XPRO-DESIGN: RATIONAL PROTEIN ENGINEERING FRAMEWORK USING EXPLAINABLE AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein engineering seeks to rationally tailor proteins to achieve specific structural and functional objectives. These objectives encompass enhancing catalytic efficiency, modifying substrate specificity, improving binding affinity, reducing immunogenicity, and increasing stability under adverse conditions. A major bottleneck is protein instability, as elevated temperatures often drive degradation and compromise activity. Developing thermostable proteins is therefore a key objective in engineering efforts. Here, we present XPro-Design, an explainable AI driven framework for protein optimization that integrates amino acid-level explanations of functional impact into generative modeling. Our method captures epistatic interactions and the mutational landscape by training a low-rank matrix, which biases the generative model toward high-scoring regions of sequence space. This enables targeted generation of candidate variants optimized for thermostability, while remaining extensible to other objectives. XPro-Design further uses distribution tempering and annealing to effectively balance exploration vs exploitation without compromising on structural integrity. We demonstrate rational, causality driven design of protein variants with melting temperatures nearly 2x that of their wild-type counterparts, while preserving binding pocket integrity and domain architecture. Moreover, engineered variants show up to 38% lower folding free energy relative to wild-type indicating significantly enhanced thermodynamic stability. XPro-Design establishes a generalizable strategy for explainable and controllable protein design, enabling multi-objective optimization beyond thermostability.

1 INTRODUCTION

Protein engineering is a cornerstone of modern biotechnology, enabling the rational tailoring of proteins for therapeutic, industrial, and synthetic biology applications. Engineered proteins can improve binding affinity, reduce immunogenicity, extend half-life, function in harsh environments, catalyze reactions with higher efficiency, or perform novel tasks such as biosensing and pathway modulation. Despite these diverse applications, a central challenge remains: amino acid mutations often exert complex, non-additive effects on structure and function, making the sequence–function landscape difficult to navigate. Among targeted properties, thermostability is especially critical, as proteins unstable at elevated temperatures readily unfold, aggregate, and lose function. Stabilizing determinants include hydrophobic core packing, hydrogen-bond networks, covalent linkages such as disulfide bonds, and minimization of unfavorable electrostatic or solvent-exposed hydrophobic interactions, whereas disruptions to these features often destabilize proteins. Balancing these opposing contributions defines the mutational landscape of thermostability and underscores the need for methods that accurately capture and exploit sequence–structure–function relationships.

Protein stability prediction has been approached through both physics-based and machine learning methods. Classical tools such as FoldX (Schymkowitz et al., 2005) and Rosetta (Leaver-Fay et al., 2011; Fleishman et al., 2011; Leman et al., 2020) estimate mutational effects on folding free energy by modeling structural energetics and have long served as reference points for benchmarking. More recent data-driven approaches, including DDGun/DDGun3D (Montanucci et al., 2019), and DDGemb (Savojardo et al., 2025), leverage evolutionary features or embeddings from protein language models (Rives et al., 2019; Lin et al., 2022) to predict $\Delta\Delta G$ of mutation. Other efforts such as DeepTM (Li et al., 2023), ProTstab2 (Yang et al., 2022), and DeepSTABp (Jung et al., 2023)

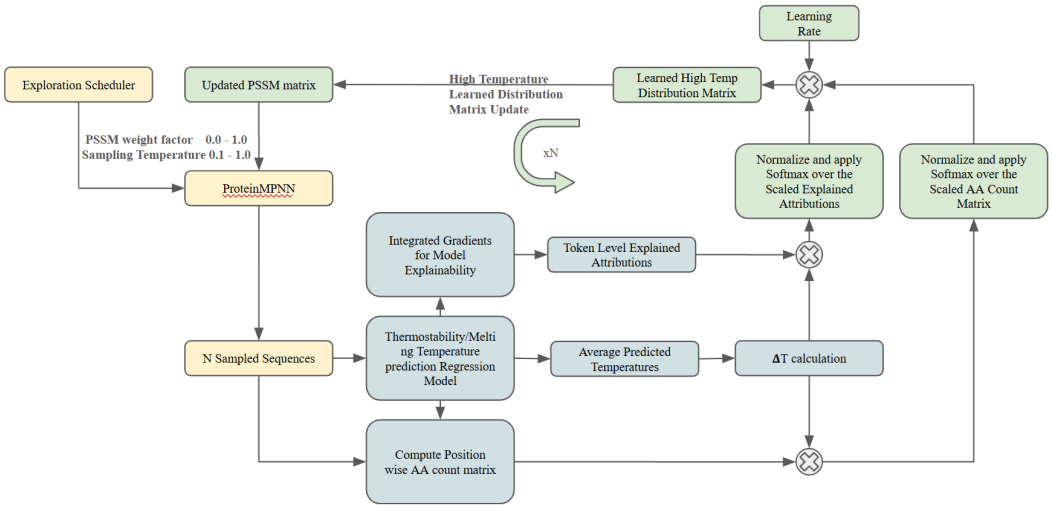

Figure 1: XPro-Design architecture

focus on predicting melting temperatures directly from sequence information, while transformer-based models like TemBERTure (Rodella et al., 2024), ESMStabP (Ramos et al., 2025), TemStaPro (Pudžiuvelytė et al., 2024), and ESMTherm/EsmTemp (Chu et al., 2024; Sułek et al., 2024) build on large pretrained protein models to capture thermostability across diverse families. Structure-aware neural networks, including ThermoMPNN (Dieckhaus et al., 2024), ThermoMPNN-D (Dieckhaus & Kuhlman, 2025), and SPURS (Li & Luo, 2025), further incorporate backbone geometry to refine predictions for thermostability and T_m. While these tools provide valuable guidance for engineering, most have been trained primarily on single- or double-point mutations and often show reduced accuracy when generalized to multi-site or structure-wide mutational designs, limiting their utility in large-scale protein re-engineering.

In parallel, generative modeling approaches have shifted the focus from prediction to design. Sequence-to-structure inverse folding methods such as ProteinMPNN (Dauparas et al., 2022), HyperMPNN (Ertelt et al., 2024), ESM-IF (Hsu et al., 2022), AlphaDesign (Jendrusch et al., 2025), DivPro (Zhou et al., 2025) and PiFold (Gao et al., 2022) generate sequences compatible with given backbones, and in some cases have been experimentally validated for stable protein design. Diffusion and flow-based frameworks, including RFdiffusion (Watson et al., 2023), MapDiff (Bai et al., 2025), RareFold (Li et al., 2025), and ADFLIP (Yi et al., 2025) introduce probabilistic sampling strategies that enable exploration of novel topologies and controlled backbone design. Specialized models such as AntiFold (Høie et al., 2024) extend generative principles and modeling to antibody design or specialized protein families. However, most of these approaches are trained on broad protein structure datasets (Berman et al., 2000a) and are not optimized for task-specific objectives such as thermostability. For this reason, we adopt ProteinMPNN as a general-purpose baseline and HyperMPNN as a task-specific prior for high-temperature stability design, ensuring that comparisons are grounded in models directly relevant to our objective.

XPro-Design enhances inverse folding frameworks such as ProteinMPNN by overcoming biases from mesophile-dominated training data, which undersample rare but functionally important substitutions. By correcting this bias, XPro-Design increases access to underrepresented residues while preserving backbone compatibility and foldability. In addition, it leverages tempering and annealing to broaden exploration of sequence space while refining promising directions, enabling more effective navigation of the mutational landscape. Sequences are sampled in paired batches: one modified by a learned low-rank biasing matrix and the other left unbiased as a control. Each batch is evaluated with melting-temperature (Tm) predictors, wrapped with Integrated Gradients (Sundararajan et al. (2017)) modules to generate residue-level attribution maps. By aggregating attribution signals across multiple predictors trained on diverse datasets, XPro-Design reduces model-specific biases and builds robust residue-level guidance for updating biasing weights, while the unbiased batch preserves exploration. The framework also supports position masking, allowing functional domains

such as catalytic residues or binding pockets to remain fixed while surrounding regions are optimized. In experiments, XPro-Design generated sequences with reduced conservation relative to mesophilic proteins yet achieved substantial improvements in predicted thermostability. In some scaffolds, engineered variants exhibited predicted Tm increases of up to 90 °C compared to wild type. All redesigned variants were validated using Boltz-2 (Passaro et al., 2025) and AlphaFold3-based predictors like Chai-1 (team et al., 2024) and AlphaFold3 (Abramson et al., 2024), which confirmed correct folding into stable structures. Further validation with BioEmu (Lewis et al., 2025) equilibrium sampling, short molecular dynamics simulations (Hollingsworth & Dror, 2018), and MM/GBSA (Sun et al., 2014; Genheden & Ryde, 2015) folding-energy calculations supported favorable $\Delta\Delta G$ changes relative to wild type and baseline generative approaches. While thermostability serves as a case study, XPro-Design establishes a generalizable, explainable, and controllable framework for multi-objective protein engineering, with applications ranging from altered substrate selectivity and enhanced cofactor binding to optimized hinge dynamics and reduced immunogenic epitopes.

2 METHODOLOGY

The XPro-Design Architecture (Fig. 1) consists of 3 main components: Sampling, Explanations and Optimization. Each of which are going to be detailed below. For the targets, we selected two proteins with distinct thermostability profiles. The first was Candida Antarctica Lipase B (CalB) (Uniprot ID: P41365; (Berman et al., 2000b)), a widely used biocatalyst in esterification, transesterification, and hydrolysis reactions (PDB ID: 4K6G; (Berman et al., 2000b; Xie et al., 2014a)). Wild-type CalB has a reported melting temperature of 45–60 °C ((Xie et al., 2014b; Le et al., 2012; Qian et al., 2009)) depending on variant, solvent and assay conditions, making it moderately stable but suboptimal for high-temperature industrial processes. To preserve activity during optimization, catalytic pocket residues (Ser105, His224, Asp187, and surrounding binding-site residues) were conserved by leaving them unmasked. The second target was Superoxide Reductase (SOR) (Uniprot ID: P82385) from Pyrococcus Furiosus (PDB ID: 1DQI; (Yeh et al., 2000)), a hyperthermophilic enzyme stable up to 95 °C in oligomeric form (75 °C as monomer). SOR was used as a control scaffold to benchmark improvements from XPro-Design against baseline models (ProteinMPNN, HyperMPNN).

2.1 SAMPLING

XPro-Design leverages inverse folding models to map three-dimensional protein backbones to sequence space. While we use ProteinMPNN and HyperMPNN for ablation studies, the framework is general and can incorporate any inverse folding model. Depending on the design objective, critical residues can be preserved by leaving them unmasked during preprocessing; in our experiments, the catalytic and substrate binding residues were explicitly conserved (Fig. 7), though in practice this can extend to entire substrate-binding domains. Given a protein backbone, the inverse folding model generates conditional probabilities over masked positions while respecting frozen residues. We first sample from the baseline model to obtain amino acid distributions, which serve as priors. XPro-Design then tempers this distribution once by applying a temperature, broadening support and reducing initialization bias. Training thereafter proceeds normally, with the tempered distribution gradually sharpening to a newer distribution. This step mitigates the bias of training data dominated by mesophilic proteins, raises entropy, and prevents the model from becoming trapped in local optima early in optimization.

We represent the amino acid distributions across an aligned protein of length $L = 320$ as a matrix

$$P = (p_{i,\ell}) \in \mathbb{R}^{K \times L},\ K = 20, \quad \sum_{i=1}^{K} p_{i,\ell} = 1\ \ \forall \ell,\ p_{i,\ell} \geq 0 \tag{1}$$

To reduce bias toward highly frequent residues (e.g., conserved amino acids) and to increase the chance of sampling low-probability substitutions, we apply temperature scaling independently to each column. For temperature $T > 0$ and smoothing constant $\varepsilon > 0$, the tempered probabilities are defined as

$$\tilde{p}_{i,\ell} = (p_{i,\ell} + \varepsilon)^{1/T}, \qquad p_{i,\ell}(T) = \frac{\tilde{p}_{i,\ell}}{\sum_{j=1}^{K} \tilde{p}_{j,\ell}} \tag{2}$$

Notation.

- $K = 20$: number of categories (amino acids).
- $L = 320$: sequence length (positions).
- $p_{i,\ell}$: normalized probability of amino acid i at position ℓ.
- $\varepsilon > 0$: smoothing constant ensuring nonzero support for all categories.
- T: temperature parameter; $T = 1$ recovers the original distribution, $T > 1$ broadens the distribution so that rare amino acids become more likely at a given site, and $0 < T < 1$ sharpens the distribution, reinforcing the dominant residue choices.
- $p_{i,\ell}(T)$: tempered probability of amino acid i at position ℓ.

After the one-time tempering step, we applied a linear annealing schedule (from T=1.0 to T=0.1) during sampling, progressively sharpening the distribution and enabling early exploration followed by exploitation. Alternatively, sequences can be sampled in parallel from fixed temperatures (e.g., [0.1, 0.5, 1.0]) to balance exploration and exploitation. Sequence generation uses two complementary strategies: direct sampling from the baseline model and sampling guided by a learned position-specific scoring matrix (PSSM) that biases toward desired residue preferences. Together, these yield a diverse baseline of sequences drawn from both the prior and a tempered distribution. A key advantage of this approach is that model weights remain unchanged. This avoids catastrophic forgetting and prevents convergence to narrow local optima, while also preserving the structural fidelity of the inverse folding model; something that can degrade under fine-tuning on limited datasets.

2.2 Temperature prediction and Explainable AI

Each generated protein sequence $x = (x_1, \ldots, x_L), \quad |x| = L$, is evaluated using Integrated Gradients (IG) to attribute residue-level contributions to the predicted melting temperature (T_m). While we primarily employ three variants of `TemBERTure` for prediction, the framework can incorporate any differentiable T_m or $\Delta\Delta G$ model. Predictions from `DeepSTABp` and `TemStaPro` are also used to establish cross-model correlations and derive consensus thermostability estimates.

For a differentiable predictor f_θ, the IG for residue i is defined as

$$IG_i(x) = (x_i - x_i') \int_0^1 \frac{\partial f_\theta\big(x' + \alpha(x - x')\big)}{\partial x_i}\, d\alpha \tag{3}$$

where x_i is the embedding vector of residue i in sequence x , x_i' is a baseline embedding for residue i (e.g., all-zero or reference amino acid), $\alpha \in [0, 1]$ is the interpolation coefficient along the path from baseline to input, $f_\theta(x)$ is the predicted Tm from model θ and $IG_i(x)$ is the contribution of residue i to the predicted Tm

Signed attributions $IG_i(x)$ indicate whether a residue increases or decreases predicted T_m. Pairwise effects can be captured by

$$IG_{ij}^{\text{pair}} = \frac{1}{|S|} \sum_{x \in S} IG_i(x) \cdot IG_j(x) \tag{4}$$

where positive values denote synergistic contributions and negative values denote antagonistic interactions. Averaging across sequences helps suppress predictor noise, yielding sharper and more reliable attribution signals.

In scenarios where the predictor is non-differentiable—such as black-box outputs from molecular dynamics simulations or docking—the framework falls back to correlation-based surrogates. Define the positional frequency of amino acid a at site i across sequences S:

$$p_i(a) = \frac{1}{|S|} \sum_{x \in S} \mathbf{1}\{x_i = a\}, \quad a \in AA \tag{5}$$

with correlation to predicted T_m:

$$r_i = \text{corr}\big(p_i, f_{\text{oracle}}(x)\big) \tag{6}$$

Although correlation-based optimization is less efficient and converges more slowly, it can still reveal residue–property associations. By contrast, attribution-guided exploration via IG improves convergence efficiency and produces distributions that closely resemble those of naturally thermostable proteins.

2.3 Learning Sequence Preferences

Each generated batch of sequences is converted into a position-specific scoring matrix (PSSM) of shape $L \times 20$, where L is the sequence length and 20 corresponds to the amino acid types. To update the PSSM, we first compute the mean-centered stability signal for each sequence:

$$\Delta T_m(x) = T_m(x) - \mu_{T_m} \tag{7}$$

where $T_m(x)$ is the predicted melting temperature of sequence x, and μ_{T_m} is the batch mean.

To enhance signal separation, ΔT_m values are rescaled on an exponential scale. For a batch of ΔT_m values, let

$$\Delta T_m^{\max} = \max_{x \in S} |\Delta T_m(x)| \tag{8}$$

We then define

$$\widetilde{\Delta T}_m(x) = \Delta T_m(x) \cdot \left(\frac{|\Delta T_m(x)|}{\Delta T_m^{\max}} \right)^{\gamma} \tag{9}$$

where $\gamma > 0$ is an exponent hyperparameter.

This transformation ensures that large deviations from the mean are amplified, small deviations are attenuated and the sign of $\Delta T_m(x)$ is preserved.

The weighted ΔT_m is then combined with residue-wise attribution scores to form the update term:

$$\Delta\text{PSSM}_{i,a} = \frac{1}{|S|} \sum_{x \in S} \widetilde{\Delta T}_m(x) \cdot IG_{i,a}(x) \tag{10}$$

where $IG_{i,a}(x)$ is the attribution score for amino acid a at position i in sequence x, and S is the batch of sequences.

This procedure ensures that amino acids contributing to low T_m are penalized with amplified negative updates, amino acids contributing to high T_m are rewarded with amplified positive updates, context-dependent effects (epistasis) are captured naturally through batch-level averaging.

Finally, the PSSM is updated iteratively as

$$\text{PSSM}^{(t+1)} = \text{PSSM}^{(t)} + \eta(t) \cdot \Delta\text{PSSM} \tag{11}$$

where the learning rate $\eta(t)$ follows a linear decay schedule from 0.01 to 0.001. Over successive batches, this update biases sampling toward regions enriched in stabilizing mutations while maintaining exploration of diverse sequence space.

2.4 Evaluation of Fold Integrity and Stability

After convergence of the sampling distribution, we generated N sequences for evaluation. Thermostability was first predicted using the methods in Section 2.2, followed by structure prediction with Boltz-2. Predicted structures were aligned to the reference backbone, and Cα RMSD was calculated to assess fold preservation. Additional descriptors like packing density, solvent-accessible surface area (SASA), and inter-residue interaction networks—were computed on both Boltz-2 and energy-relaxed structures to evaluate packing and interaction integrity. For thermodynamic validation, sequences were analyzed with BioEmu to sample 50 equilibrium conformations, subjected to molecular dynamics simulations, and evaluated by MM/GBSA free-energy calculations in OpenMM ((Eastman et al., 2017)). These MM/GBSA energies (denoted as ΔG in tables and plots) are used as relative stability proxies of folded conformations; they should not be interpreted as absolute folding free energies (ΔG_{fold}). The resulting $\Delta\Delta G$ values quantified relative stability across variants. This integrated framework ensured that designed proteins preserved structural topology while exhibiting favorable energetic and thermodynamic profiles.

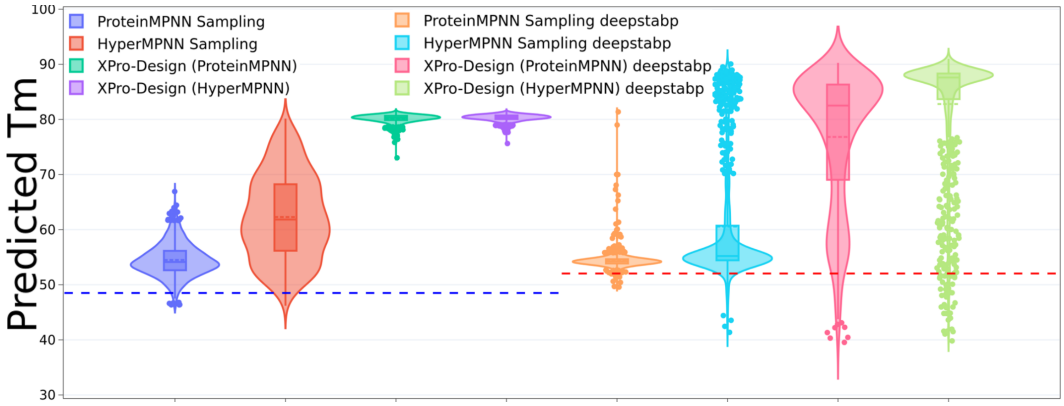

Figure 2: Predicted melting temperature values for CalB variants in °C for the different methods using TemBERTure and DeepSTABp models. We Observe the spread of High-Tm sequences is considerably narrower for XPro-Design compared to even HyperMPNN which was specifically trained to generate thermophile variants.

Table 1: Predicted Temperatures and Sequence Recovery

methods	TemBERTure Tm (°C) mean, max ↑	DeepSTABp Tm (°C) mean, max ↑	Sequence Recovery	Sequence Diversity
ProteinMPNN	54.4, 66.9	54.4, 81.3	$56.3\% \pm 2.5\%$	0.750
HyperMPNN	62.2, 80.1	60.5, 90.0	$48.7\% \pm 1.5\%$	0.732
XPro-Design(P) ours	80.1, 81.5	76.8, 90.2	$48.2\% \pm 1.5\%$	0.734
XPro-Design(H) ours	80.3, 81.5	82.7, 90.9	$34.8\% \pm 1\%$	0.602

3 Results

3.1 Melting Temperature Predictions

Once XPro-Design had learned the optimized protein sequence space, we generated 1,000 sequences each from ProteinMPNN, HyperMPNN, XPro-Design(P) (with ProteinMPNN sampling), and XPro-Design(H) (with HyperMPNN sampling) at sampling temperatures of 0.1 and 0.5.

As illustrated in Figure 2, our approach outperformed even the specialized HyperMPNN model in generating sequences with substantially higher predicted melting temperatures (Tm). Thermostability was independently verified using both TemBERTure and DeepSTABp predictors, with DeepSTABp evaluated at a growth temperature of 37°C. Consistent with our hypothesis, XPro-Design reliably identified high-Tm sequence variants for the given backbones regardless of the underlying sampling model. The difference between XPro-Design(P) and (H) was negligible in terms of predicted temperature performance and sampled AA space, indicating strong generalization.

Among the baseline methods, ProteinMPNN exhibited the weakest performance, with the lowest mean and maximum predicted T_m across both TemBERTure and DeepSTABp (Tab 1). HyperMPNN generated sequences with higher maximum temperatures than ProteinMPNN, though mean values remained significantly below those achieved by our methods.

In contrast, both XPro-Design variants consistently produced sequences with markedly elevated thermostability. TemBERTure predicted mean T_m values above 80 °C for both samplers, while DeepSTABp predictions reached up to 91 °C. The temperature distribution was slightly sharper for XPro-Design(H) as the sampler, though the overall improvement over XPro-Design(P) was marginal when considering sequence recovery and diversity as shown in Table 1.

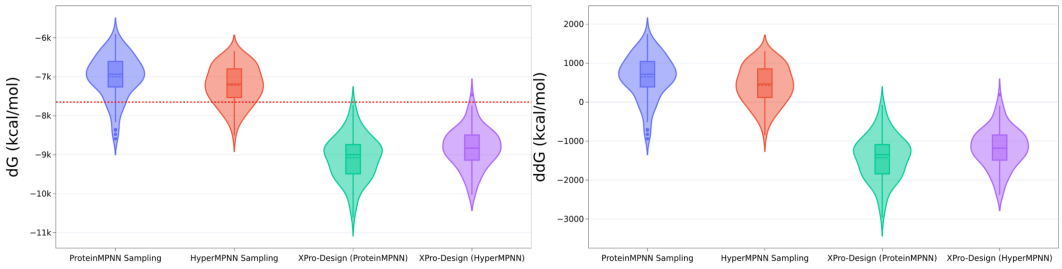

Figure 3: Violin plots showing ΔG (left) and $\Delta\Delta G$ (right) distributions of variants generated from the different methods computed against CalB WT

Table 2: Structure Predictions

methods	Alphafold PTM	Alphafold pLDDT	RMSD (Å) ↓
ProteinMPNN	0.966 $\pm$0.004	0.94 $\pm$0.01	1.94$\pm$0.25
HyperMPNN	0.958 $\pm$0.007	0.92 $\pm$0.01	2.02 $\pm$0.33
XPro-Design (P) ours	0.947 $\pm$0.009	0.90 $\pm$0.01	2.24 $\pm$0.69
XPro-Design (H) ours	0.915 $\pm$0.015	0.85 $\pm$0.02	3.12 $\pm$0.56

3.2 Sequence Diversity and Recovery

Sequence diversity was quantified using a k-mer–based Jaccard similarity approach ($k = 3$), which efficiently captures local compositional differences without requiring full pairwise alignments. Diversity is expressed as 1 minus the Jaccard similarity (Brohee & Van Helden, 2006), averaged over all sequence pairs. As shown in Table 1, all variants exhibit comparable diversity except XPro-Design(H), which converged on a narrower sequence space.

Sequence recovery was computed as the fraction of residues matching the wild-type (WT) sequence at aligned positions, reflecting the balance between conservation and exploration. ProteinMPNN achieved the highest recovery (56.3%), indicating strong preservation of WT residues but limited mutational diversity, which corresponds to weaker thermostability and $\Delta\Delta G$ improvements. HyperMPNN (48.7% recovery) explores more of sequence space, yielding more thermophilic designs, though gains remain modest.

XPro-Design variants achieved recoveries of 48.2% (P) and 34.7% (H) while generating sequences with markedly improved thermostability and reduced $\Delta\Delta G$. Lower recovery for HyperMPNN-based designs reflects its bias toward hyperthermophilic residues. XPro-Design with ProteinMPNN strikes an optimal balance, maintaining high sequence coverage while producing superior designs without model-specific fine-tuning. Both XPro-Design variants converge to a distinct amino acid distribution, clearly separating them from the baseline models.

3.3 Structure Prediction and RMSD

We predicted the structures for all generated sequences using the Boltz-2 model, with multiple sequence alignments (MSAs) obtained from the ColabFold (Mirdita et al., 2022) MSA server. Wild-type templates were not provided during the prediction to ensure unbiased folding assessments. Across all methods, the designed sequences were predicted to fold correctly, with XPro-Design variants consistently demonstrated successful folding. Backbone RMSD values of the designed sequences relative to the wild-type backbone showed minimal deviations, confirming structural conservation despite extensive sequence redesign. As illustrated in Figure 6, the variant V_2372 folded nearly identically to the wild-type CalB backbone while exhibiting a 38.7% reduction in predicted $\Delta\Delta G$ and a 63% improvement in predicted melting temperature. Predicted Aligned Error (PAE) and predicted Local Distance Difference Test (pLDDT) scores are summarized in Table 2, all within acceptable confidence thresholds, further supporting the structural reliability of the generated sequences.

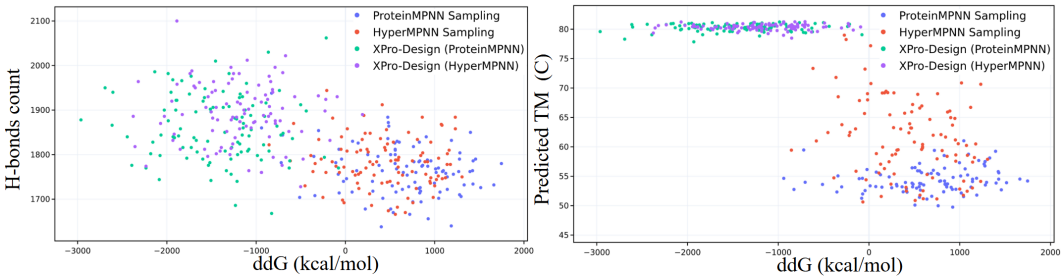

Figure 4: Scatter Plot of the Predicted melting Temperatures vs ddG

Table 3: Folding free energy, $\Delta\Delta G$ and

Methods	ΔG mean, best (kcal · mol^{-1}) ↓	$\Delta\Delta G$ mean, best (±kcal·mol^{-1}) ↓	Normalized entropy (J·mol^{-1}·K^{-1}) ↓	Normalized entropy Core (J·mol^{-1}·K^{-1}) ↓
ProteinMPNN	-6999, -8588	650, -939	0.707 ± 0.010	0.696 ± 0.013
HyperMPNN	-7215, -8504	433, -855	0.682 ± 0.009	0.681 ± 0.011
XPro-Design(P)	-9075,-10612	-1426, -2962	0.673 ± 0.009	0.670 ± 0.012
XPro-Design(H)	-8828, -10026	-1179, -2377	0.660 ± 0.007	0.658 ± 0.010

3.4 Free Energy and Packing Entropy Analyses

We evaluated thermodynamic stability using folding free energies (ΔG) and relative stability changes ($\Delta\Delta G$) with respect to the wild type (Tab 3). Stabilizing variants were defined by lower ΔG and negative $\Delta\Delta G$. Baseline ProteinMPNN generated stabilizing variants in only 14% of cases, with mean ΔG = –6999 kcal·mol^{-1} and mean $\Delta\Delta G$ = +650 kcal·mol^{-1}, indicating overall destabilization. HyperMPNN modestly improved performance (21% stabilizing), with mean ΔG = –7215 kcal·mol^{-1} and mean $\Delta\Delta G$ = +433 kcal·mol^{-1}. These results confirm that baseline models rarely introduce consistently stabilizing substitutions.

In contrast, XPro-Design produced near-universal stabilization (Fig. 3). With ProteinMPNN as the sampler, all variants were stabilizing, with mean ΔG = –9075 kcal·mol^{-1} and mean $\Delta\Delta G$ = –1426 kcal·mol^{-1}. Using HyperMPNN yielded similarly strong results (99% stabilizing, mean ΔG = –8828 kcal·mol^{-1}, mean $\Delta\Delta G$ = –1179 kcal·mol^{-1}). Both samplers achieved substantially more favorable $\Delta\Delta G$ values than either baseline, consistent with the enhanced thermostability and packing analyses (Fig. 4).

To further probe stability, we computed packing-derived residue entropies with PACKMAN (Khade, 2024) (Voronoi/Delaunay geometry → packing fraction → entropy). Normalized entropy values (removing length bias)(Fig 10) showed clear reductions for XPro-Design: ProteinMPNN (0.707), HyperMPNN (0.682), XPro-Design(P) (0.673), and XPro-Design(H) (0.660). Hydrophobic-core entropies followed the same trend (0.696, 0.681, 0.670, 0.658, respectively), confirming that XPro-Design variants adopt tighter, less flexible packing. These paired results indicate that improved thermostability arises from redistributed packing patterns rather than simple global compression of the core volume.

3.5 Bond Analysis

The analysis of non-covalent interactions revealed that XPro-Design variants consistently formed a higher number of hydrogen bonds relative to baseline ProteinMPNN and HyperMPNN designs (Fig. 5). In the same figure, a comparable trend was observed for salt bridges, π-cation interactions, and π-stacking contacts, all of which increased significantly in the redesigned variants. Disulfide bond counts remained largely unchanged across methods, indicating that the global covalent connectivity of the proteins was preserved.

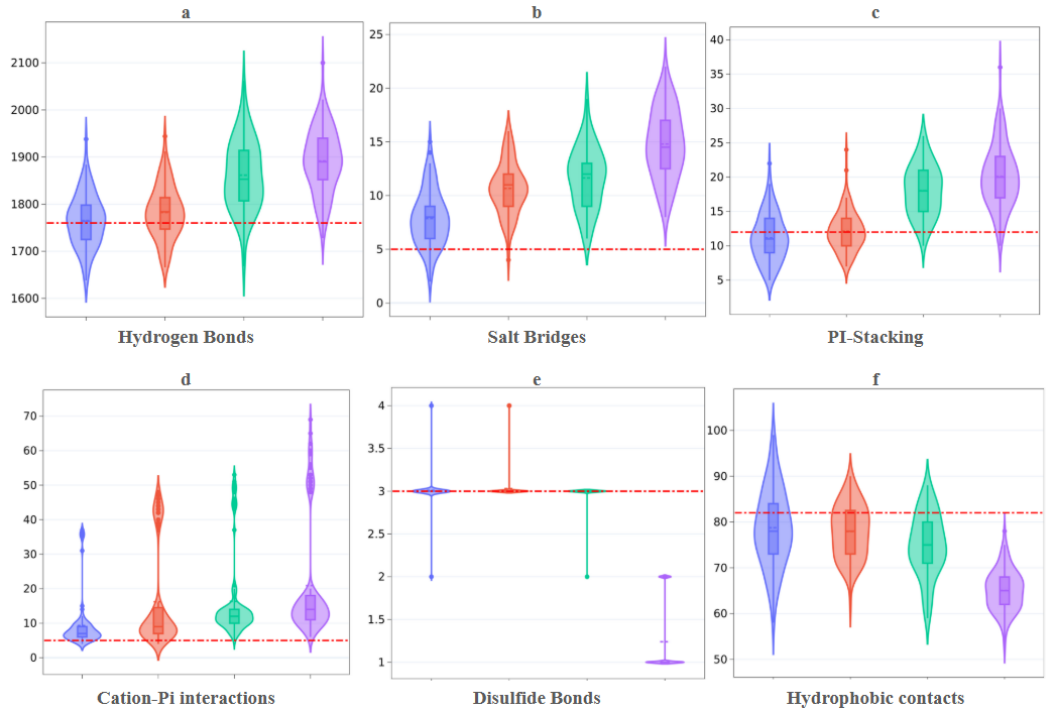

Figure 5: Inter residue Interactions: ProteinMPNN (blue), HyperMPNN (red), XPro-Design(P) (green), XPro-Design(H) (purple)

Interestingly, this redistribution of interactions was accompanied by a modest reduction in hydrophobic contacts. However, rather than reflecting destabilization, this shift appears to be a compensatory effect: the gain in directional, energetically favorable interactions such as hydrogen bonds and electrostatic or aromatic contacts outweighs the slight decrease in non-specific hydrophobic packing.

Taken together, these results indicate that XPro-Design variants achieve improved stability not by maximizing hydrophobic burial alone, but by reinforcing a diverse network of stabilizing noncovalent interactions. This richer interaction landscape likely contributes to the enhanced thermostability and folding robustness observed in our designs.

4 Conclusion

We introduced XPro-Design, a novel framework for protein sequence optimization that leverages explainable AI in a gradient-free setting. By using attribution methods such as Integrated Gradients, the approach provides residue-level interpretability while guiding optimization without task-specific predictors. Unlike conventional baselines, XPro-Design requires no fine-tuning, operating directly on inverse folding models while retaining their structural fidelity and broad applicability. The framework balances exploration and exploitation through tempered initialization and annealed sampling, systematically uncovering stabilizing mutations missed by baseline methods. As a result, XPro-Design designed orders of magnitude more stable sequences than ProteinMPNN and HyperMPNN. It consistently yielded higher predicted thermostability, near-universal shifts toward stabilizing $\Delta\Delta G$, and reduced packing entropies indicative of more favorable folds; all while preserving sequence diversity. These results highlight XPro-Design as a transformative step in protein engineering: scalable, interpretable, and capable of delivering unprecedented improvements in stability beyond the reach of existing generative models.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Peizhen Bai, Filip Miljković, Xianyuan Liu, Leonardo De Maria, Rebecca Croasdale-Wood, Owen Rackham, and Haiping Lu. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nature Machine Intelligence*, pp. 1–13, 2025.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000a.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000b.

Sylvain Brohee and Jacques Van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1):488, 2006.

Simon KS Chu, Kush Narang, and Justin B Siegel. Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset. *PLoS computational biology*, 20(7):e1012248, 2024.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Henry Dieckhaus and Brian Kuhlman. Protein stability models fail to capture epistatic interactions of double point mutations. *Protein Science*, 34(1):e70003, 2025.

Henry Dieckhaus, Michael Brocidiacono, Nicholas Z Randolph, and Brian Kuhlman. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the national academy of sciences*, 121(6):e2314853121, 2024.

Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

Moritz Ertelt, Phillip Schlegel, Max Beining, Leonard Kaysser, Jens Meiler, and Clara T Schoeder. Hypermpnn–a general strategy to design thermostable proteins learned from hyperthermophiles. *bioRxiv*, 2024.

Sarel J Fleishman, Andrew Leaver-Fay, Jacob E Corn, Eva-Maria Strauch, Sagar D Khare, Nobuyasu Koga, Justin Ashworth, Paul Murphy, Florian Richter, Gordon Lemmon, et al. Rosettascripts: a scripting language interface to the rosetta macromolecular modeling suite. *PloS one*, 6(6):e20161, 2011.

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Samuel Genheden and Ulf Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert opinion on drug discovery*, 10(5):449–461, 2015.

Magnus Haraldson Høie, Alissa Hummer, Tobias H Olsen, Broncio Aguilar-Sanjuan, Morten Nielsen, and Charlotte M Deane. Antifold: Improved antibody structure-based design using inverse folding. *arXiv preprint arXiv:2405.03370*, 2024.

Scott A Hollingsworth and Ron O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6): 1129–1143, 2018.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *ICML*, 2022. doi: 10.1101/2022.04.10.487779. URL https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779.

Michael A Jendrusch, Alessio LJ Yang, Elisabetta Cacace, Jacob Bobonis, Carlos GP Voogdt, Sarah Kaspar, Kristian Schweimer, Cecilia Perez-Borrajero, Karine Lapouge, Jacob Scheurich, et al. Alphadesign: A de novo protein design framework based on alphafold. *Molecular Systems Biology*, pp. 1–24, 2025.

Felix Jung, Kevin Frey, David Zimmer, and Timo Mühlhaus. Deepstabp: a deep learning approach for the prediction of thermal protein stability. *International Journal of Molecular Sciences*, 24(8): 7444, 2023.

Pranav Khade. PACKMAN: Python package built around protein structure and dynamics. https://github.com/Pranavkhade/PACKMAN, 2024. Accessed: 2025-09-24.

Quang Anh Tuan Le, Jeong Chan Joo, Young Je Yoo, and Yong Hwan Kim. Development of thermostable candida antarctica lipase b through novel in silico design of disulfide bridge. *Biotechnology and bioengineering*, 109(4):867–876, 2012.

Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian W Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in enzymology*, volume 487, pp. 545–574. Elsevier, 2011.

Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17 (7):665–680, 2020.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025.

Mengyu Li, Hongzhao Wang, Zhenwu Yang, Longgui Zhang, and Yushan Zhu. Deeptm: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences. *Computational and Structural Biotechnology Journal*, 21:5544–5560, 2023.

Qiuzhen Li, Diandra Daumiller, and Patrick Bryant. Rarefold: Structure prediction and design of proteins with noncanonical amino acids. *bioRxiv*, pp. 2025–05, 2025.

Ziang Li and Yunan Luo. Rewiring protein sequence and structure generative models to enhance protein stability prediction. In *International Conference on Research in Computational Molecular Biology*, pp. 255–259. Springer, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Ludovica Montanucci, Emidio Capriotti, Yotam Frank, Nir Ben-Tal, and Piero Fariselli. Ddgun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. *BMC bioinformatics*, 20(Suppl 14):335, 2019.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pp. 2025–06, 2025.

Ieva Pudžiuvelytė, Kliment Olechnovič, Egle Godliauskaite, Kristupas Sermokas, Tomas Urbaitis, Giedrius Gasiunas, and Darius Kazlauskas. Temstapro: protein thermostability prediction using sequence representations from protein language models. *Bioinformatics*, 40(4):btae157, 2024.

Zhen Qian, John R Horton, Xiaodong Cheng, and Stefan Lutz. Structural redesign of lipase b from candida antarctica by circular permutation and incremental truncation. *Journal of molecular biology*, 393(1):191–201, 2009.

Marcus Ramos, Robert L Jernigan, and Mesih Kilinc. Esmstabp: A regression model for protein thermostability prediction. *bioRxiv*, 2025.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

Chiara Rodella, Symela Lazaridi, and Thomas Lemmin. Temberture: advancing protein thermostability prediction with deep learning and attention mechanisms. *Bioinformatics Advances*, 4(1): vbae103, 2024.

Castrense Savojardo, Matteo Manfredi, Pier Luigi Martelli, and Rita Casadio. Ddgemb: predicting protein stability change upon single-and multi-point variations with embeddings and deep learning. *Bioinformatics*, 41(1):btaf019, 2025.

Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

Adam Sułek, Jakub Jończyk, Patryk Orzechowski, Ahmed Abdeen Hamed, and Marek Wodziński. Esmtemp-transfer learning approach for predicting protein thermostability. In *International Conference on Computational Science*, pp. 187–194. Springer, 2024.

Huiyong Sun, Youyong Li, Sheng Tian, Lei Xu, and Tingjun Hou. Assessing the performance of mm/pbsa and mm/gbsa methods. 4. accuracies of mm/pbsa and mm/gbsa methodologies evaluated by various simulation protocols using pdbbind data set. *Physical Chemistry Chemical Physics*, 16(31):16719–16729, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pp. 2024–10, 2024.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Yuan Xie, Jiao An, Guangyu Yang, Geng Wu, Yong Zhang, Li Cui, and Yan Feng. Enhanced enzyme kinetic stability by increasing rigidity within the active site. *Journal of Biological Chemistry*, 289 (11):7994–8006, 2014a.

Yuan Xie, Jiao An, Guangyu Yang, Geng Wu, Yong Zhang, Li Cui, and Yan Feng. Enhanced enzyme kinetic stability by increasing rigidity within the active site. *Journal of Biological Chemistry*, 289 (11):7994–8006, 2014b.

Yang Yang, Jianjun Zhao, Lianjie Zeng, and Mauno Vihinen. Protstab2 for prediction of protein thermal stabilities. *International journal of molecular sciences*, 23(18):10798, 2022.

Andrew P Yeh, Yonglin Hu, Francis E Jenney Jr, Michael WW Adams, and Douglas C Rees. Structures of the superoxide reductase from pyrococcus furiosus in the oxidized and reduced states. *Biochemistry*, 39(10):2499–2508, 2000.

Kai Yi, Kiarash Jamali, and Sjors HW Scheres. All-atom inverse protein folding through discrete flow matching. *arXiv preprint arXiv:2507.14156*, 2025.

Xinyi Zhou, Guibao Shen, Yingcong Chen, Guangyong Chen, and Pheng Ann Heng. Divpro: diverse protein sequence design with direct structure recovery guidance. *Bioinformatics*, 41 (Supplement_1):i382–i390, 2025.

A Appendix

A.1 Future Work

Future work planned around XPro-Design involves validating some of the top variants in a lab along with quantifying their half lives at an elevated range of temperatures compared to the WT proteins. Also we are currently working on redesigning several other industry relevant proteins to operate at elevated temperatures.

Moreover we have already started testing XPro-Design towards substrate binding selectivity for enzymes, optimizing protein-protein binding affinity by improving the interaction profiles on the binding domains as well as evaluating XPro-Design towards improving enzymatic kinetics by re-designing hinge regions towards higher catalytic efficiency.

A.2 Shannon Entropy

The Shannon entropy of a categorical distribution p is defined as

$$H(p) = -\sum_{i=1}^{K} p_i \, \log p_i. \tag{12}$$

Applying tempering yields the distribution $p(T)$, whose entropy is

$$H(p(T)) = -\sum_{i=1}^{K} p_i(T) \, \log p_i(T) \tag{13}$$

For $T > 1$, it follows that

$$H(p(T)) \geq H(p),$$

with equality if and only if p is uniform.

Notation:

- $H(p)$: Shannon entropy of distribution p.
- $\log$: natural logarithm.
- $p_i(T)$: tempered probability of category i.

A.3 Biological intuition behind Temperature Scaling of prior distribution

Temperature scaling with $T > 1$ increases the probability of rare substitutions that might otherwise be ignored because of under representation in the training data, thus encouraging exploration of sequence diversity. Conversely, setting $T < 1$ amplifies the dominance of conserved residues, reinforcing evolutionary constraints. This single parameter therefore provides a biologically interpretable knob to balance between conservation and diversity in sampling.

A.4 CalB Structural Results

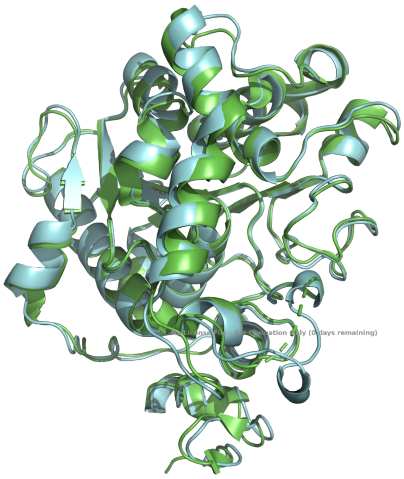

Figure 6: lowest ddG variant V_2372 (blue) overlaid over WT 4K6G (green) structure

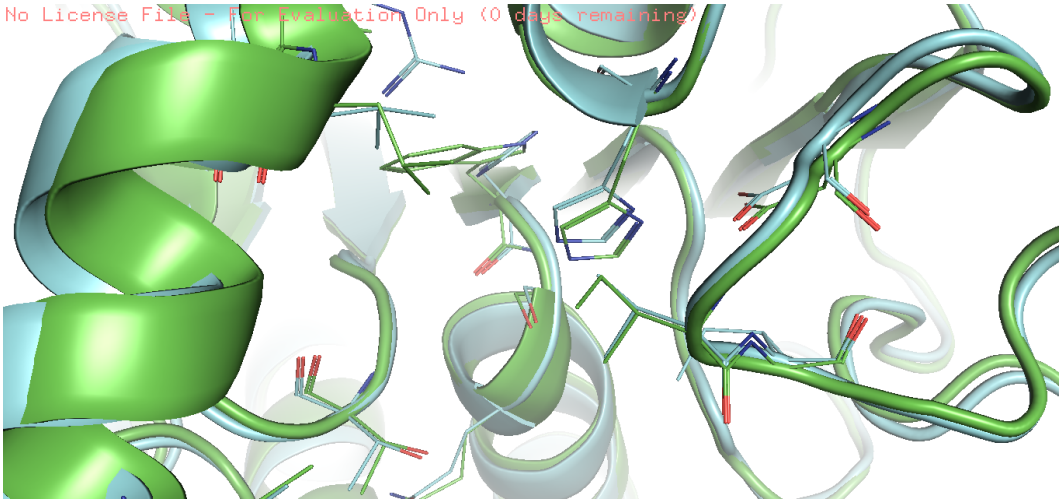

Figure 7: lowest ddG variant V_2372 (blue) overlaid over WT 4K6G (green) structure shows that the substrate bidning pocket is well conserved. Catalytic Triad Residues S105, D187, H224 and Substrate binding residues T40, E188, L278 were specifically conserved. Residues I189, V190, I285 were masked out, yet we see that XPro-Design substitutions represents a conservative change (Ile $\rightarrow$ Leu) within the substrate binding pocket, and is not expected to drastically alter the overall hydrophobic character even for redesigned residues not explicitly conserved, though subtle changes in side-chain packing or pocket geometry may occur.

A.5 AA-wise Distribution Shifts

To evaluate how design strategies altered amino acid usage, we analyzed the distribution of residue classes across surface, core, and overall regions of the protein (Figure 9 & 8). Both baseline samplers (ProteinMPNN and HyperMPNN) preserved broad compositional trends but differed in their bias toward polar residues on the surface and hydrophobic residues in the core.

XPro-Design introduced a clear shift in these distributions. On the surface, it reduced excessive polar enrichment while slightly increasing charged and special residues, suggesting more balanced solvent exposure. In the core, XPro-Design produced a higher fraction of hydrophobic residues and a modest rise in glycines, consistent with tighter packing and increased conformational adaptability. When averaged over the full sequence, the distributions from both XPro-Design variants diverged from the baselines in a consistent manner, indicating that optimization not only improved thermostability but also drove distinct residue-level preferences aligned with thermophilic design principles.

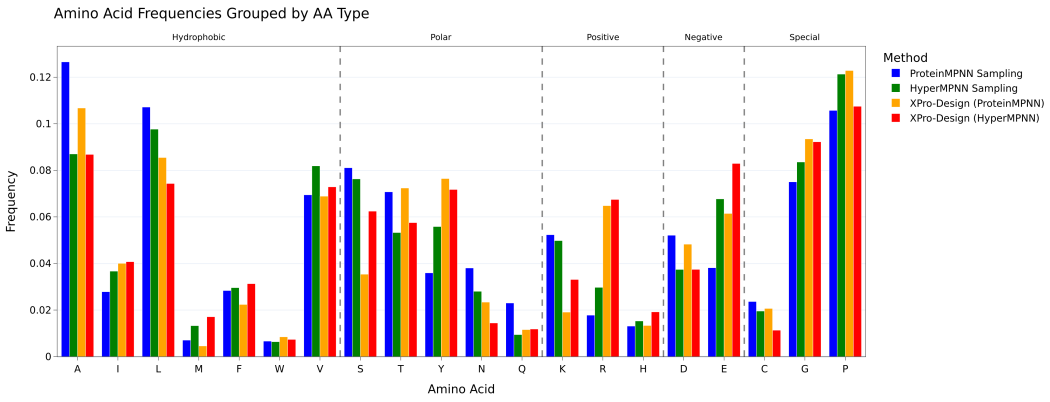

Figure 8: AA wise distribution shift of XPro-Design from baseline models

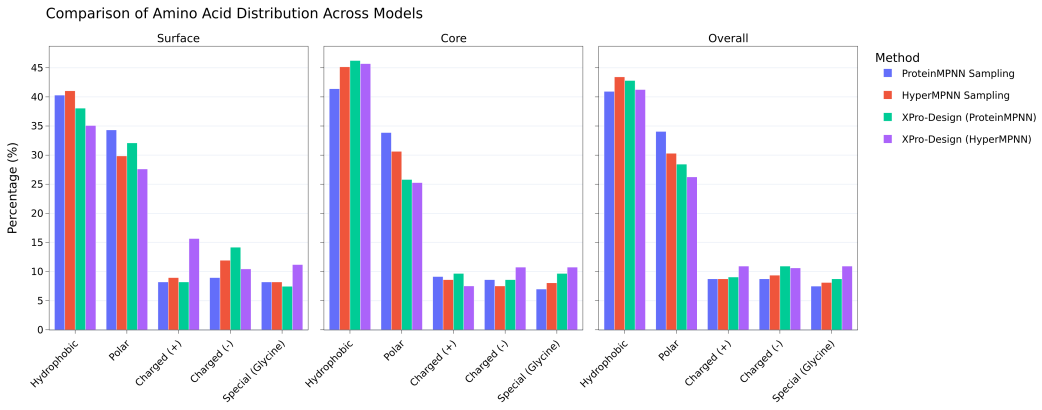

Figure 9: Surface vs Core AA Type distribution from the different methods

A.6 Entropy Reduction

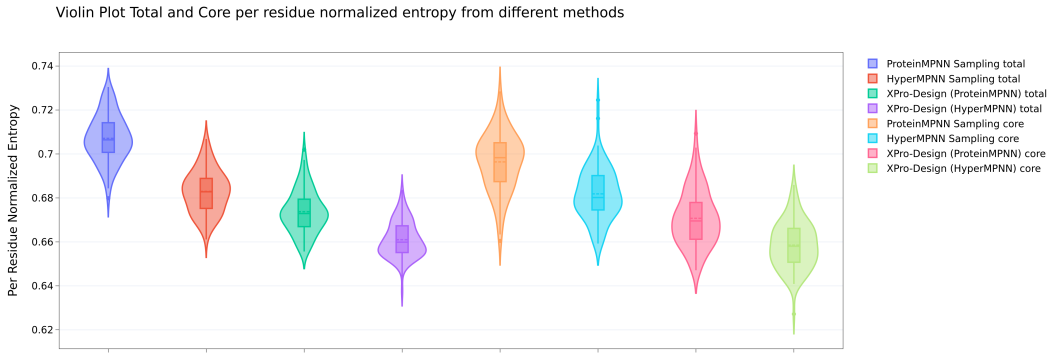

Figure 10: Normalized Per Residue wise Entropy for the full protein and the core

A.7 SorA Results

Despite the SorA protein being a small protein having only 124 AAs with not much scope towards optimization since it is already a hyper thermophile, we observe a clear and similar trend here as well. Our XPro-Design method considerably outperforms even the finetuned HyperMPNN model at designing more thermostable variants. This is clear by the clear upward shift in the predicted melting temperatures from different methods as well as the MM/GBSA based energy calculations.

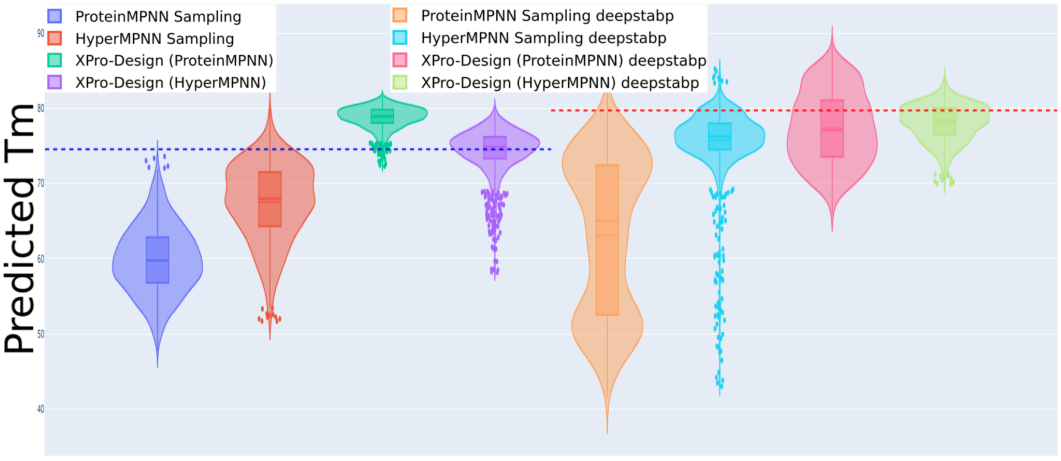

Figure 11: Predicted melting temperature values for SorA variants in °C for the different methods using TemBERTure and DeepSTABp models. We Observe the spread of High-Tm sequences is considerably narrower for XPro-Design compared to even HyperMPNN which was specifically trained to generate thermophile variants.

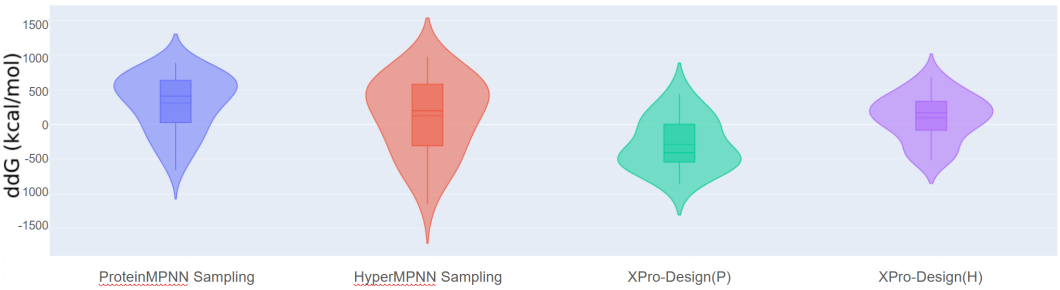

Figure 12: Violin plots showing $\Delta\Delta G$ distributions of variants generated from the different methods computed against SorA WT

Similar to trends seen with the CalB target, all the generated variants from all the methods folded correctly despite having sequence coverage of only around 53% across all methods.

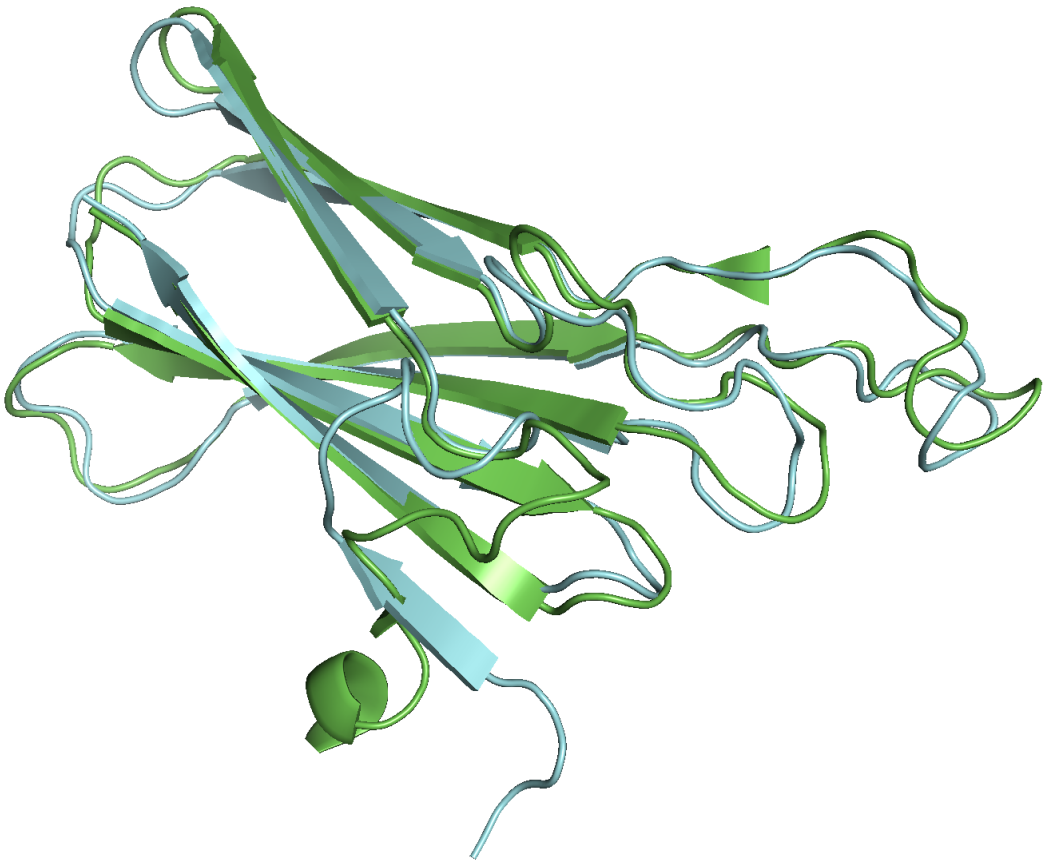

Figure 13: Top SorA designed variant V_5984 (Xpro-Design(P) in Blue overlaid over SorA WT 1DQI

A.8 LLM Use Disclaimer

The authors used a large language model (ChatGPT, OpenAI) to assist in polishing grammar and improving conciseness of the manuscript text. The model was not used for data analysis, generation of scientific content, or drawing conclusions. All scientific content and interpretations are solely the responsibility of the authors.