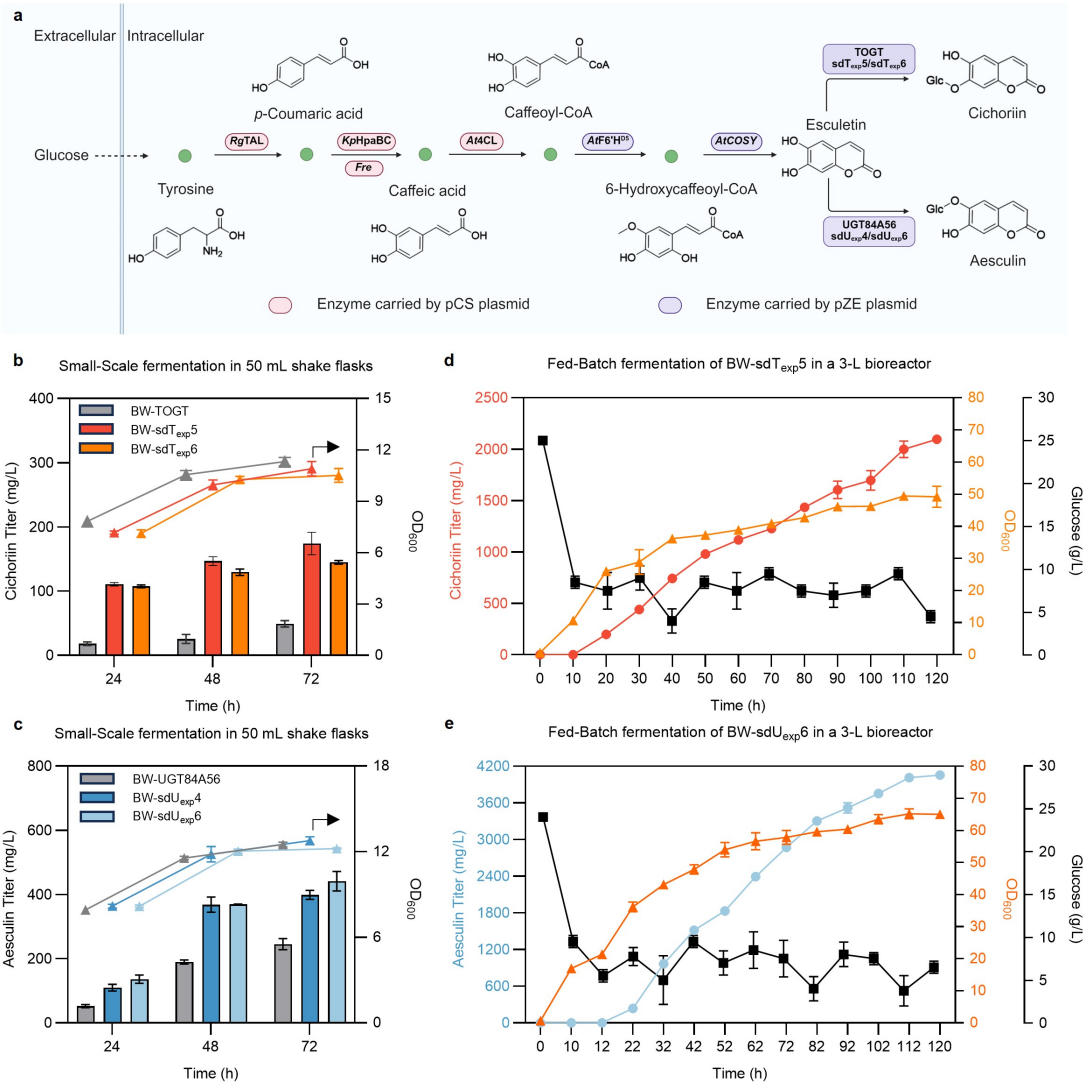


Fig. 6. Microbial production of cichoriin and aesculin with the redesigned enzymes. a, Biosynthetic pathways for cichoriin and aesculin from glucose. Abbreviations: *Rg*TAL, tyrosine ammonia lyase from *Rhodotorula glutinis*; *Kp*hpaBC, 4-hydroxyphenylacetate-3-hydroxylase from *Klebsiella pneumoniae*; Fre, NAD(P)H-flavin reductase from *E. coli*; *At*4CL, 4-coumarate: CoA ligase from *Arabidopsis thaliana*; *At*F6'H, feruloyl-CoA6'-hydroxylase from *Arabidopsis thaliana*; *At*COSY, coumarin synthase from *Arabidopsis thaliana*. **b,** 50-mL shake-flask fermentation of cichoriin. Bars in grey, red and orange represent strains BW-TOGT, BW-sdT$_{exp}$5 and BW-sdT$_{exp}$6, respectively. **c,** 50-mL shake-flask fermentation of aesculin. Bars in grey, deep blue and light blue correspond to strains BW-UGT84A56, BW-sdU$_{exp}$4, and BW-sdU$_{exp}$6, respectively. **d,** 3-L fed-batch fermentation for cichoriin production using strain BW-sdT$_{exp}$5. **e,** 3-L fed-batch fermentation for aesculin production using strain BW-sdU$_{exp}$6. The error bars in figures **b-e** represent standard deviations from two independent measurements.

Discussion

Efficient expression of heterologous enzymes remains a major challenge in constructing microbial cell factories for functional molecule synthesis[3,4], particularly when recreating natural product biosynthetic pathways using enzymes derived from higher plants and animals. Such efforts often require extensive, empirically driven optimization of host strains, codon usage, expression conditions, and transcriptional and translational regulation[13,14,15,16]. In this study, we developed a ProteinMPNN-assisted sequence redesign strategy to systematically optimize the sequences of enzymes with poor native expression levels. Using glycosyltransferases TOGT and UGT84A56 as examples, this strategy successfully enhanced their soluble expression in *E. coli* without compromising—and in many cases, even boosting—their enzymatic activity. When implemented in whole-cell catalysis, the redesigned enzymes outperformed conventional expression-enhancing strategies, such as fusion with solubility tags and co-expression with molecular chaperones. Using the top-performing designs, we synthesized cichoriin and aesculin from glucose in *E. coli* BW1 cells, achieving record-high titers of up to 2110 mg/L and 4050 mg/L, respectively. Notably, these improvements were achieved without any prior mechanistic knowledge of the enzymes, demonstrating the simplicity and broad applicability of the proposed strategy.

Our approach achieves a balance between sequence fixation and regeneration. A key factor in this success is the strategic fixation of residues based on both evolutionary and functional constraints. Meanwhile, these constraints are implemented through adjustable parameters, such as the radius of the second shell and the conservation level of the residues. As demonstrated in our ablation experiments, fine-tuning the designable space is critical for striking the optimal balance among expressibility, stability, and activity. We believe the same principle could be equally applicable to the redesign of other enzymes.

Designs generated by ProteinMPNN consistently exhibited relatively good solubility, as demonstrated in both the TOGT and UGT84A56 cases. The native TOGT enzyme showed low expression levels, whereas the native UGT84A56 enzyme already exhibited moderate soluble expression, although inclusion bodies remained predominant. From the small pool of

candidates generated by our design workflow, we successfully identified enzymes with improved soluble expression and enhanced both in vitro and in vivo activities. These results highlight the effectiveness of our strategy for both cases, although the likelihood of activity improvement was higher for TOGT. This might reflect the trade-off between the foldability and activity. Future developments could incorporate customized evaluation functions to optimize template enzymes with different expression levels and baseline activities.

This study echoes the well-established understanding that protein structures and functions are far more conserved than sequences[47]. Previous bioinformatics studies have established that 60% sequence identity serves as a reliable threshold for accurate functional annotation[48], indicating that evolution tolerates large-scale variations in protein sequence while maintaining a relatively conserved structure and function[49,50]. For the two enzymes we redesigned, the well-preserved enzymatic activity following extensive sequence mutagenesis (altering more than 20% of the full sequence) demonstrates that non-conserved regions in enzymes are highly tolerant to mutations. While this tolerance provides opportunities for enzyme optimization, natural evolution alone is too slow to fully exploit this potential[51]. In contrast, inverse folding models, such as ProteinMPNN, which are trained to capture sequence features essential for successful folding[32], can propose large sets of reasonable mutation combinations at non-conserved residues. This enables *in silico* exploration of sequence space on a scale order of magnitude beyond the reach of conventional evolution-based strategies. Combined with efficient computational filtering, this approach facilitates the engineering of superior enzymes with tailored properties for industrial applications[52]. We anticipate that such deep learning-assisted approaches will prove broadly applicable across diverse enzyme classes, thereby offering a powerful framework for advancing microbial cell factory design and biocatalysis.

Methods

Sequence Redesign and Filtering

Sequences and Structural Modeling. The sequence of TOGT from *Nicotiana tabacum* was fetched from the Non-redundant Protein Sequences (nr) database (NCBI accession: Q9AT54.1). The UGT84A56 sequence of *Aesculus chinensis* was obtained from the report by Sun et al[28]. The structures of all these sequences were predicted using the AlphaFold2 webserver[40].

Structure-Based Fixation. Molecular docking was performed using AutoDockTools (1.5.7)[53], and the substrate-enzyme conformations with the lowest ΔG_{bind} were selected. Residues in direct contact with the ligands were designated as the first-shell residues, and residues interacting with these residues, either via sidechain packing or polar contacts, were defined as the second-shell residues. This structure-based strategy effectively constrained all sites within 8 Å of the ligand-binding pocket. Specifically, the 8 Å radius preserved catalytically essential residues, while allowing ProteinMPNN to redesign solvent-exposed regions distant from the active site.

Sequence-Based Fixation. Evolutionary conservation analysis was conducted using the ConSurf server[38], which evaluates amino acid conservation through multiple sequence alignment (MSA) and evolution-based algorithms. The sequence of the target protein was submitted to the ConSurf with the WAG evolutionary model. Subsequently, 300 sequence alignments were generated to assess site conservation. Residues with conservation score $\geqslant 5$ (scale: 1-9) were designed as fixed sites. This approach reflects the intrinsic conservation of each site, providing reliable natural evolutionary constraints.

Sequence Redesign with ProteinMPNN. During sequence generation, ProteinMPNN preserved the fixed sites, while allowing designable sites to mutate among 19 amino acids (excluding cysteine) to generate optimal sequences[34,37]. The sampling temperature was set to 0.1. Additional details are provided in the Supporting Information.

Sequence Filtering. The generated designs were filtered using a two-tier scoring: the Global Score and the Weighted Score. Designs were ranked in ascending order by the Global Score, then in descending order by the Weighted Score. The top-ranked designs were selected as candidates for

experimental testing.

The Global Score derived from the ProteinMPNN, measures how well a predicted protein sequence fits the desired structural fold and has the potential to be stable in a biological context[39]. Lower scores usually correlate with better structural and functional potential for the protein.

The Weighted Score was calculated from four metrics:

(1) pLDDT (predicted Local Distance Difference Test): Calculated by AlphaFold2, pLDDT assesses the confidence of local residue-level structure predictions[54]. The score ranges from 0 to 100, with higher values indicating greater confidence.

(2) pTM (predicted TM-score): evaluates the structural similarity between the predicted and native structures[40,54]. A pTM score closer to 1.0 suggests a higher similarity, indicating that the predicted structure closely matches the native fold.

(3) RMSD (Root Mean Square Deviation): measures the average distance between atoms in the predicted and native structures[55,56]. Lower RMSD values indicate better structural alignment, with values typically ranging from 0 to several Ångströms.

(4) Solubility prediction by SoluProt[41], a tool developed using gradient boosting machine techniques with the TargetTrack database[57] as a training set, estimates the likelihood of the protein being soluble in *E. coli*. The solubility score ranges from 0 to 1, with higher values indicating a greater probability of soluble expression.

The Weighted Score was a weighted sum of the above four scores after normalization to 0-1, calculated as follows, with the weighting factors empirically determined:

$$\text{Weighted Score} = 0.1 \times pTM_{norm} + 0.2 \times pLDDT_{norm} + 0.3 \times solubility_{norm} + 0.4 \times RMSD_{inv\text{-}norm}.$$

Experiments and Simulation

Heterologous Expression and Purification. Recombinant plasmids encoding the enzymes were transformed into *E. coli* BL21(DE3). When OD600 reached 0.6-0.8, protein expression was induced with 0.5 mM IPTG at 16°C, 220 rpm for 18 hours. Purification was performed using Ni-NTA affinity

chromatography to isolate the His-tagged proteins. For more detailed information on the experimental section, please refer to the Supplementary Information.

Protein Expression Quantification. Protein expression levels were quantified by the BCA Protein Assay Kit. The expression of target proteins was confirmed by SDS-PAGE and Western blotting using Anti-His-tag antibodies.

Crude Enzyme Lysate Activity Determination. Enzyme activity in crude lysates was assessed by incubating the lysates with UDP-glucose and the respective substrates at 30°C. Reactions were terminated after 60 minutes and analyzed by HPLC.

Kinetic Parameters Measurement. The kinetic parameters were measured in 400 μL reaction mixtures containing 2 mM UDPG, 50 mM Tris-HCl pH 8.0, a certain concentration of enzymes (0.5-1 μM), and different concentrations of substrate (10-500 μM) at 30 °C for 30 or 60 minutes. Michaelis–Menten constants (K_m) and turnover numbers (k_{cat}) were determined by fitting the initial steady-state velocity to the Michaelis-Menten equation.

Melting Temperature Determination. The protein melting temperature (T_m) was determined using Nano Differential Scanning Fluorimetry (NanoDSF) with a Prometheus NT.48 instrument, which measures protein thermal stability by monitoring changes in intrinsic fluorescence. The samples were heated from 20°C to 95°C at a rate of 1°C/min.

Culture Media and Conditions in Feeding/*De Novo* Synthesis Experiments. M9Y medium was used for feeding and *de novo* synthesis experiments. Recombinant plasmids were electroporated into the host strains, and protein expression was induced with IPTG. Product formation was analyzed by HPLC.

Culture Media and Conditions in Fed-Batch Experiments. Seed cultures with OD_{600} of 8-10 were transferred to a 3 L bioreactor with 1 L of medium at a 10% inoculum. The fermentation medium contained Na_2HPO_4, KH_2PO_4, $(NH4)_2SO_4$, yeast extract, $CaCl_2$, $MgSO_4$, VB1, ampicillin, kanamycin, and TES, with glucose at 25 g/L initially. After reaching OD_{600} of 15-20, 0.5 mM IPTG was added to induce protein expression. The pH was controlled at 7.0, and dissolved oxygen was maintained at 30%. Glucose was fed to maintain a concentration from exceeding 10 g/L. Two independent experiments were performed.

High-Performance Liquid Chromatography (HPLC) Analysis. Target product identification and quantification were conducted by HPLC using a Shimadzu system with reverse-phase chromatography. The separation was performed on a ChromCore 120 C18 column, and the mobile phase consisted of 0.2% TFA and methanol. UV detection was performed at 345 nm, where the absorption is primarily attributed to the conjugated aromatic ring systems and the phenolic hydroxyl groups (-OH) present in cichoriin and aesculin.

Molecular Dynamics Simulations. Molecular dynamics simulations were conducted using GROMACS 2018 package[58] with the CHARMM36 force field[59]. The system was solvated, energy minimized, equilibrated under the NVT ensemble for 100 ps, followed by 100 ps in the NPT ensemble, and the production run was performed for 100 ns with a time step of 2 fs. Structural stability and conformational changes were evaluated by calculating the root mean square deviation (RMSD), root mean square fluctuation (RMSF), solvent-accessible surface area (SASA), and radius of gyration (*Rg*) using built-in GROMACS tools. Hydrophobicity/hydrophilicity analysis was conducted with the Protscale tool[60]. Simulation trajectories were visualized and analyzed with PyMOL 2.5.4 and VMD 1.9.3.

References

1. Cravens, A., Payne, J. & Smolke, C. D. Synthetic biology strategies for microbial biosynthesis of plant natural products. *Nat. Commun*. **10**, 2142 (2019).
2. Kim, G. B., Kim, H. R., & Lee, S. Y. Comprehensive evaluation of the capacities of microbial cell factories. *Nat. Commun.* **16**, 2869 (2025).
3. Pitera, D. J., Paddon, C. J., Newman, J. D. & Keasling, J. D. Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. *Metab. Eng.* **9**, 193–207 (2007).
4. Nielsen, J., Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185-1197 (2016).
5. Zhang, J. et al. A microbial supply chain for production of the anti-cancer drug vinblastine. *Nature* **609**, 341–347 (2022).
6. Gao, J. et al. Biosynthesis of catharanthine in engineered *Pichia pastoris*. *Nat. Synthesis.* **2**, 231–242 (2023).
7. Lv, Y., Marsafari, M., Koffas, M., Zhou, J. & Xu, P. Optimizing oleaginous yeast cell factories for flavonoids and hydroxylated flavonoids biosynthesis. *ACS Synth. Biol*. **8**, 2514–2523 (2019).
8. Li, H. et al. Synergetic engineering of multiple pathways for *de novo* (2S)-naringenin biosynthesis in *Saccharomyces cerevisiae*. *ACS Sustainable Chem. Eng*. **12**, 59–71 (2024).
9. Chang, M. C. Y., Eachus, R. A., Trieu, W., Ro, D.-K., & Keasling, J. D. Engineering *Escherichia coli* for production of functionalized terpenoids using plant P450s. *Nat. Chem. Biol.* **3**, 274–277 (2007).
10. Chan, P. P., Cooney, M. J., Ding, Y., Arkin, A. P. & Church, G. M. Enhancing the translational capacity of *E. coli* by resolving codon bias. *ACS Synth. Biol.* **7**, 2656–2664 (2018).
11. Rosano, G. L. & Ceccarelli, E. A. Folding of heterologous proteins in bacterial cell factories: Cellular mechanisms and engineering strategies. *Microb. Cell Fact.* **21**, 170 (2022).
12. Fang, X., Baum, D. A., Yang, M. C., Shen, C. R. & You, L. Omics-guided bacterial engineering of *Escherichia coli* ER2566 for recombinant protein expression. *Biotechnol. Bioeng.* **120**, 689–702 (2023).
13. Nishihara, K., Kanemori, M., Kitagawa, M., Yanagi, H. & Yura, T. Chaperone coexpression

plasmids: Differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of *Japanese cedar* pollen, Cryj2, in *Escherichia coli*. *Appl. Environ. Microbiol.* **64**, 1694–1699 (1998).

14. Schmidt, M. et al. Maximizing heterologous expression of engineered type I polyketide synthases: investigating codon optimization strategies. *ACS Synth. Biol*. **12**, 3366–3380 (2023).
15. Jiang, R. et al. Strategies to overcome the challenges of low or no expression of heterologous proteins in *Escherichia coli*. *Biotechnol. Adv*. **75**, 108417 (2024).
16. Kotopka, B. J., Li, Y. & Smolke, C. Synthetic biology strategies toward heterologous phytochemical production. *Nat. Prod. Rep*. **35**, 902-920 (2018).
17. Ki, M.-R. & Pack, S. P. Fusion tags to enhance heterologous protein expression. *Microbiol. Biotechnol.* **104**, 2411–2425 (2020).
18. Lee, J.-M., Kim, J. H., Kim, J. Y., Oh, M.-K., & Kim, B.-G. Enhancing the soluble expression of α1,2-fucosyltransferase in *Escherichia coli* using high-throughput flow cytometry screening coupled with a split GFP system. *J. Biotechnol*. **387**, 49–57 (2024).
19. Jhamb, K., & Sahoo, D. K. Production of soluble recombinant proteins in *Escherichia coli*: Effects of process conditions and chaperone co-expression on cell growth and production of xylanase. *Bioresour. Technol*. **123**, 135–143 (2021).
20. Köppl, C. et al. Fusion tag design influences soluble recombinant protein production in *Escherichia coli*. *Int. J. Biol. Macromol.* **23**, 7678 (2022).
21. Jayakumar, T. et al. Activation of Nrf2 by esculetin mitigates inflammatory responses through suppression of NF-κB signaling cascade in RAW 264.7 cells. *Molecules.* 2**7**, 5143 (2022).
22. Luo, Y. et al. Protective effects and mechanisms of esculetin against H_2O_2-induced oxidative stress, apoptosis, and pyroptosis in human hepatoma HepG2 cells. *Molecules.* **29**, 1415 (2024).
23. Kwak, J.-H., Kim, Y., Staatz, C. E. & Baek, I.-H. Oral bioavailability and pharmacokinetics of esculetin following intravenous and oral administration in rats. *Xenobiotica*. **51**, 811–817 (2021).
24. Singh, A. K., Garg, A., & Sagar, B. P. S. Determination of in vitro produced cichoriin from callus culture of *Cichorium intybus* L. *Asian J. Chem*. **24**, 2935–2937 (2012).
25. Ju, S. et al. Antioxidant and anti-inflammatory effects of esculin and esculetin. *Exp. Ther. Med*. **27**, 248 (2024).

26. Moon, K. et al. Synthesis of α-cichoriin using *Deinococcus geothermalis* amylosucrase and its antiproliferative effect. *Microbiol. Biotechnol. Lett*. **50**, 218–227 (2022).
27. Park, S., Moon, K., Park, C. S., Jung, D. H., & Cha, J. Synthesis of aesculetin and aesculin glycosides using engineered *Escherichia coli* expressing *Neisseria polysaccharea* amylosucrase. *J. Microbiol. Biotechnol*. **28**, 566-570 (2018).
28. Sun, W. et al. Characterization of the horse chestnut genome reveals the evolution of aescin and aesculin biosynthesis. *Nat. Commun*. **14**, 6470 (2023).
29. Xie, C. et al. Establishing a coumarin production platform by protein and metabolic engineering. *Metab. Eng*. **86**, 89-98 (2024).
30. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
31. Shanker, V. R., Bruun, T. U. J., Hie, B. L. & Kim, P. S. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* **385**, 46–53 (2024).
32. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
33. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol*. **41**, 1099–1106 (2023).
34. Sumida, K. H. et al. Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
35. Hansen, A. L. et al. Carving out a glycoside hydrolase active site for incorporation into a new protein scaffold using deep network hallucination. *ACS Synth. Biol*. **13**, 862–875 (2024).
36. Ding, Y., Zhang, S., Kong, X., Hess, H. & Zhang, Y. Replicating PET hydrolytic activity by positioning active sites with smaller synthetic protein scaffolds. *Adv. Sci*. **12**, 2500859 (2025).
37. King, B. R., Sumida, K. H., Caruso, J. L., Baker, D. & Zalata, J. Computational stabilization of a non-heme iron enzyme enables efficient evolution of new function. *Angew. Chem. Int. Ed*. **64**, e202414705 (2024).
38. Goldenberg, O., Erez, E., Nimrod, G. & Ben-Tal, N. The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res*. **37**, D323–D327 (2008).
39. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).

40. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

41. Hon, J. et al. SoluProt: Prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics.* **37**, 23–28 (2021).

42. Zhang, Q. & Liu, H. Improving the soluble expression of difficult-to-express proteins in common prokaryotic expression systems: Strategies and advances. *Crit. Rev. Biotechnol*. **43**, 456–478 (2023).

43. Kaur, J., Kumar, A. & Kaur, J. Strategies for optimization of heterologous protein expression in *E. coli*: Roadblocks and reinforcements. *Int. J. Biol. Macromol*. **803**, (2017).

44. Chen, L. et al. A fusion protein strategy for soluble expression of *Stevia* glycosyltransferase UGT76G1 in *Escherichia coli*. *3 Biotech.* **7**, 356 (2017).

45. Zhu, C. et al. Multiple regulatory mechanisms synergistically control the soluble expression of CsCE for enhanced enzymatic productivity of lactulose in *E. coli*. *J. Agric. Food Chem.* **72**, 41 (2024).

46. Wang, M. et al. Structure–function and engineering of plant UDP-glycosyltransferase. *Comput. Struct. Biotechnol. J*. **21**, 5358–5371 (2023).

47. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins.* **77**, 499–508 (2009).

48. Addou, S., Rentzsch, R., Lee, D. & Orengo, C. A. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *Bioinformatics.* **25**, 1364–1371 (2009).

49. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol*. **19**, 596–604 (2009).

50. Escobedo, A., Voigt, G., Faure, A. J., & Lehner, B. Genetics, energetics, and allostery in proteins with randomized cores and surfaces. *Science* **389**, eadq3948 (2025).

51. Maddamsetti, R. Universal constraints on protein evolution in the long-term evolution experiment with *Escherichia coli*. *Genome Biol. Evol*. **13**, evab070 (2021).

52. Johnson, S. R. et al. Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nat. Biotechnol*. **43**, 396–405 (2025).

53. Morris, G. M. et al. AutoDock4 and AutoDockTools4: Automated docking with selective

receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

54. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

55. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31**, 3370–3374 (2003).

56. Zhang, Y., & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

57. Berman, H. M., Gabanyi, M. J., Kouranov, A., Micallef, D. I., & Westbrook, J. Protein Structure Initiative—TargetTrack 2000–2017—All Data Files. *Zenodo.* (2017).

58. Lemkul, J. A. From proteins to perturbed Hamiltonians: A suite of tutorials for the GROMACS-2018 molecular simulation package. *Living J. Comput. Mol. Sci.* **1**, 5068 (2019).

59. Huang, J. et al. Biomolecular simulations: Theory and applications. *Biophys. J.* **112**, 175a–176a (2017).

60. Gasteiger, E. et al. Protein Identification and Analysis Tools on the ExPASy Server. In: Walker, J.M. (eds) The Proteomics Protocols Handbook. Springer Protocols Handbooks. *Humana Press*, pp 571–607 (2005).

Acknowledgement

This work was supported by the National Nature Science Foundation of China under grant number 32371325, Fundamental Research Funds for the Central Universities (QNTD2023-01), and Interdisciplinary Research Center of Beijing University of Chemical Technology (No. XK2025-05).

Author information

Authors and Affiliations

State Key Laboratory of Chemical Resources Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Yifei Zhang, Shanshan Zhang, Yujing Ding, Xiaofan Lin, Zhipeng Yuan, Xinxiao Sun & Qipeng Yuan

Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China

Yifei Zhang, Shanshan Zhang, Yujing Ding, Xiaofan Lin, Zhipeng Yuan

Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

Yilei Han, Zheng Liu

College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Xinxiao Sun, Qipeng Yuan

Contributions

Y.Z. conceived and designed the research. S.Z., Y.D. and X.L. performed the computational design. S.Z. and Z.Y. performed the experiments. S.Z., Y.H. and Y.Z. wrote the initial manuscript. S.Z., Y. H., Y.Z., X.S, Q.Y., and Z. L. discussed the results and revised the manuscript.

Declaration of interests

The authors declare no competing financial interests.

Data availability

All data supporting the findings of this study are included within the Article and its Supplementary information. Correspondence and requests for materials should be addressed to Y.Z.

Supplementary information

Supplementary information includes details of the materials, experimental methods, computational design and analysis, as well as Supplementary Figures, Tables, Notes, and References.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationSequenceRedeisgn.pdf
- rs.pdf