

EZPro-Fit: A Thermophile-Informed Cross-Modal Generative Framework Integrating DNA Language and Protein Structure for Enzyme Mutant Design

Ran Xu

Macao Polytechnic University

Linyan Zhao

Nanjing Agricultural University

Ronghua Pei

Nanjing Agricultural University

Yan Xu

Nanjing Agricultural University

Zhiyuan Zhou

Nanyang Technological University

Hongliang Duan

Macao Polytechnic University

Leyi Wei

Macao Polytechnic University

Liangzhen Zheng

`zhenglz@zelixir.com`

Shenzhen Zelixir Biotech Company Ltd.

Dongyang Liu

`liudongyang@njau.edu.cn`

Nanjing Agricultural University

Jingjing Guo

`jguo@mpu.edu.mo`

Macao Polytechnic University <https://orcid.org/0000-0002-4632-4364>

Research Article

Keywords: Protein engineering, DNA, thermostability, generative model, Flow-matching, cel5E

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11377962/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

EZPro-Fit: A Thermophile-Informed Cross-Modal Generative Framework Integrating DNA Language and Protein Structure for Enzyme Mutant Design

Ran Xu^{*1}, Linyan Zhao^{*2}, Ronghua Pei², Yan Xu², Zhiyuan Zhou³, Hongliang Duan¹, Leyi Wei¹, Liangzhen Zheng^{4,*}, Dongyang Liu^{2,*}, and Jingjing Guo^{5,*}

¹Centre in Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macao, China

²College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing, China

³School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

⁴Shenzhen Zelixir Biotech Company Ltd., Shenzhen, China

⁵Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence, Macao Polytechnic University, Macao, China

Abstract

Generative models are transforming enzyme engineering by exploring vast protein sequence spaces while reducing experimental screening. However, most approaches optimize proteins without considering their genetic encoding and host expression. Here, we introduce EZPro-Fit, a multimodal generative framework that jointly learns protein structures, coding sequences (CDSs), and taxonomic information from thermophilic microorganisms, coupling structural design with nucleotide-level implementation. Pretraining on thermophilic proteins followed by *Escherichia coli*-conditioned fine-tuning captures signals related to protein conformation, thermal resilience, codon usage, and host compatibility. EZPro-Fit enables bidirectional inference between structures and CDSs and generates host-adapted coding sequences while preserving protein-level constraints. Cross-species benchmarks demonstrate improved codon adaptation and learning of structure-associated coding principles of thermostable proteins. For Cel5E engineering, EZPro-Fit identified novel variants, including D259E and N291D, that experimentally improved both melting temperature and catalytic activity. EZPro-Fit thus provides a genetically executable framework for host-aware enzyme design and industrial biocatalysis.

Keywords: Protein engineering, DNA, thermostability, generative model, Flow-matching, cel5E

^{*}These authors contributed equally to this work.

Industrial enzymes play a central role in biotechnology, green chemistry, and renewable energy conversion, where catalytic efficiency, thermostability, and production robustness largely determine the economic viability and sustainability of industrial bioprocesses [1], [2], [3], [4]. Consequently, enhancing enzyme thermostability while preserving catalytic performance has become a major objective of modern protein engineering. Recent advances in AI-driven generative protein design, including inverse-folding models such as ESM-IF1 and ProteinMPNN and diffusion-based models such as RFDiffusion, have substantially expanded the accessible protein design space by enabling structure-guided generation of protein sequences and conformations [5], [6], [7]. However, these approaches remain insufficiently aligned with the practical requirements of industrial enzyme engineering.

Most existing generative frameworks are designed to optimize protein-level representations, focusing primarily on structural compatibility or sequence plausibility rather than explicit optimization for thermostability. More importantly, they largely overlook the complementary biological information encoded in coding DNA sequences. Protein sequences define the three-dimensional architecture underlying enzyme function, while coding DNA sequences encode nucleotide-level characteristics associated with codon usage, mRNA structure, translation dynamics, and organism-specific adaptation [8], [9], [10], [11]. These characteristics are not explicitly represented in protein structures or amino acid sequences, but can influence heterologous protein expression, co-translational folding, and the functional implementation of engineered enzymes [12], [13]. Incorporating complementary protein structural and nucleotide-level representations therefore provides an opportunity to move beyond conventional protein-centric design and establish a cross-modal framework for thermostable enzyme engineering.

Computational tools specifically developed for protein stability provide a complementary set of capabilities. Physics- and energy-based methods, including FoldX and Rosetta, estimate mutation-induced changes in folding free energy and are commonly used to rank candidate substitutions [14], [15]. FireProt combines energy calculations with evolutionary information to design stabilized single- and multiple-point mutants [16], whereas ThermoMPNN uses structure-derived representations to predict stability changes caused by point mutations [17]. Although these methods are valuable for reducing the burden of experimental screening, most function primarily as scoring, prediction, or candidate-ranking tools and generally require a predefined wild-type sequence or an enumerated mutation set. Consequently, they primarily explore the sequence space locally and do not directly generate coding DNA sequences from protein structures.

These limitations reveal a fundamental disconnect between general-purpose protein generation and task- and host-specific enzyme optimization. Although generative models efficiently explore protein sequence or structural space [7], [18], they rarely incorporate objectives that jointly account for stability, catalytic activity, expression efficiency, and host adaptation. In contrast, property-oriented engineering tools can prioritize mutations associated with specific traits but often lack a unified generative framework for balancing multiple functional requirements. This is important because optimizing a single property in isolation may compromise overall enzyme performance [19], [20]; for instance, excessive stabilization can reduce the conformational flexibility needed for catalysis, producing the well-known stability–activity trade-off. Moreover, both paradigms mainly represent proteins at the amino acid sequence or structural level, leaving the biological information encoded in CDSs, including codon usage, nucleotide composition, translational features, and organism-specific adaptation, largely unexplored [21].

Here, we developed EZPro-Fit, a thermophile-informed cross-modal generative framework that establishes a direct bridge between protein structural representations and coding DNA sequences for task- and host-adaptive enzyme engineering. Rather than generating amino acid sequences alone, EZPro-Fit maps protein three-dimensional backbones directly to coding DNA sequences, enabling the model to jointly exploit geometric constraints and nucleotide-level information associated with evolutionary adaptation. The framework integrates a geometry-aware MPNN encoder with a conditional flow-matching decoder to learn the cross-modal correspondence between protein structures and coding sequences. To capture both evolutionarily shaped protein-level constraints and host-adaptive coding patterns, EZPro-Fit is pretrained on paired protein structure-CDS data from thermophilic microorganisms and subsequently fine-tuned using highly expressed *Escherichia coli* genes through a dynamically weighted multi-objective optimization strategy. Using the industrial endoglucanase Cel5E as a test case, we show that ec.EZPro-Fit can discover variants with gains in both thermostability and catalytic activity. These findings highlight structure-to-DNA generation as a promising strategy for integrating protein-level design with genetic implementation in host-aware enzyme engineering.

RESULTS AND DISCUSSION

EZPro-Fit jointly models protein geometry, coding sequences and taxonomic information

EZPro-Fit is a thermophile-informed end-to-end cross-modal generative framework for designing high-performance enzyme mutants by integrating DNA language, protein structure and taxonomic information (Figure 1A). The input data comprised structure-CDS pairs, with 3D coordinates from the AlphaFold database and corresponding DNA sequences from UniProt cross-references (Figure 1B). In contrast to conventional two-step strategies that decouple protein sequence design from downstream nucleotide optimization, EZPro-Fit jointly models structural compatibility and genetic constraints within a unified framework. By learning structure-sequence mappings of proteins from thermophilic species, EZPro-Fit enables rational design of thermostability-enhanced variants while avoiding performance degradation caused by overlooking information at the translational level, circumventing the conventional structure-sequence-DNA pipeline.

As shown in Figure 1C, the framework of EZPro-Fit comprises four sequential modules: multidimensional input integration, Message Passing Neural Network (MPNN) encoding, latent flow-matching, and a dual-decoder feedback loop. Specifically, the pipeline initiates with a multidimensional input layer that synergistically combines structural protein topologies with taxonomic (TAXON) embeddings. These complementary features are concatenated to provide both spatial geometry and evolutionary constraints. Subsequently, the fused representations are processed by an MPNN Encoder. Through T iterations of message-passing operations, the encoder extracts node-level chemical environments to output local residue latent vectors and globally pooled vectors. This encoding phase establishes the source latent distribution, denoted as $q_0(z)$. To bridge the source representation with the desired functional space, a Latent Flow-matching module executes a continuous flow transformation over the time interval $t \in [0, 1]$. This formulation deterministically maps the initial distribution $q_0(z)$ to the target latent distribution $q_1(z)$ via iterative integration. Finally, vectors sampled from the target space, $Z_{\text{target}} \sim q_1(z)$, are routed into a Dual Decoder Feedback loop. One pathway translates the latent codes into amino acid sequences, which a pre-trained ESM2 model then evaluates to verify the biological fidelity and feeds back into the continuous flow. Concurrently, the alternative pathway guides the explicit generation of CDS. This dual-decoder mechanism establishes a closed-loop optimization, ensuring high-fidelity, structure-conditioned sequence generation.

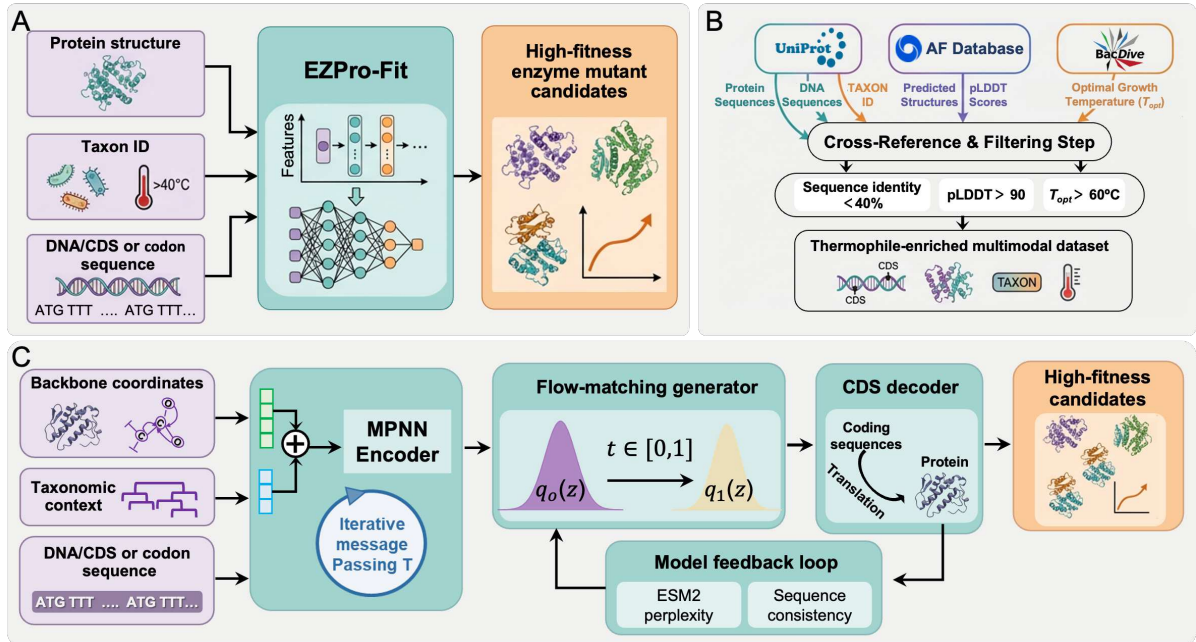


Figure 1: Architecture of EZPro-Fit, a cross-modal generative model trained on thermophilic organism-derived data for enzyme mutant design. (A) EZPro-Fit integrates DNA language, protein structure, and taxonomic context to generate coding sequences and fitness-guided enzyme mutants. (B) Protein sequences, coding sequences, predicted structures, taxonomic annotations, and optimal growth temperatures were collected and cross-referenced from UniProt, AlphaFold DB, and BacDive. Records were filtered by optimal growth temperature ($T_{\text{opt}} \geq 40^\circ\text{C}$), pairwise sequence identity ($< 40\%$), and structural confidence (mean pLDDT > 90) to construct a high-confidence thermophile-enriched dataset. (C) The model architecture integrates protein backbone geometry, DNA/codon-language information, and TAXON embeddings through an MPNN encoder. The resulting latent representation conditions a flow-matching generator to produce coding sequences and mutation candidates. A model-based feedback module evaluates generated variants and promotes candidates with improved predicted fitness.

Overall performance of EZPro-Fit

To rigorously evaluate EZPro-Fit against established protein design methodologies, we conducted comprehensive benchmarking on a held-out test set of thermostable protein structures. The evaluation spanned sequence fidelity, nucleotide-level recovery, structural compatibility, and biological plausibility (Figure 2A). Among baseline approaches, conventional structure-to-sequence SOTA models such as ESM-IF1 and ProteinMPNN achieved competitive residue recovery rates of 51.47% and 47.31%, respectively. However, both models fundamentally lack DNA-level prediction capabilities. While CodonMPNN introduced nucleotide-level design, it yielded limited performance (53.59% residue recovery and 28.41% CDS recovery) accompanied by a notable degradation in structural fidelity (TM-score: 0.78, Mean RMSD: 4.80 Å). In contrast, our full EZPro-Fit framework demonstrated superior overall performance, achieving the highest residue recovery of 55.54% (+17.4% relative improvement over ProteinMPNN) and CDS recovery of 32.34% (+13.8% relative improvement over CodonMPNN). Simultaneously, EZPro-Fit maintained robust structural compatibility, yielding a high TM-score of 0.89 and a low mean RMSD of 3.16 Å, alongside natural-like sequence likelihood (Perplexity: 1.42).

Ablation analysis further highlights the synergistic benefit of integrating the ESM2 continuous feedback mechanism. The ablated variant (w/o ESM2) already surpassed baselines in structural preservation (TM-score: 0.91, RMSD: 3.10 Å); however, the introduction of the ESM2 dual-decoder feedback loop (EZPro-Fit) significantly boosted sequence and CDS recovery (improving CDS recovery from 29.23% to 32.34%). By simultaneously co-optimizing amino acid sequence and codon usage within an end-to-end continuous flow, EZPro-Fit directly addresses expression bottlenecks while safeguarding structural integrity, effectively bridging the design-to-expression gap in rational protein engineering.

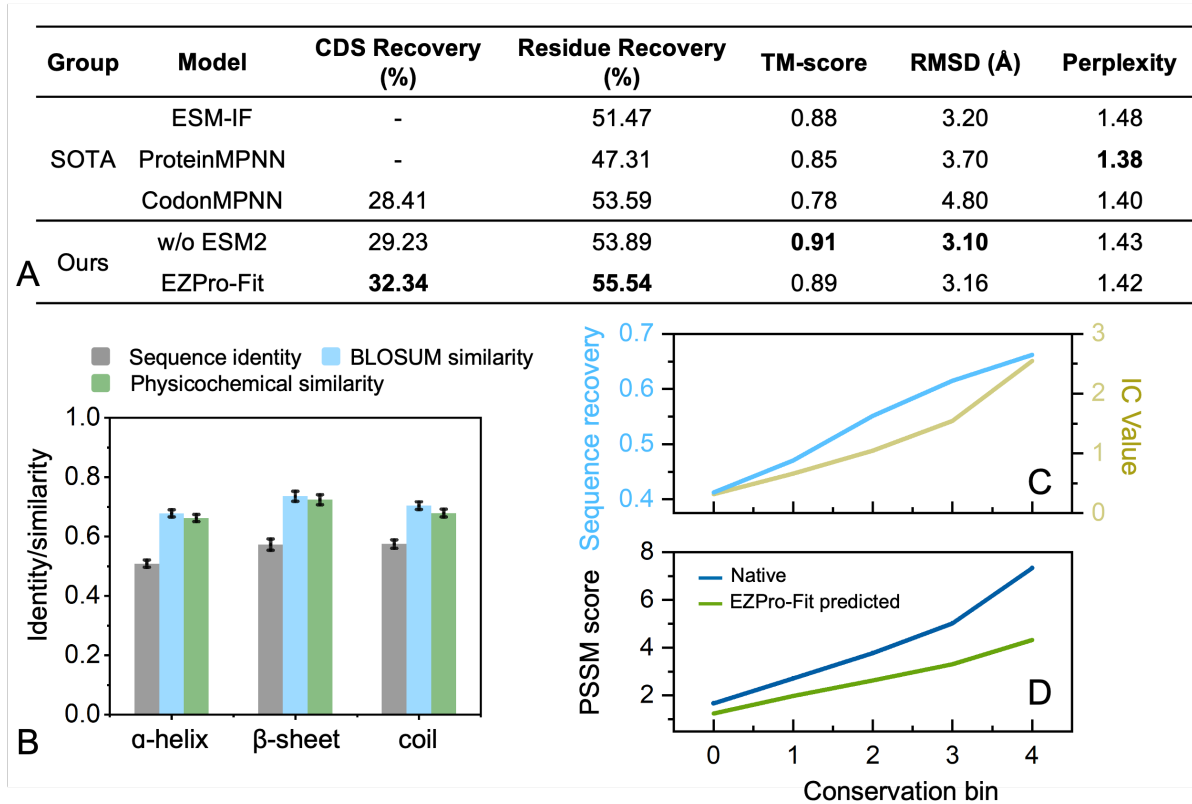


Figure 2: Performance evaluation of EZPro-Fit on sequence recovery, structural fidelity, and evolutionary consistency. (A) Benchmark comparison of EZPro-Fit with representative models. CDS recovery and residue recovery measure nucleotide- and amino acid-level reconstruction accuracy, respectively. TM-score and mean RMSD evaluate structural fidelity, and perplexity measures sequence modeling quality. Bold values indicate the best performance within each metric. (B) Secondary-structure-specific recovery, including sequence identity, BLOSUM similarity, and physicochemical similarity. (C) Relationship between residue recovery and evolutionary conservation. Conservation bins were defined according to information content values, with higher bins indicating more conserved positions. (D) Mean PSSM scores of native and EZPro-Fit-generated sequences across conservation bins.

Sequence recovery across structural contexts

To investigate whether EZPro-Fit preserves local structural constraints, we evaluated sequence recovery across three major secondary structure elements, including α -helices, β -sheets, and coils (Figure 2B). Sequence identity was

comparable between β -sheet and coil regions (57.3% and 57.5%, respectively), whereas α -helices showed relatively lower recovery (50.9%). Beyond exact residue recovery, BLOSUM similarity and physicochemical conservation revealed that β -sheet regions achieved the highest similarity (73.6% and 72.4%, respectively), indicating a greater tendency toward conservative substitutions. Amino acid confusion matrices further revealed structure-dependent substitution patterns (Figure S1). In β -sheets, substitutions were enriched among hydrophobic residues with similar properties, such as valine and isoleucine, whereas α -helices showed more frequent exchanges among residues with similar charge or polarity. Coil regions displayed broader substitution patterns involving polar, charged, and flexible residues, consistent with their greater sequence variability. Together, these results suggest that EZPro-Fit captures secondary structure-associated residue preferences and preserves key physicochemical characteristics, although distinguishing between closely related amino acids remains challenging.

Generalizability across functional categories

Given that enzymes (EC 1–6) represent major targets for protein engineering, we further evaluated whether EZPro-Fit maintains consistent performance across different functional categories (Figure S2). Violin plots showed highly comparable distributions between enzyme and non-enzyme proteins for codon recovery, sequence recovery, and sequence perplexity, with no apparent performance bias toward either category. We further examined model performance across individual Enzyme Commission (EC) classes and observed similar distributions of recovery metrics and perplexity among EC 1–6 enzymes. These results indicate that EZPro-Fit exhibits robust sequence generation performance across diverse functional categories and is not strongly dependent on specific enzymatic classes or functional annotations.

Correlation between sequence recovery and evolutionary conservation

To evaluate whether EZPro-Fit captures evolutionary constraints beyond structural compatibility, we analyzed sequence recovery and position-specific scoring matrix (PSSM) profiles across residues with different levels of evolutionary conservation (Figure 2C–D). Conservation bins were defined based on positional information content (IC), where higher IC values indicate stronger evolutionary constraints and reduced tolerance to amino acid substitutions.

To evaluate whether EZPro-Fit captures evolutionary constraints beyond structural compatibility, we analyzed sequence recovery and position-specific scoring matrix (PSSM) profiles across residues with different levels of evolutionary conservation (Figure 2C–D). PSSM-derived profiles have been widely used to characterize position-specific amino acid preferences and evolutionary constraints in proteins [22], [23]. Conservation bins were defined based on positional information content (IC), where higher IC values indicate stronger evolutionary constraints and reduced tolerance to amino acid substitutions [23], [24], [25]. Sequence recovery provides a complementary measure of the extent to which structure-conditioned sequence design recapitulates native-like amino acid preferences [26].

As shown in Figure 2C, sequence recovery increased progressively with IC, reaching over 65% at highly conserved positions (Bin 4). This trend suggests that EZPro-Fit preferentially preserves residue identities at evolutionarily constrained sites while allowing greater sequence diversity at less conserved positions. We next compared the PSSM profiles of native and EZPro-Fit-designed sequences across the same conservation bins (Figure 2D). PSSM scores, derived from homologous sequence alignments, reflect the evolutionary preference for amino acids at individual positions. Across all conservation levels, EZPro-Fit-generated sequences consistently exhibited lower PSSM scores than native sequences, indicating that there remains room for improvement in reproducing native evolutionary preferences. Nevertheless, the PSSM profiles of designed and native sequences displayed similar trends across conservation bins, particularly at highly conserved positions, suggesting that EZPro-Fit captures the major evolutionary constraints governing residue preferences while only partially recapitulating the sequence characteristics of natural protein families.

Case studies for single-site design and full sequence generation

To demonstrate the practical versatility of EZPro-Fit, we evaluated its performance across two distinct yet complementary engineering paradigms: site-specific refinement and holistic sequence redesign. In industrial enzyme engineering, while *de novo* generation explores novel sequence spaces, targeted mutations often provide higher experimental success rates and immediate production value by preserving the functional essence of a known scaffold. EZPro-Fit addresses both needs by enabling either full-sequence optimization or a “partial-fixing” strategy, in which critical structural motifs are anchored while specific hotspots are prioritized for mutation.

To rigorously assess the predictive accuracy of EZPro-Fit, we conducted a zero-shot evaluation using experimental Deep Mutational Scanning (DMS) datasets, specifically focusing on the lipase EstA from *B. subtilis* (P37957) and

human myoglobin (P02144) (Figure 3), where thermostability served as the primary fitness metric. By analyzing the prioritized residue distributions recommended by EZPro-Fit, we quantified the proportion of variants effectively enhancing thermostability and benchmarked these results against the physics-based modeling tools Rosetta and FoldX. Our results demonstrate a significant enrichment of stabilizing mutations through EZPro-Fit: the model achieved a precision of approximately 0.30 across the global sequence. Notably, in solvent-exposed surface regions, the predictive precision increased to over 0.41, underscoring the model's unique capability to capture "molecular armor" effects, such as optimized surface salt-bridge networks. Comparative analysis reveals that EZPro-Fit significantly outperforms Rosetta in surface regions while maintaining competitive performance with FoldX. This robust capacity for targeted local refinement allows for the compression of vast mutational landscapes into a handful of high-probability candidates, effectively bridging the gap between computational selection and laboratory screening with high industrial production value.

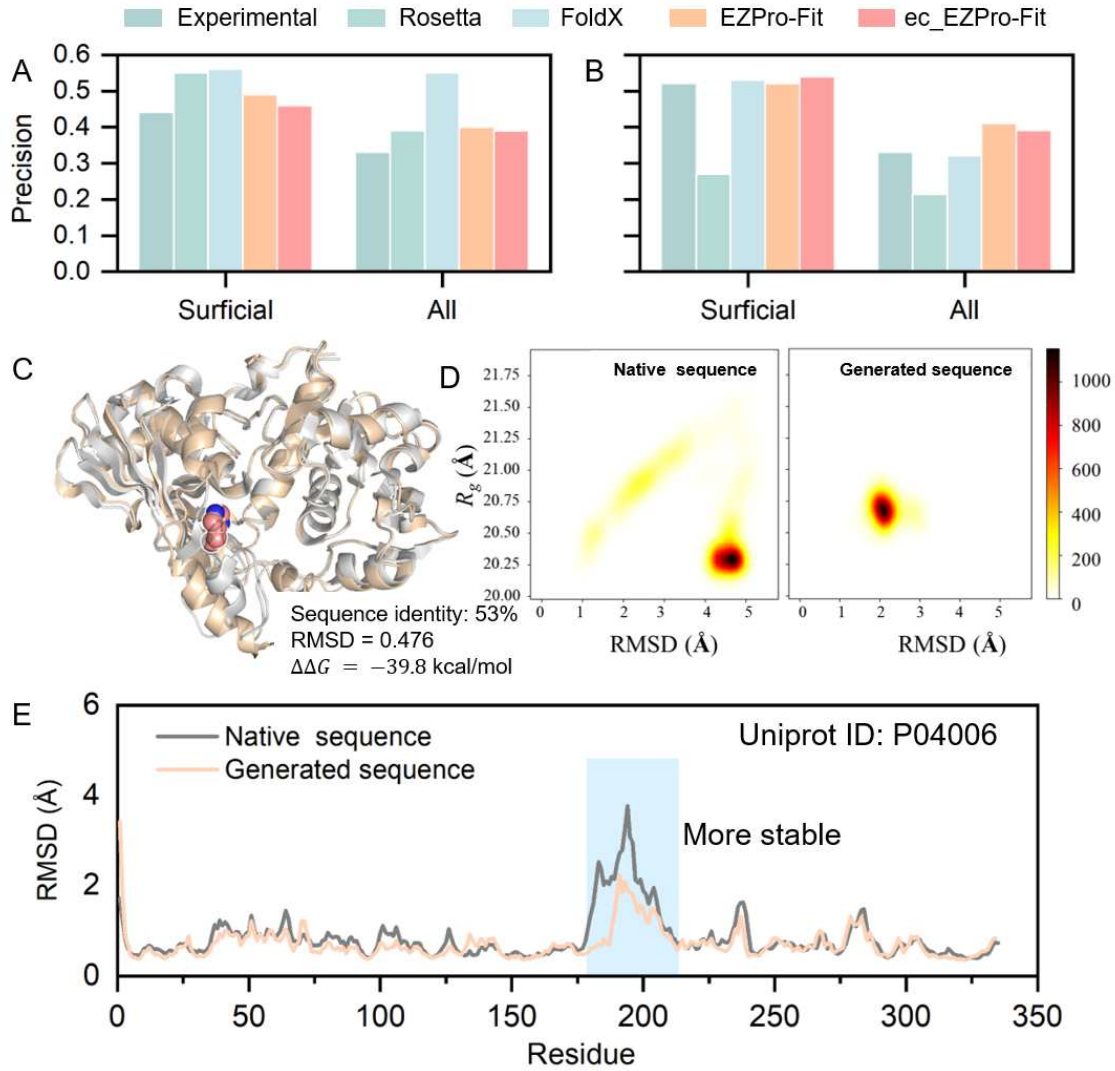


Figure 3: Case studies demonstrating the capabilities of EZPro-Fit in single-mutation and full-sequence design for enhancing thermostability. (A, B) Performance comparison of thermostability prediction methods (Rosetta, FoldX, EZPro-Fit, and ec_EZPro-Fit) evaluated across all residues and surface-exposed residues for two proteins (UniProt IDs: P37957 and P02144), respectively. (C) Structural superposition of wild-type (WT, UniProt ID: P04406) and EZPro-Fit-optimized GAPDH, highlighting key evaluation metrics: 53% sequence identity, 0.476 Å backbone RMSD, and $\Delta\Delta G = -39.8$ kcal/mol. The sphere indicates H179, a catalytic histidine that activates the thiol group and remains invariant after full sequence generation. (D) MD simulation heatmaps of RMSD versus radius of gyration (R_g), demonstrating that the optimized variant adopts a more compact and stable conformational state. (E) Residue-wise RMSF profiles over a 100 ns simulation for a protein (UniProt ID: P04406), showing reduced flexibility in the optimized variant, which is consistent with enhanced thermostability.

Furthermore, to evaluate the global optimization capacity of EZPro-Fit, we redesigned the mesophilic protein P04406 for high-temperature industrial applications. The model introduced 94 substitutions (53% sequence

identity) while maintaining exceptional structural fidelity (RMSD = 0.476 Å) and achieving substantial predicted stabilization ($\Delta\Delta G = -39.8$ kcal/mol). All-atom molecular dynamics (MD) simulations (100 ns) were performed to validate the enhanced thermostability (Figure 3). While the wild-type protein underwent progressive unfolding-evidenced by the expansion of the radius of gyration (R_g) from 18.3 Å to 22.7 Å and an RMSD increase to 4.2 Å-the EZPro-Fit-designed variant remained structurally intact. The designed protein's R_g stabilized at 18.5 ± 0.3 Å with an RMSD plateau of 1.8 Å. RMSF analysis further revealed a 62% reduction in global flexibility, with uniform rigidification observed across catalytic sites and binding pockets. These results confirm that EZPro-Fit effectively internalizes structure-stability relationships to generate sequences with genuine thermal robustness, bridging the gap between computational design and performance under extreme conditions.

Host adaptation of EZPro-Fit through fine-tuning on highly expressed *E. coli* proteins

While pretrained EZPro-Fit effectively captures thermostability-conferring nucleotide token space patterns, thermophilic proteins frequently exhibit suboptimal expression in *E. coli* due to fundamental mismatches in nucleotide token space, tRNA abundance, and translation kinetics between thermophiles and mesophilic hosts. We address this barrier through organism-specific fine-tuning on high-expression *E. coli* genes, adapting nucleotide token selection for maximal heterologous expression while preserving learned thermostability principles (Figure 4A). This two-stage strategy enables simultaneous optimization of structural fidelity and host-adaptive translation efficiency, bridging the design-expression gap in industrial enzyme engineering. Figure 4B summarizes expression-related metrics for sequences generated by pretrained versus fine-tuned EZPro-Fit models on a held-out test set.

Based on the comparison of model performance metrics between EZPro-Fit (pre-fine-tuning) and Finetune (post-fine-tuning) models on the test dataset, the following results were observed: the fine-tuned model demonstrated superior performance across multiple evaluation metrics. For the all score, the reported values ranged from approximately 41.4 to 68.3 across the evaluated cases. The CSI (Codon Similarity Index) showed a clear pattern of enhancement, with the fine-tuned model achieving values of approximately 0.819–0.866, compared with 0.751–0.775 for EZPro-Fit, indicating improved CDS usage optimization. The CAI (Codon Adaptation Index) exhibited a similarly pronounced improvement, with the fine-tuned model attaining values of approximately 0.817–0.870 compared with 0.754–0.782 for EZPro-Fit. This increase suggests that the fine-tuned model generates DNA sequences with better adaptation to the host organism's codon usage bias. However, the %MINMAX metric showed a slight decrease after fine-tuning, with values of approximately 76.5–77.8 compared with 79.2–81.8 for EZPro-Fit.

Engineered cel5E via ec_EZPro-Fit

Previous evaluation established that EZPro-Fit accurately maps thermostable protein sequences from their structures. To validate its utility for real-world enzyme engineering, we applied ec_EZPro-Fit to discover novel endoglucanase variants with improved thermostability without compromising catalytic activity (Figure 5A).

As the primary component of plant cell walls and the world's most vital carbon source, cellulose comprises long chains of β -1,4-linked D-glucose molecules [27], driving a surge of recent interest in its bioconversion into renewable biofuels [28]. Within this process, endoglucanases (such as cel5E) specifically cleave internal β -1,4-glucosidic bonds to generate novel chain ends [29] (Figure 5 B), while thermostable variants offer distinct industrial advantages, including reduced enzyme replacement costs, enhanced specific activity, and greater operational flexibility. To address these limitations, we first utilized the three-dimensional structure of the wild-type enzyme (UniProt ID: P16218) as a spatial template [30]. A structure-guided generative framework, ec_EZPro-Fit, was then deployed for sequence optimization, yielding a curated set of mutation candidates. To narrow this design space and minimize false positives, these candidates underwent a rigorous dual-filter consensus ranking based on free energy change ($\Delta\Delta G$) computations and position-specific scoring matrices (PSSM), thereby enriching for variants with superior thermodynamic stability and evolutionary conservation. Concurrently, top-ranking variants were advanced to wet-lab validation, where heterologous expression, purification, and enzymatic characterization were performed to confirm their enhanced properties (Figure 5A).

Experimental characterization of the prioritized cel5E variants revealed a broad spectrum of mutational effects on thermostability and catalytic activity (Figure 5C,D), indicating that the AI-guided design strategy effectively sampled diverse regions of the enzyme fitness landscape. Among the thirteen experimentally evaluated variants, most substitutions either improved one property while maintaining the other or produced only marginal effects, whereas two variants, N291D and D259E, exhibited simultaneous enhancements in both thermostability and catalytic performance. Specifically, N291D increased the melting temperature (T_m) from 63.16 °C for the wild-type enzyme to 69.58 °C while enhancing the relative activity by approximately 35%. Likewise, D259E increased the

T_m to 72.35 °C while retaining a 17% improvement in catalytic activity. These two variants therefore represent the most balanced improvements among all candidates, demonstrating that ec_EZPro-Fit can identify mutations that simultaneously enhance structural robustness and enzymatic function.

In contrast, the double mutant DS (D218E/S262I) achieved the highest thermal stability of all tested variants, with a T_m of 78.12 °C, but suffered a pronounced loss of catalytic activity, retaining only approximately one-third of the wild-type activity. This observation highlights the classical stability–activity trade-off frequently encountered in protein engineering, suggesting that excessive rigidification may impair the conformational flexibility required for efficient catalysis. Similarly, variants such as V227I and VT (V220I_T299E) substantially increased thermostability but showed little or no improvement in catalytic activity, whereas S152E, A265Y, S84A, and G88D primarily enhanced activity while producing only modest changes in thermal stability. Other substitutions, including H205F and N108Y, exhibited only minor effects on both properties.

Collectively, these results demonstrate that the proposed AI-guided design framework generated a diverse set of functional variants spanning stability-biased, activity-biased, and dual-beneficial mutations. More importantly, the successful identification of D259E and N291D indicates that ec_EZPro-Fit is capable of enriching mutations that simultaneously improve thermostability and catalytic performance, rather than merely maximizing protein stability, thereby providing experimentally validated candidates with enhanced industrial applicability.

CONCLUSIONS

EZPro-Fit establishes an end-to-end generative paradigm that maps protein tertiary structures directly to expression-optimized coding sequences via a unified neural network. Unlike conventional “decoupled” strategies, which first design amino acid sequences and then perform post hoc codon optimization, EZPro-Fit simultaneously accounts for structural constraints and translation-related properties. At its architectural core, the framework integrates an MPNN with flow-matching generative modeling, utilizing taxonomic identifiers and ESM2-derived perplexity regularization to ensure biological plausibility and sequence “naturalness”. To meet the rigorous demands of industrial biocatalysis, a two-stage transfer learning strategy was employed: pre-training on a diverse thermophilic proteome to internalize universal thermostability principles, followed by organism-specific fine-tuning on *E. coli* genomic data to capture host-dependent DNA biases. This study establishes a closed-loop engineering pipeline integrating computational rational design and wet-lab validation. Using the three-dimensional structure of wild-type cel5E as the template, two variants (N291D and D259E) are identified to simultaneously and synergistically improve both thermostability and catalytic activity. The N291D mutation located at the periphery of the catalytic pocket remodels the pocket electrostatic environment and hydrogen-bond network, elevating the melting temperature to 69.58 °C and boosting relative enzymatic activity by 35%. The D259E substitution resides on the protein surface far from the active site; the extra methylene group of glutamic acid reinforces local surface rigidity, yielding a temperature of approximately 72.35 °C with a 17% activity increment without disrupting the core catalytic architecture.

METHODS

ONLINE METHODS

Multi-modal data curation and preprocessing

First, a high-fidelity dataset of thermophilic bacterial sequences was curated by integrating physiological metadata from BacDive [31] with genomic and proteomic annotations from UniProt [32]. To bridge sequence and structural dimensions, 3D coordinates predicted by AlphaFold2 [33] were cross-referenced with the AlphaFold Protein Structure Database [34]. Next, candidate species were strictly filtered based on an optimal growth temperature ($T_{opt} \geq 40^\circ\text{C}$), with phylogeny restricted to the bacterial domain. Finally, only entries with unambiguous coding sequence-to-protein mappings were retained.

To mitigate homology-induced bias and ensure dataset representativeness, we first performed sequence dereplication using CD-HIT (v4.8.1) [35] with a 40% amino acid identity threshold ($-c \ 0.4 \ -n \ 2$). Selecting the longest representative from each cluster yielded a non-redundant core dataset, minimizing computational overhead for downstream structural analysis. To account for structural constraints on translational kinetics, we subsequently evaluated the structural fidelity of the representative sequences. Utilizing the Biopython framework [36], global mean Predicted Local Distance Difference Test (pLDDT) scores were calculated from AlphaFold2-predicted PDB files. To ensure high-confidence modeling, the dataset was restricted to representatives exhibiting a global average

pLDDT > 90 (denoting "very high" confidence), forming the final high-quality training corpus optimized for thermophilic-specific host-adapted nucleotide bias analysis.

Additionally, the fine-tuning dataset comprised high-expression proteins from *E. coli* K-12 strain MG1655, selected to represent optimal codon usage patterns for this industrial workhorse organism. High-expression genes were defined by stringent criteria: transcripts per million (TPM) values exceeding the 90th percentile (TPM > 500) across biological replicates and three-dimensional structures predicted by AlphaFold2 with quality filtering requiring average pLDDT > 85 to ensure ample and reliable structural features for the MPNN encoder.

Protein structure encoding via MPNN

To accurately capture the 3D geometric context of the target protein, our model employs the encoder architecture derived from ProteinMPNN. In this framework, the protein backbone is represented as a spatial graph, where each amino acid acts as a node, and the distances between C_α atoms define the edges. Through multiple layers of message passing, the network updates the hidden state of each node by aggregating geometric and relational features from its k -nearest neighbors. This process generates highly robust, rotationally and translationally invariant structure embeddings, providing a precise physical blueprint for the subsequent sequence design.

Sequence generation via Flow-matching

To formalize the non-autoregressive generation process, we employ the conditional flow-matching framework. Let x_1 represent the target variable (the native thermostable sequence) sampled from the empirical data distribution $q(x_1)$, and x_0 denote the initial noise state sampled from a tractable prior distribution $p_0(x_0)$ (e.g., standard Gaussian $\mathcal{N}(0, I)$). The continuous time variable is defined as $t \in [0, 1]$.

The generative process is governed by an Ordinary Differential Equation (ODE), which constructs a deterministic mapping from p_0 to $q(x_1)$ via a time-dependent vector field $v_t(x)$. To achieve structure-guided sequence design, we introduce a conditional vector field $v_\theta(x_t, t, c)$ parameterized by a neural network with weights θ . Here, the condition c represents the concatenated embeddings derived from the MPNN structure encoder and the specific TAXON ID.

Following the Optimal Transport (OT) probability path formulation, the intermediate state x_t at time t is linearly interpolated between the noise x_0 and the target data x_1 :

$$x_t = tx_1 + (1 - t)x_0 \quad (1)$$

Consequently, the target marginal vector field $u_t(x_t|x_1)$ guiding this transformation is simply a constant velocity:

$$u_t(x_t|x_1) = x_1 - x_0 \quad (2)$$

The neural network is then trained to approximate this target vector field. The conditional flow matching objective function is defined as the expected mean squared error between the predicted vector field and the target vector field:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim U[0,1], q(x_1), p(x_0)} |v_\theta(x_t, t, c) - (x_1 - x_0)|^2 \quad (3)$$

During the inference phase, given a novel 3D protein backbone (encoded as condition c) and an initial noise vector $x_0 \sim \mathcal{N}(0, I)$, the final engineered sequence x_1 is generated by solving the ODE using a numerical solver (e.g., Euler or Runge-Kutta methods):

$$x_1 = x_0 + \int_0^1 v_\theta(x_t, t, c) dt \quad (4)$$

This mathematical formulation ensures that the generation of sequences is globally conditioned on the spatial geometry provided by the MPNN, circumventing the locality biases inherent in autoregressive decoding.

Multi-objective constraints for model fine-tuning

During fine-tuning, three auxiliary objective functions were introduced to preserve translation fidelity, adapt the generated coding sequences to the codon usage of *Escherichia coli*, and retain thermophile-associated amino acid characteristics.

First, an amino acid consistency loss was used to ensure that the generated codons translated into the target amino acid sequence. A fixed codon-to-amino-acid mapping matrix $M \in \{0, 1\}^{K_c \times K_a}$ was constructed, where K_c and K_a denote the numbers of codon and amino acid categories, respectively. Given the predicted codon probability distribution $P_\theta^{\text{codon}}(i)$ at position i , the corresponding amino acid probability distribution was calculated as

$$P_\theta^{\text{AA}}(i) = P_\theta^{\text{codon}}(i)M. \quad (5)$$

The amino acid consistency loss was defined as

$$\mathcal{L}_{\text{AA}} = -\frac{1}{\sum_{i=1}^L m_i} \sum_{i=1}^L m_i \log P_\theta^{\text{AA}}(i, a_i^*), \quad (6)$$

where a_i^* is the target amino acid and m_i indicates a valid sequence position. This term was assigned a weight of $\lambda_{\text{AA}} = 15.0$.

Second, an *E. coli* codon-frequency regularization term was introduced to encourage host-adapted codon usage. Let $q_{\text{Ec}}(c)$ denote the normalized empirical frequency of codon c in the predefined *E. coli* codon-usage prior. The loss was computed as

$$\mathcal{L}_{\text{Ec}} = \frac{1}{\sum_{i=1}^L m_i} \sum_{i=1}^L m_i D_{\text{KL}} [q_{\text{Ec}}(c) \| P_\theta^{\text{codon}}(i, c)]. \quad (7)$$

This objective regularized the predicted codon distribution toward the global codon-usage characteristics of *E. coli*, with a maximum weight of $\lambda_{\text{Ec}}^{\text{max}} = 0.3$.

Third, a thermophile-associated amino acid preference loss was applied using a predefined residue-weight vector $\mathbf{w}^{\text{stab}} \in \mathbb{R}^{K_a}$. Positive weights were assigned to residues favored by the thermophile-associated composition prior, whereas negative weights were assigned to residues considered unfavorable. The expected residue-preference loss was defined as

$$\mathcal{L}_{\text{stab}} = -\frac{1}{\sum_{i=1}^L m_i} \sum_{i=1}^L m_i \sum_{a=1}^{K_a} P_\theta^{\text{AA}}(i, a) w_a^{\text{stab}}. \quad (8)$$

Minimizing this term increased the probability of residues with positive preference weights and suppressed residues with negative weights. Its maximum coefficient was set to $\lambda_{\text{stab}}^{\text{max}} = 0.2$. This term represents an amino acid composition regularizer rather than a direct predictor of melting temperature or folding free-energy changes.

The three auxiliary constraints were combined with the primary sequence-generation loss as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda_{\text{AA}} \mathcal{L}_{\text{AA}} + \lambda_{\text{Ec}}(E) \mathcal{L}_{\text{Ec}} + \lambda_{\text{stab}}(E) \mathcal{L}_{\text{stab}}, \quad (9)$$

where the weights of the *E. coli* codon-frequency and thermophile-associated amino acid preference losses were gradually increased during fine-tuning.

Molecular dynamics simulation

Duplicate 100-ns all-atom MD simulations were executed for wild-type and predicted sequences via AMBER 22 [37] in explicit solvent environments. System construction utilized Tleap [38] with ff14SB parameterization [39]. Electroneutrality was maintained through Na^+/Cl^- addition, with TIP3P water molecules filling rectangular simulation boxes extending at least 10 Å beyond the protein. Multi-stage energy minimization preceded a 10,000-step heating protocol (0→300 K), followed by a 200-ps NPT equilibration (300 K, 1 atm). Production protocols featured 2-fs integration, 8-Å nonbonded truncation, and periodic boundary treatment. Electrostatic interactions beyond the cutoff were evaluated via PME summation [40], with SHAKE-constrained [41] hydrogen bonds. Trajectories were recorded at 4-ps resolution.

Codon usage bias metrics

To quantitatively evaluate codon usage bias and translational adaptation, three complementary metrics were employed: the Codon Adaptation Index (CAI), the Codon Similarity Index (CSI), and the %MinMax profile [42].

These metrics characterize codon preference from global and local perspectives and have been widely adopted in codon optimization, synthetic biology, and recent protein language model studies.

CAI [43] is the most widely used metric for evaluating the adaptation of a coding sequence to the codon usage preference of highly expressed genes in a host organism. It quantifies the geometric mean of the relative adaptiveness of all codons in a gene. The CAI ranges from 0 to 1, with higher values indicating stronger adaptation to the host translational machinery and generally higher expected translation efficiency.

For a coding sequence containing L codons, the CAI is defined as

$$\text{CAI} = \exp \left(\frac{1}{L} \sum_{k=1}^L \ln w_k \right) = \left(\prod_{k=1}^L w_k \right)^{1/L}, \quad (10)$$

where the relative adaptiveness of codon i is

$$w_i = \frac{f_i}{\max(f_j)}, \quad (11)$$

with f_i denoting the frequency of codon i in a reference set of highly expressed genes, and $\max(f_j)$ representing the maximum frequency among synonymous codons encoding the same amino acid.

CSI extends the concept of CAI by replacing the reference set of highly expressed genes with genome-wide codon usage frequencies. This genome-level normalization provides a more general measure of codon preference and reduces the bias introduced by selecting a specific subset of highly expressed genes.

CSI is computed as

$$\text{CSI} = \exp \left(\frac{1}{L} \sum_{k=1}^L \ln w_{k,\text{genome}} \right), \quad (12)$$

where

$$w_{k,\text{genome}} = \frac{x_{ij}}{x_{i,\max}}, \quad (13)$$

in which x_{ij} denotes the genome-wide frequency of the j -th synonymous codon encoding amino acid i , and $x_{i,\max}$ is the maximum frequency among all synonymous codons for amino acid i .

Unlike CAI and CSI, which evaluate the overall codon adaptation of an entire coding sequence, the %MinMax metric characterizes local variations in codon usage using a sliding-window strategy [44]. It is particularly useful for identifying clusters of rare or non-optimal codons that may introduce translational pauses and thereby facilitate co-translational protein folding.

For a sliding window of size w , the normalized codon usage score of the i -th window is calculated as

$$\% \text{MinMax}_i = \frac{\bar{x}_i - x_{\min,i}}{x_{\max,i} - x_{\min,i}} \times 100\%, \quad (14)$$

where $\bar{x}_i$ is the average codon usage frequency within the window, while $x_{\min,i}$ and $x_{\max,i}$ denote the minimum and maximum synonymous codon frequencies, respectively.

Alternatively, many recent implementations compute the normalized average codon preference as

$$\% \text{MinMax}_i = \frac{1}{w} \sum_{k \in \text{window}_i} \frac{f_k - f_{\min}(AA_k)}{f_{\max}(AA_k) - f_{\min}(AA_k)}, \quad (15)$$

where f_k is the frequency of codon k , and $f_{\min}(AA_k)$ and $f_{\max}(AA_k)$ are the minimum and maximum synonymous codon frequencies for the corresponding amino acid. The complete %MinMax profile is represented as a sequence of sliding-window values, allowing local codon usage patterns to be analyzed and compared across genes.

Experimental validation of the cel5E mutant designed by finetuned EZPro-Fit

Site-directed mutagenesis and vector construction

Specific mutant strains were engineered via site-directed mutagenesis using primers designed to incorporate target substitutions (Tables S1-S2). The pET-28a(+) plasmid served as the template for PCR amplification, which was performed using Phanta Flash Master Mix (Vazyme, China). The thermal cycling protocol consisted of an initial denaturation at 98°C for 30 s, followed by 35 cycles of denaturation at 98°C for 10 s, annealing at 72°C for 35 s, and extension at 72°C for 5 min.

Following amplification, PCR products were verified by agarose gel electrophoresis and recovered using a DNA purification kit. The purified amplicons were assembled into the expression vector via homologous recombination using the ClonExpress II One Step Cloning Kit (Vazyme). The resulting recombinant plasmids were transformed into chemically competent *E. coli* BL21 (DE3) cells. Positive transformants were screened and validated through Sanger sequencing to confirm the successful introduction of the desired mutations.

Protein expression and purification

The wild-type (WT) and mutant plasmids were transformed into *E. coli* BL21 (DE3) via the heat-shock method (42 °C for 45 s). After a 1 h recovery in LB medium (37 °C, 220 rpm), the cells were harvested and plated onto LB agar supplemented with kanamycin. Single colonies were inoculated into LB broth and cultured at 37°C until they reached the exponential growth phase.

Protein expression was induced by adding isopropyl β -D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 mM, followed by incubation at 16°C and 170 rpm for 16 h. The biomass was harvested by centrifugation at 10,000 g for 5 min at 4°C. The cell pellets were resuspended in lysis buffer (50 mM Tris-HCl, 150 mM NaCl, pH 8.0) and disrupted via ultrasonication (400 W, 3 s pulse/4 s interval, total duration 30 min). The lysate was clarified by centrifugation at 10,000 g for 20 min at 4°C, and the resulting supernatant was collected as the crude enzyme extract.

Purification was conducted using Immobilized Metal Affinity Chromatography (IMAC) with a Ni-NTA column. The purity and molecular weight of the purified proteins were assessed via SDS-PAGE analysis.

Protein stability and thermal shift assay

The thermostability of the protein was evaluated using Differential Scanning Fluorimetry (DSF). Protein samples were prepared at a final concentration of 0.25 μ M in a buffer containing 100 mM NaCl and 50 mM of the respective buffer pair at the target pH. To monitor protein unfolding, the fluorescent dye SYPRO Orange (Sigma-Aldrich) was utilized. Specifically, 20 * SYPRO Orange was added to 20 μ L of the protein solution, followed by thorough mixing and incubation at room temperature for 1 h in the dark to ensure stable dye-protein interaction. All samples were prepared and measured in triplicate to ensure reproducibility. Thermal denaturation was performed using a quantitative real-time PCR (qPCR) system. The fluorescence intensity was monitored using excitation and emission wavelengths of 470 nm and 570 nm, respectively. The temperature was increased from 25°C to 95°C at a constant ramp rate of 1°C/30 s.

DATA AND CODE AVAILABILITY

The complete source code implementing the EZPro-Fit framework is openly available on GitHub to enable reproducibility and methodological extension. The repository can be accessed at: https://github.com/Ran-Xu1213/EZPro_Fit.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization: J.G., L.Z., D.L.; Methodology: R.X., LY.Z., R.P., Z.Z.; Computational analysis: R.X.; Experimental validation: LY.Z., R.P., Z.Z.; Data analysis: R.X., L.Z.; Writing-original draft: R.X., LY.Z., J.G.; Writing-review and editing: R.X., LZ.Z., H.D., L.W., J.G.; Supervision: D.L., LZ.Z., J.G. All authors discussed the results and approved the final manuscript.

ACKNOWLEDGMENTS

This research is supported by the Science and Technology Development Fund of Macau (0004/2025/RIA1) and Macao Polytechnic University (RP/CAI-01/2023).

REFERENCES

- [1] M. S. Mesbahuddin, A. Ganesan, and S. Kalyaanamoorthy, "Engineering stable carbonic anhydrases for co2 capture: A critical review," *Protein Engineering, Design and Selection*, vol. 34, gza021, 2021.
- [2] S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, and U. T. Bornscheuer, "Biocatalysis: Enzymatic synthesis for industrial applications," *Angewandte Chemie International Edition*, vol. 60, no. 1, pp. 88–119, 2021.
- [3] S. Shan et al., "Deep learning guided optimization of human antibody against sars-cov-2 variants with broad neutralization," *Proceedings of the National Academy of Sciences*, vol. 119, no. 11, e2122954119, 2022.
- [4] J. Sui, R. Xu, H. Sun, H. Duan, L. Zheng, and J. Guo, "Ezpro-multi: Contrastive learning-enhanced multi-property prediction for enzyme engineering," *Journal of Chemical Theory and Computation*, 2026.
- [5] C. Hsu et al., "Learning inverse folding from millions of predicted structures," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 8946–8970.
- [6] J. Dauparas et al., "Robust deep learning-based protein sequence design using proteinmpnn," *Science*, vol. 378, no. 6615, pp. 49–56, 2022. doi: 10.1126/science.add2187.
- [7] J. L. Watson et al., "De novo design of protein structure and function with RFdiffusion," *Nature*, vol. 620, no. 7976, pp. 1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.
- [8] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, "Coding-sequence determinants of gene expression in *Escherichia coli*," *Science*, vol. 324, no. 5924, pp. 255–258, 2009. doi: 10.1126/science.1170160.
- [9] D. B. Goodman, G. M. Church, and S. Kosuri, "Causes and effects of n-terminal codon bias in bacterial genes," *Science*, vol. 342, no. 6157, pp. 475–479, 2013. doi: 10.1126/science.1241934.
- [10] G. Boël et al., "Codon influence on protein expression in *E. coli* correlates with mrna levels," *Nature*, vol. 529, no. 7586, pp. 358–363, 2016. doi: 10.1038/nature16509.
- [11] C.-H. Yu et al., "Codon usage influences the local rate of translation elongation to regulate co-translational protein folding," *Molecular Cell*, vol. 59, no. 5, pp. 744–754, 2015.
- [12] F. Zhao, C.-H. Yu, and Y. Liu, "Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells," *Nucleic Acids Research*, vol. 45, no. 14, pp. 8484–8492, 2017. doi: 10.1093/nar/gkx501.
- [13] C. Outeiral et al., "Codon language embeddings provide strong signals for protein engineering," *Nature Machine Intelligence*, vol. 6, 2024. doi: 10.1038/s42256-024-00791-0.
- [14] R. Guerois, J. E. Nielsen, and L. Serrano, "Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations," *Journal of Molecular Biology*, vol. 320, no. 2, pp. 369–387, 2002. doi: 10.1016/S0022-2836(02)00442-4.
- [15] E. H. Kellogg, A. Leaver-Fay, and D. Baker, "Role of conformational sampling in computing mutation-induced changes in protein structure and stability," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 3, pp. 830–838, 2011. doi: 10.1002/prot.22921.
- [16] M. Musil et al., "Fireprot: Web server for automated design of thermostable proteins," *Nucleic Acids Research*, vol. 45, no. W1, W393–W399, 2017. doi: 10.1093/nar/gkx285.
- [17] H. Dieckhaus, M. Brocidiaco, N. Z. Randolph, and B. Kuhlman, "Transfer learning to leverage larger datasets for improved prediction of protein stability changes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, no. 6, e2314853121, 2024. doi: 10.1073/pnas.2314853121.
- [18] A. Madani et al., "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, vol. 41, no. 8, pp. 1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.
- [19] S. D. Stimple, M. D. Smith, and P. M. Tessier, "Directed evolution methods for overcoming trade-offs between protein activity and stability," *AIChE Journal*, vol. 66, no. 3, e16814, 2020. doi: 10.1002/aic.16814.
- [20] S. R. Miller, "An appraisal of the enzyme stability–activity trade-off," *Evolution*, vol. 71, no. 7, pp. 1876–1887, 2017. doi: 10.1111/evo.13275.
- [21] J. B. Plotkin and G. Kudla, "Synonymous but not the same: The causes and consequences of codon bias," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 32–42, 2011. doi: 10.1038/nrg2899.
- [22] S. F. Altschul et al., "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997. doi: 10.1093/nar/25.17.3389.
- [23] J. Pei and N. V. Grishin, "Al2co: Calculation of positional conservation in a protein sequence alignment," *Bioinformatics*, vol. 17, no. 8, pp. 700–712, 2001. doi: 10.1093/bioinformatics/17.8.700.
- [24] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, 2007. doi: 10.1093/bioinformatics/btm270.

- [25] H. Ashkenazy et al., “Consurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules,” *Nucleic Acids Research*, vol. 44, no. W1, W344–W350, 2016. doi: 10.1093/nar/gkw408.
- [26] J. Dauparas et al., “Robust deep learning-based protein sequence design using ProteinMPNN,” *Science*, vol. 378, no. 6615, pp. 49–56, 2022. doi: 10.1126/science.add2187.
- [27] L. R. Lynd, P. J. Weimer, W. H. Van Zyl, and I. S. Pretorius, “Microbial cellulose utilization: Fundamentals and biotechnology,” *Microbiology and molecular biology reviews*, vol. 66, no. 3, pp. 506–577, 2002.
- [28] D. B. Wilson, “The first evidence that a single cellulase can be essential for cellulose degradation in a cellulolytic microorganism,” *Molecular microbiology*, vol. 74, no. 6, pp. 1287–1288, 2009.
- [29] S. Ichikawa, M. Yoshida, S. Karita, M. Kondo, and M. Goto, “Carbohydrate-binding modules influence substrate specificity of an endoglucanase from clostridium thermocellum,” *Bioscience, biotechnology, and biochemistry*, vol. 80, no. 1, pp. 188–192, 2016.
- [30] R. Zamani et al., “Enhancing the thermostability of cellulase from clostridium thermocellum via salt bridge interactions,” *Biotechnology and Bioprocess Engineering*, vol. 28, no. 4, pp. 684–694, 2023.
- [31] L. C. Reimer et al., “Bacdiv in 2022: The bacterial diversity metadatabase in the digital age,” *Nucleic Acids Research*, vol. 50, no. D1, pp. D741–D746, 2022.
- [32] U. Consortium, “Uniprot: The universal protein knowledgebase in 2023,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D523–D531, 2023.
- [33] J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. doi: 10.1038/s41586-021-03819-2.
- [34] M. Varadi et al., “Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models,” *Nucleic acids research*, vol. 50, no. D1, pp. D439–D444, 2022.
- [35] L. Fu et al., “Cd-hit: Accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.
- [36] P. J. Cock et al., “Biopython: Freely available python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.
- [37] V. Cruzeiro et al., “Amber 2022,” *University of California, San Francisco*, 2022.
- [38] D. A. Case et al., “The amber biomolecular simulation programs,” *Journal of computational chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.
- [39] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, “Ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb,” *Journal of chemical theory and computation*, vol. 11, no. 8, pp. 3696–3713, 2015.
- [40] T. Darden, D. York, L. Pedersen, et al., “Particle mesh ewald: An n log (n) method for ewald sums in large systems,” *Journal of chemical physics*, vol. 98, pp. 10 089–10 089, 1993.
- [41] C. R. JP, “Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes,” *Journal of Computational Physics*, vol. 23, p. 15, 1977.
- [42] A. Roth, M. Anisimova, and G. M. Cannarozzi, “Measuring codon usage bias,” *Codon evolution: mechanisms and models*, pp. 189–217, 2012.
- [43] A. Carbone, A. Zinovyev, and F. Képes, “Codon adaptation index as a measure of dominating codon bias,” *Bioinformatics*, vol. 19, no. 16, pp. 2005–2015, 2003.
- [44] T. F. Clarke IV and P. L. Clark, “Rare codons cluster,” *PloS one*, vol. 3, no. 10, e3412, 2008.

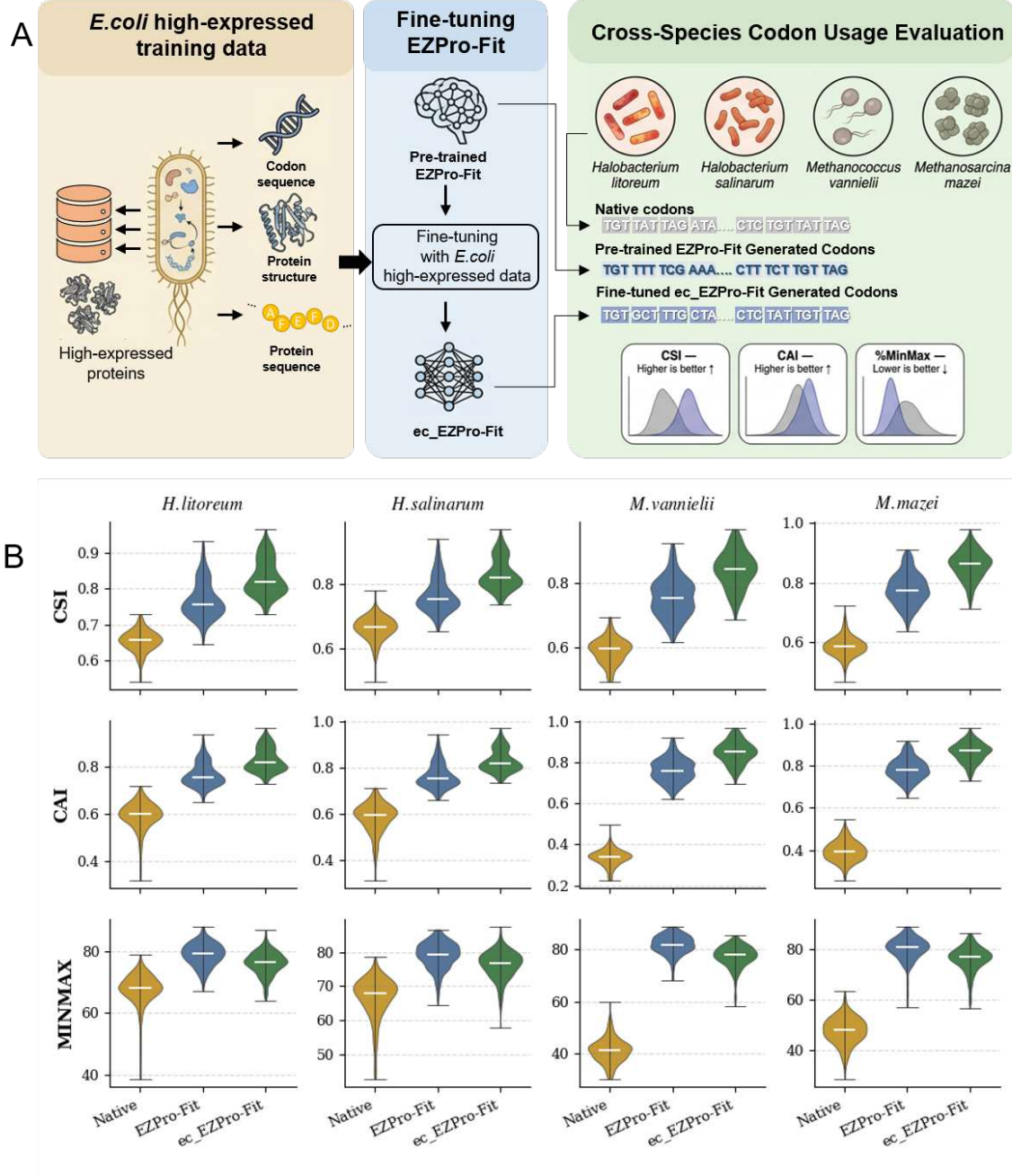


Figure 4: Fine-tuning EZPro-Fit with highly expressed *E. coli* genes for host-adapted enzyme design. (A) Workflow for fine-tuning the pre-trained EZPro-Fit model using coding sequences from highly expressed *E. coli* genes. (B) Comparison of codon adaptability for heterologous expression in *E. coli* between native sequences and sequences generated by the base and fine-tuned EZPro-Fit models across four test datasets: *Halobacterium litoreum*, *Halobacterium salinarum*, *Methanococcus vannielii*, and *Methanosarcina mazel*.

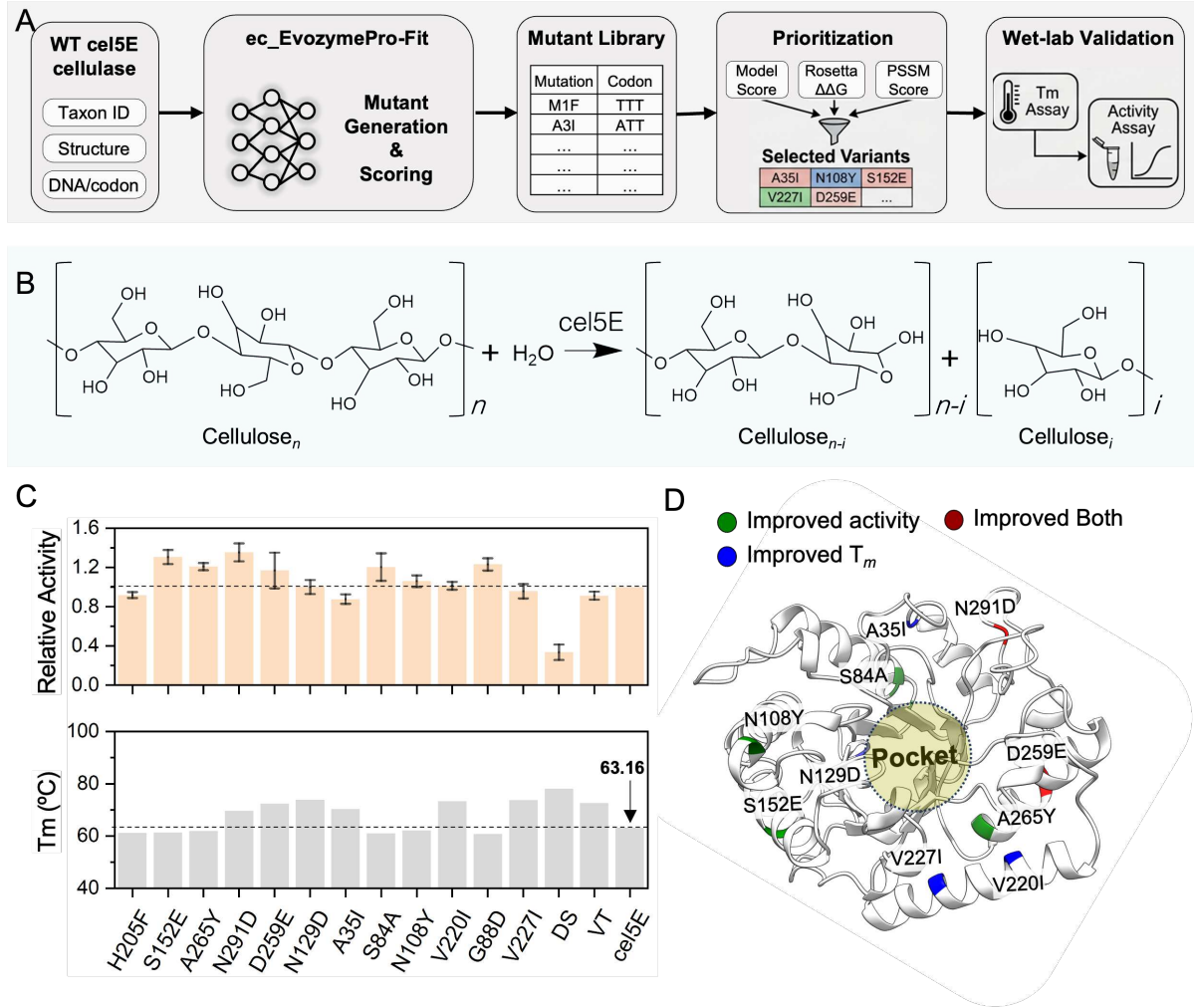


Figure 5: Schematic workflow and experimental validation of the computer-aided engineering of cel5E for enhanced thermostability and catalytic activity. (A) Workflow of the computational protein engineering strategy. (B) Schematic of the cel5E-catalyzed hydrolysis of cellulose. (C) Experimental validation of ec_EZPro-Fit-designed cel5E variants. Relative activity (top) and melting temperature (T_m , bottom) were measured for individual variants. Mutations exhibiting simultaneous improvements in catalytic activity and thermal stability are highlighted, whereas variants showing increased stability but reduced activity illustrate the intrinsic stability-activity trade-off. (D) Structural representation of the cel5E cellulase protein, highlighting spatial distribution of mutated residues mapped onto the 3D structure, with color-coding indicating improvements in activity (green), T_m (blue), or both (red) relative to the wild-type.

SUPPORTING INFORMATION

Table S1: Nucleotide sequence of the cel5E.

Protein	Sequence
cel5E	ATGAAAGCGGTTGATCCGTTTGAAATGGTGCGCAAAATGGGCATGGGCACCAATCTGGGCAATACCCTGGAAGCGC CGTATGAAGGCAGCTGGAGCAAAAGCGCGATGGAATATTATTTTGACGACTTCAAGGCGGCGGGCTATAAGAATGT GCGCATTCCGGTGCGCTGGGATAATCATACCATGCGCACCTATCCGTATACCATTGATAAAGCGTTTCTGGACCGC GTGGAACAGGTGGTGGATTGGAGCTTAAGCCGCGGCTTTGTTACCATTATTAATAGCCATCATGACGACTGGATCA AGGAGGATTATAACGGCAATATCGAACGCTTCGAGAAGATCTGGGAGCAGATTGCGGAACGCTTTAAAAATAAAAG CGAGAACCTGCTGTTTCGAGATCATGAACGAACCGTTTGGCAATATTACCGATGAACAGATTGATGACATGAACAGC CGCATTCTGAAAATTATCCGCAAAACCAATCCGACCCGATTGTGATTATTGGCGGCGGCTATTGGAATAGCTATA ATACCCTGGTGAATATCAAGATCCCGGATGACCCGTATCTGATCGGCACCTTTCATTATTATGATCCGTATGAGTT CACCCACAAGTGGCGCGGCACCTGGGGTACTCAAGAAGATATGGATACCGTGGTGCGCGTGTGTTGATTTTGTGAAA AGCTGGAGCGATCGCAATAATATTCCGGTGATTTTGGCGAATTTGCGGTGATGGCGTATGCGGATCGCACCAGCC GTGTTAAATGGTATGATTTTATTAGCGACGCGGCGCTGGAACGCGGCTTTGCGTGTAGTGTGTTGGGATAATGGCGT GTTTGGCAGCCTGGATAATGATATGGCGATTATAATCGCGACACCCGCACCTTTGATACCGAAATTCTGAATGCG CTGTTTAAC

Table S2: Oligonucleotide primer sequences used for site-directed mutagenesis.

Name	Primer sequence
H205F_F	GTTACCTTTAAGTGGCGCGGCACC
H205F_R	CCACTTAAAGGTGAACCTCATACGGATC
S84A_F	TGGATTGGCGTTGAACGCCGGCGTTT
S84A_R	GGCTTAACGCCCAATCCACCACCTGT
A351_F	GCAAAAAGCATTATGGAATATTATTTTGACGACTTCA
A351_R	ATTTTAATCCTTTGTCCTCACGGTG
D259E_F	TTAAATGGTATGAATTTATTAGCGACGCGGC
D259E_R	CGCTAAATTAATCATACATTTAAACGGCTGG
N129D_F	AAGCGAGTCTGCTGTTGAGATCATG
N129D_R	CAGCAGATCCTCGCTTTATTTTAAAGCGTTC
N108Y_F	TATAACGGCTATATCGAACGCTTCGAGAAG
N108Y_R	CGTTCGATAGCGTTATAGCCGTTATATCTT
G88D_F	AGCGCGATTTTGTTACCATTATTAATAGCCATC
G88D_R	GTACAAATCGCGGCTTAGCT
T299E_F	CACCCCTTGATGAAGAAATTACCTGTG
T299E_R	GAATTTCTTCATCAAAGGTGCG
S152E_F	GCAATGAACAGCAATCTGAAAATTATCCG
S152E_R	GAATGCGTTCGTTTCATGTCATCAATCTGTTC
N106D_F	GGAGGATTATGATGGCAATATCGAACGCTC
N106D_R	GATATTGCCATCAATCCTCATCCGTGACGT
V227L_F	GTTTGATTTTATTTAAAGCTGGAGCGATCG
V227L_R	CCAGCTTTTAATAAAATCAAAACGCGCAC
V220L_F	GATCCATTGTCGCGCGTATTGATAGTT
V220L_R	CGCACAATGGTATCCATATCTTCTTGAG
N291D_F	GCGATTTATGATCGCGACACCCGCACC
N291D_R	GTGCTGGCGATATAAACGCCATATTCCAG
D218E_F	AAGATATGGAAACCGTGG TGC GCGTG
D218E_R	CCACGGTTTCATATCTTCTTGAGTACCC
S262I_F	ATTTTAGGATCGCGCGCTGGA
S262I_R	CCGCGTCAATAATAAACATCATACCATTAAACACGG

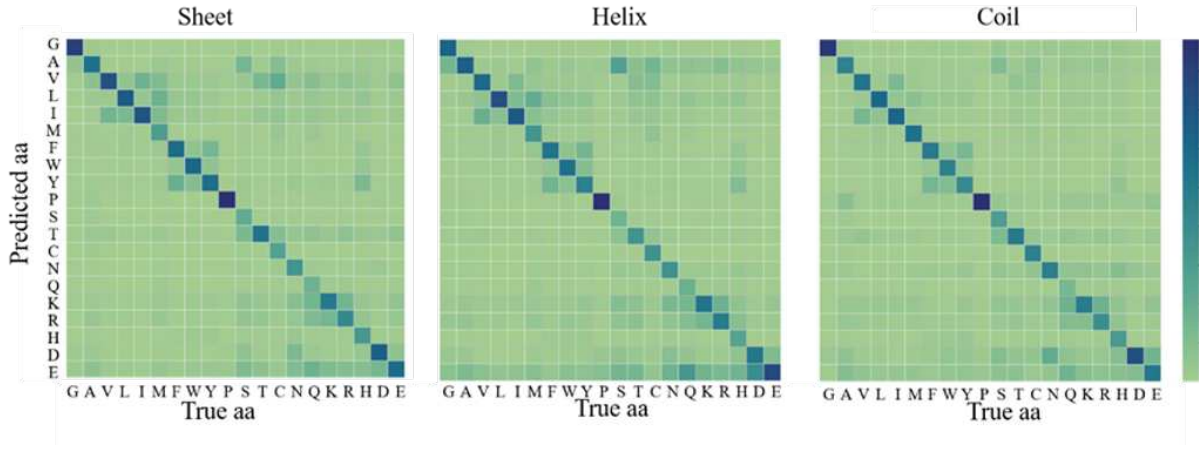


Figure S1: Heatmap illustrating amino acid replacement preferences between generated and true sequences across Sheet, Helix, and Coil secondary structures. Darker diagonal elements indicate high recovery accuracy for individual residues, and structured off-diagonal clusters highlight biochemically conservative mutations maintained by EZPro-Fit.

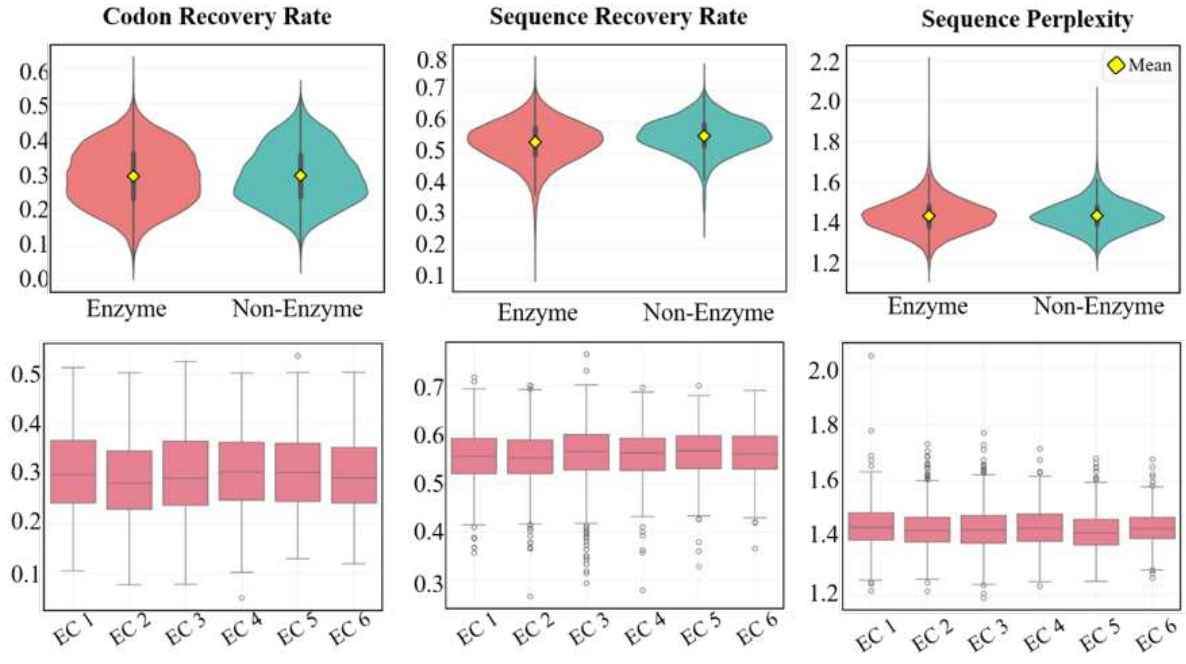


Figure S2: Sequence recovery metrics and perplexity distributions of EZPro-Fit. (Top) Violin plots comparing codon recovery, sequence recovery, and sequence perplexity between enzyme and non-enzyme datasets. (Bottom) Box plots showing consistent performance across the six Enzyme Commission (EC 1–6) classes, indicating uniform generative capacity across diverse catalytic functions.

Supplementary Files

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

- [7736940supp180099tk05c6convrt.pdf](#)