Decoding the physicochemical basis of taxonomy preferences in protein design models

Laura Dillon[1,2] Aaron Maiwald[3] Oliver Crook[1,2]

[1]Department of Chemistry, University of Oxford, Oxford, UK [2] Kavli Institute for Nanoscience Discovery, University of Oxford, Oxford, UK[3] Department of Statistics, University of Oxford, Oxford, UK

Abstract

Protein design models have transformed protein engineering by enabling computational exploration of sequence spaces far exceeding experimental capacity. Yet different models applied to identical structures generate divergent outputs. These systematic differences reflect training data characteristics rather than random variation, but their patterns and malleability remain uncharacterised. We quantified how training data modality shapes systematic preferences across six architectures. Structure-conditioned models retain minimal unexplained taxonomic variance (<3%) after controlling for measurable protein properties, while sequence-only models show substantial residual taxonomic dependence (19-25%). This difference stems from distinct preference patterns: Structure-conditioned models favour compactness and stability, properties enriched in experimentally tractable structures; sequence-only models bias reflect organism identity more heavily when evaluating sequence quality. These systematic preferences constrain design space exploration in predictable ways. Generated sequences converge toward model-specific biophysical profiles, with both input template properties and training-derived preferences jointly determining outcomes. Fine-tuning demonstrates these patterns are malleable: training ProteinMPNN on alkaline-adapted proteins redirects preferences toward elevated isoelectric points and basic residue content while preserving structural compatibility. Systematic bias characterisation enables both prediction of design outcomes and strategic modification for target applications. As protein design tackles increasingly complex challenges, understanding how training procedures encode systematic preferences becomes essential for reliable engineering.

Main

The central challenge in protein engineering is determining which amino acid sequences will reliably fold into stable structures and carry out specific functions. This sequence to structure to function relationship underlies enzyme catalysis, signalling, and immunity, but the astronomical size of sequence space makes direct exploration infeasible [1–3]. For much of the last century, this "Protein Folding Problem" [4, 5] remained largely intractable [6]. To design new proteins, researchers modified existing structures through rational design and directed evolution, or relied on computationally expensive physics-based energy functions [7–12].

The recent confluence of high-throughput experimental techniques and computational resources has transformed the field from localised, hypothesis-driven modifications to global, data-driven exploration powered by machine learning [7, 13–20]. Deep learning techniques now underpin most cutting-edge protein engineering methods. However, this advancement involves trading traceable, stepwise logic for potentially more powerful but opaque neural networks. While models capture sequence-structure relationships with accuracy comparable to experimental methods, researchers cannot fully determine how models arrive at predictions, or whether learned patterns represent genuine biological principles rather than artifacts of training data composition [21].

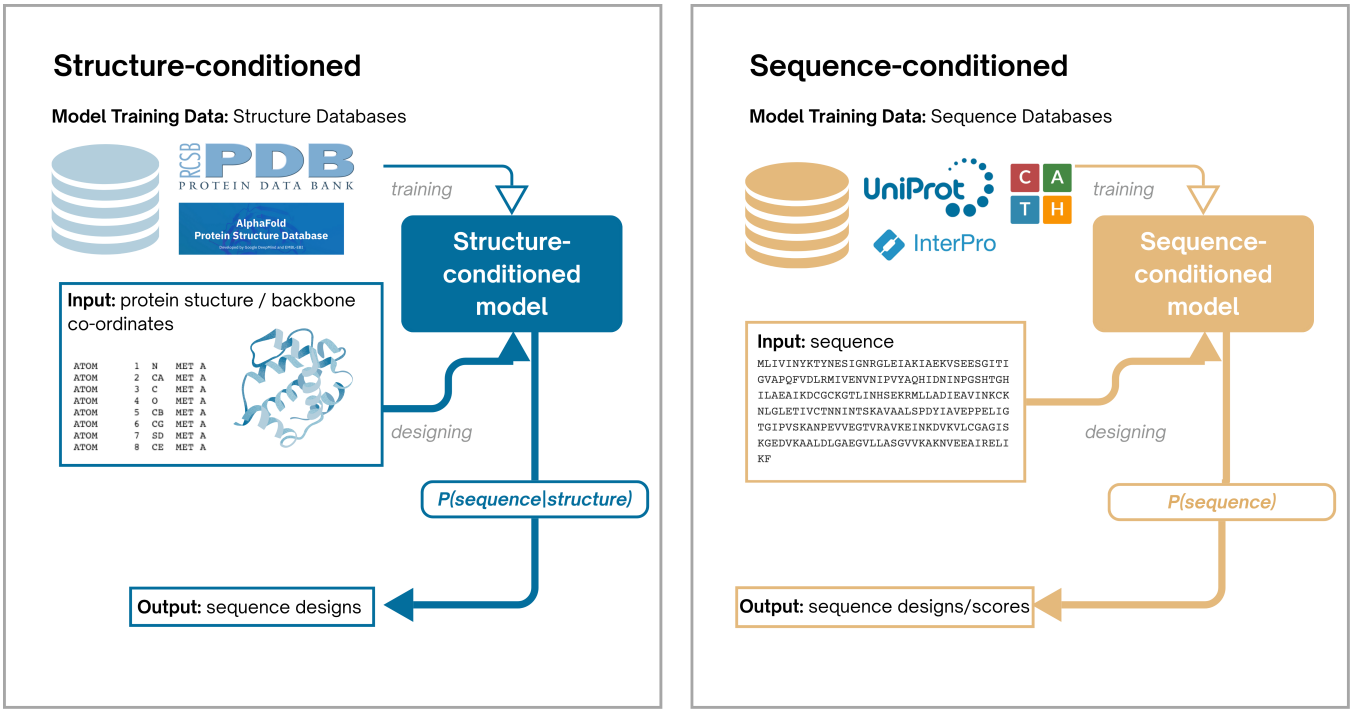

Figure 1: Training modalities of structure- and sequence-conditioned protein design models. Left, structure-conditioned models such as ProteinMPNN and ESM-IF are trained on structural databases (e.g., PDB, AlphaFold DB) and learn the conditional probability $P(\text{sequence} \mid \text{structure})$, predicting amino acid sequences compatible with a given backbone. **Right**, sequence-conditioned models such as ESM2 and CARP-640M are trained on sequence databases (e.g., UniProt, CATH) and learn the unconditional probability $P(\text{sequence})$, capturing statistical regularities across evolution. This difference in data modality, 3D structure versus linear sequence, determines what information each model encodes and constrains the kinds of biases they may develop.

Such interpretability gaps make evaluation metrics that capture biological implications of outputs essential. Standard metrics like perplexity or sequence recovery summarise fit to training-like data but obscure model *biases* that matter for design [22]. Similar issues have long existed across other machine learning domains, from computer vision [23], to natural language processing [24–27], such as the demographic disparities in facial recognition [28], and existing inequalities re-enforced by healthcare algorithms for treatment recommendations [29].

In the context of protein design, we define bias as a systematic, reproducible deviation in model scoring or generation that reflects training data properties or architectural constraints rather than biological fitness. Unlike natural language processing or image recognition where users can quickly identify problematic outputs, protein design biases typically remain hidden until resource-intensive experimental validation reveals them.

Ding and Steinhardt (2024) demonstrated that ESM2 and other protein language models (pLMs) assign consistently higher likelihoods to proteins from model organisms such as *E. coli* and *H. sapiens,* penalising functionally comparable proteins from underrepresented species [30]. This bias persists despite masked language modelling treating each sequence independently, indicating that training data composition shapes model preferences in ways architecture cannot fully overcome [31]. The implication is that species effects in sequence-only models are present even when models appear accurate on aggregate. Similarly, Shaw et al. (2023) quantified how sequence-only pLMs limit design diversity by converging on high-frequency amino acid patterns present in training data [32]. Using entropy and diversity metrics across generated sequences, they showed that pLMs systematically under-sample rare residues and structural motifs, reducing the effective novelty of designs despite high perplexity scores. This demonstrates that database skew constrains not only species representation but also the accessible biochemical diversity of designs.

Similar to how biases in other ML domains are largely linked to skewed datasets the models trained on, bias in protein design largely reflects the model's training modalities. Current protein design models fall into two categories: sequence-only models trained on sequence databases, and structure-conditioned models trained on structure databases

that interpret input as three-dimensional coordinates. This data modality difference may shape what biases models develop (Figure 1).

It is unclear if the biases seen in pLMs are also affecting structure-conditioned models. Rather than learning patterns exclusively from sequences, structure-conditioned models directly incorporate three-dimensional backbone geometry to predict compatible amino acid sequences that will fold into the target structure [2, 13, 15, 33, 34]. Protein structure often shows stronger evolutionary conservation than sequence, evidenced by numerous examples of convergent evolution producing similar folds from divergent sequences [35–37]. This approach could theoretically circumvent taxonomic limitations. Because of the structure-conditioning, models like ProteinMPNN [38], ESM-IF [39], and MIF [40] potentially learn more generalisable folding principles rather than organism-specific sequence patterns.

Yet structural databases inherit distinct biases. The Protein Data Bank (PDB) contains structures which have been experimentally determined, which are typically rigid, thermostable molecules[41, 42]. Membrane proteins comprise approximately 20-30% of proteomes [43] but represent less than 3% of PDB structures [44], and intrinsically disordered regions, which play crucial roles in signalling and regulation, remain poorly characterised [45–47]. Structure-conditioned models trained on these datasets might inherit preferences for crystallisable proteins, potentially limiting their ability to design functionally diverse sequences.

These distinct bias patterns linked to training modality raise our central questions: Does incorporating structural information reduce taxonomic biases compared to sequence-only approaches? What biases do structure-conditioned models exhibit that are particular to their architecture? And can systematic characterisation of these patterns enable strategic model selection and modification?

Results

Structure-Conditioned Models Show Reduced Taxonomic Bias

To evaluate taxonomic bias across protein design models, we curated a dataset of approximately 8,000 proteins spanning 144 species across the three domains of life: eukaryotes (organisms with membrane-bound nuclei, including humans, yeast, and plants), bacteria (single-celled organisms without nuclei, such as *E. coli*), and archaea (single-celled organisms often found in extreme environments like hot springs). This dataset provides broad evolutionary coverage, sampling proteins from thermophilic organisms thriving at high temperatures, mesophilic organisms adapted to moderate conditions, and species ranging from microscopic prokaryotes to complex multicellular organisms.

We evaluated six models representing different training approaches (Supplementary Table S2): purely structure-conditioned models trained on experimental data (ProteinMPNN on PDB), structure-conditioned models trained on predicted structures (ESM-IF on AlphaFold DB), hybrid models using curated sequence-structure pairs (MIF on CATH—a hierarchically organised database of 19,000 protein domain structures classified by fold similarity), hybrid models with sequence transfer (MIF-ST combining CATH with UniRef50 embeddings from 42 million sequences), and sequence-only models (ESM2 on UniRef90/50, CARP-640M on UniRef50)[38–40, 48, 49]. Model outputs are log-likelihood values, which are numerical scores where higher values indicate the model assesses a protein sequence as more compatible with its learned distribution.

We partitioned total score variance into factors attributable to groups of related proteins sharing function, measurable biophysical characteristics like size, charge, and structural stability, structural quality, and species identity. Additional R^2 quantifies the proportion of variance explained by species after accounting for all other factors, providing an estimate of residual taxonomic bias—variance that cannot be attributed to any measured protein characteristics. Partial R^2 measures the semi-partial correlation between species and model scores relative to unexplained variance, showing the magnitude of species effects before covariate adjustment.

Table 1: Species Variance Decomposition Across Protein Design Models. Species Additional R^2 represents variance explained by species after accounting for protein family, biophysical properties, and structural quality—values below 5% indicate minimal taxonomic bias. Species Partial R^2 shows total species-score correlation before controlling for these factors. Training modality types: Structure (backbone coordinates only), Hybrid (sequence + structure), Hybrid-ST (hybrid with sequence transfer from pre-trained embeddings).

Model	Architecture	Add. R^2	Partial R^2
ProteinMPNN	Structure	2.21%	27.46%
ESM-IF	Structure	1.12%	17.06%
MIF	Hybrid	2.80%	30.08%
MIF-ST	Hybrid-ST	11.27%	43.05%
ESM2-15B	Sequence	19.47%	33.91%
CARP-640M	Sequence	21.42%	52.0%

This variance reduction has practical implications prediction reliability. When ProteinMPNN assigns a low score to a designed sequence, that score primarily reflects biophysical incompatibility with the backbone structure (only 2.21% unexplained by measurable properties), making it a more interpretable quality signal than ESM2's scores, which retain 19.47% unexplained variance tied to taxonomic origin. However, ProteinMPNN's low unexplained variance stems from reparameterising database bias through structural features. In contrast, ESM2's 19.47% unexplained species variance means a low score could indicate either poor sequence quality or that the protein originated from an underrepresented species—the two signals are confounded, reducing score interpretability for design decisions.

After controlling for protein family and biophysical covariates, structure-conditioned models explain less than 3% of score variance through species identity alone, comparable to measurement noise. Sequence-only models retain substantial taxonomic dependence (19-25% unexplained variance), suggesting these architectures learn species-specific patterns that persist even when proteins share similar functions and structural properties. Hybrid models occupy an intermediate position, with MIF showing structure-like performance (2.80% additional R^2) while MIF-ST retains elevated species variance (11.27%) (Table 1).

The collapse from high partial R^2 to low additional R^2 demonstrates that apparent species correlations reflect measurable biophysical differences rather than systematic taxonomic bias. Before covariate adjustment, all models show substantial species correlations (partial R^2 ranging from 17-61%), initially suggesting strong taxonomic preferences. However, controlling for measurable protein properties eliminates most species-related variance in structure-conditioned models, revealing that these correlations arise from biological differences between taxonomic groups (archaeal proteins tend toward higher thermostability; eukaryotic proteins tend toward larger size and lower compactness) rather than learned database artifacts. Sequence-only models retain 19-25% unexplained variance even after these adjustments, indicating true taxonomic bias independent of protein characteristics.

The MIF versus MIF-ST comparison isolates how pre-trained sequence embeddings reintroduce taxonomic skew. Both models use identical architectures and train on CATH's taxonomically balanced structural data, yet MIF-ST incorporates sequence embeddings from CARP-640M, which was pre-trained on UniRef50's 42 million sequence, which is a database heavily skewed toward highly researched organisms. The substantial increase in species variance (2.80% to 11.27%) demonstrates that incorporating representations might override balanced structural training.

Structure-conditioned models show taxonomic patterns that correlate with measurable biophysical properties. While sequence models exhibit unpredictable accuracy variations across species, such as performing well on *E. coli* and poorly on functionally equivalent proteins from underrepresented organisms, structure-conditioned models show consistent performance patterns based on measurable characteristics. Based on these results, researchers may anticipate that ProteinMPNN will favour compact proteins with low instability index, regardless of taxonomic origin, enabling strategic model selection based on target protein characteristics rather than evolutionary relationships.

Taxonomic Preferences Invert Between Training Modalities

Structure-conditioned and sequence-only models show opposing taxonomic preferences reflecting their training data composition. To quantify directional taxonomic biases, we implemented a quality-weighted ELO rating system adapted from chess rankings [50]. ELO ratings dynamically update based on head-to-head performance comparisons according to:

$$R'_A = R_A + K(S - E_A) \tag{1}$$

where R_A is species A's current rating, $K = 32$ is the update magnitude, S is the actual outcome (1 if species A wins, 0 if it loses), and $E_A = 1/(1 + 10^{(R_B - R_A)/400})$ is the expected outcome based on current ratings. If a model scores a protein from species A higher than one from species B within the same protein family, species A gains rating points while species B loses points, with larger changes occurring when outcomes go against expectations (e.g., when a low-rated species defeats a high-rated species).

Within each protein family, we conducted pairwise comparisons ranking species by model scoring preferences, weighted by structural confidence (pLDDT: predicted Local Distance Difference Test, a per-residue confidence score from AlphaFold ranging 0 to 100 where more than 70 indicates high-confidence predictions) to account for input quality disparities. This analysis reveals systematic taxonomic preferences that align with training modality differences across seven models (Table 2).

Table 2: Taxonomic Preferences by ELO Rating. Mean ELO ratings across taxonomic domains, with higher values indicating stronger model preference. The highest-rated domain for each model is bold. Structure-conditioned models consistently favour Archaea and Bacteria; sequence-only models favour Eukaryota. All domain-level differences are statistically significant ($p < 0.001$, chi-square test.)

Training Data	Model	Archaea	Bacteria	Eukaryota
Structure-only CATH (Seq+Struc)	ProteinMPNN	**1693**	1655	1247
	ESM-IF	1526	**1695**	1235
	MIF	**1668**	1589	1340
CATH + UniRef50	MIF-ST	1163	1435	**1671**
Sequence-only	ESM2-15B	1285	**1592**	1423
	CARP-640M	961	1497	**1639**

Table 3: Species-level preferences inferred from top-ranked taxa. Representative species enriched in each model's top-10 ELO rankings illustrate how training data shape systematic preferences.

Training Data	Model	Species-level Preference
Structure-only	ProteinMPNN	Favors archaeal thermophiles (*T. kodakarensis*, *P. horikoshii*) and compact extremophiles.
	ESM-IF	Enriched for pathogenic/model bacteria (*E. coli*, *Salmonella*, *Pseudomonas*).
CATH (Seq+Struc)	MIF	Tilt toward thermophilic extremophiles, with some pathogenic bacteria.
CATH + UniRef50	MIF-ST	Sharp shift to eukaryotic mammals and domesticated species
sequence-only	ESM2-15B	Dominated by enterobacteria pathogens (*Shigella*, *Yersinia*, *Salmonella*).
	CARP-640M	Strong bias toward vertebrate models

Species-level analysis reveals distinct biases within broad taxonomic preferences (Table 3; statistical statistical details in Supplementary Tables S3–S4). Although ProteinMPNN and ESM-IF show broadly similar rankings at the domain level, their species-level preferences diverge. Across all species the rankings are highly correlated (Spearman $\rho = 0.90$,

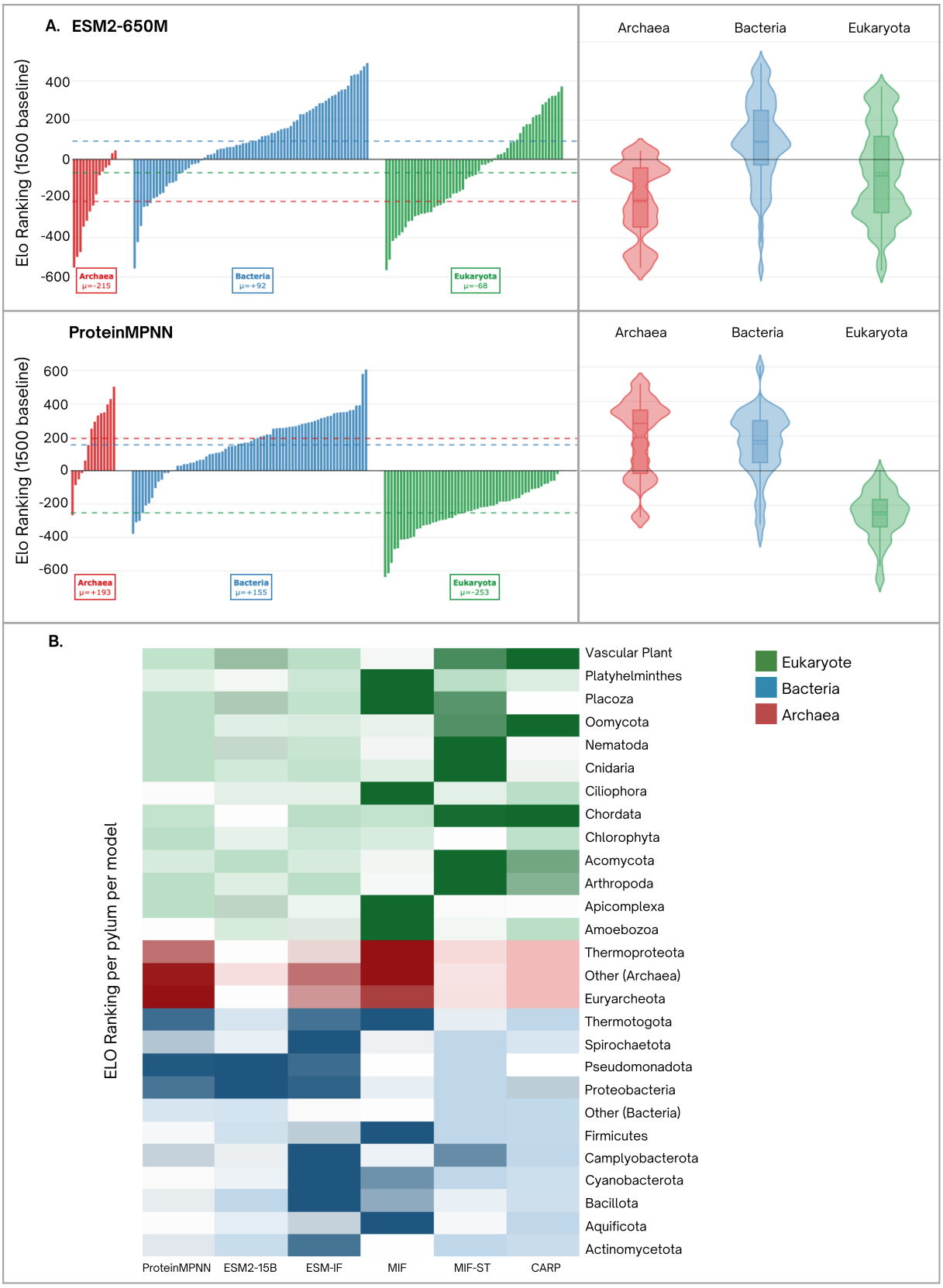

Figure 2: Training modality is linked to taxonomic preference direction. **A** ELO ratings for ProteinMPNN (structure-conditioned) versus ESM2 (sequence-only) across 144 species, coloured by domain. ProteinMPNN favours thermophilic archaea and bacteria; ESM2 favours well-studied pathogens and model organisms. The 446-point archaeal rating difference demonstrates training data composition dominates preference patterns. All species have mean pLDDT > 70. **B** Heatmap of per-phylum ELO ratings across models; darker shading indicates higher preference.

$p < 10^{-52}$), yet the overlap of their top ten favoured species is less than 20%. ProteinMPNN favouring archaeal thermophiles and ESM-IF pathogenic and model bacteria. The enrichment of pathogenic bacteria in ESM-IF's

top-ranked species is consistent with its training corpus composition. AlphaFold DB structures are predicted from UniRef50 reference proteomes, which over represent model organisms and disease relevant bacteria such as *E. coli*, *Salmonella*, and *Pseudomonas* [31, 51]. Because AlphaFold itself was trained on the PDB, these predictions likely learn both crystallographic biases (archaea preference) and UniRef's taxonomic skew, compounding both influences in ESM-IF.

Structure-conditioned models favour archaeal and bacterial proteins, while sequence-only models prefer eukaryotic species. Structure-only models consistently rank Archaea and Bacteria highest, whereas sequence-only models show the opposite pattern. CARP-640M assigns Eukaryota 678 points above Archaea and penalises Archaea severely relative to Bacteria (537-point deficit), indicating that CARP treats archaeal proteins as fundamentally incompatible with its learned sequence distributions. Hybrid models again occupy intermediate positions, with MIF resembling structure-only preferences while MIF-ST exhibits a strong eukaryotic preference inherited from its sequence embeddings.

Sequence transfer decisively reshapes taxonomic preferences despite architectural similarity. MIF and MIF-ST have nearly identical architectures and train on the same CATH structural data, yet show low agreement at species level (ρ = –0.17, top-10 overlap = 0%). Instead, MIF-ST aligns strongly with CARP-640M (ρ = 0.67, top-10 overlap = 54%) from which it inherits sequence embeddings, while MIF correlates negatively with CARP (ρ = -0.60). This demonstrates that pre-trained sequence representations from taxonomically skewed databases override balanced structural influences, importing UniRef50's compositional biases through transfer learning.

These directional preferences support our variance decomposition findings that training modality fundamentally shapes how models evaluate proteins from different evolutionary origins. All models show highly significant domain-level effects, indicating systematic rather than random taxonomic biases. ProteinMPNN assigns high scores to proteins with extremophile-associated characteristics such as compactness and low instability. Whether this reflects learned associations between these features and designability, training data composition, or actual fitness advantages cannot be determined from scoring patterns alone.

Biophysical Properties Explain Apparent Taxonomic Bias

To identify which protein characteristics drive the taxonomic preferences observed in structure-conditioned models, we analysed correlations between model scores and biophysical features, then used principal component analysis to visualise preference landscapes in low-dimensional property space.

Principal component analysis on eleven sequence-derived and structure-derived properties revealed that model preferences operate along interpretable biophysical axes. Two principal components captured 51.4% of total variance (Figure 3; complete variance analysis across all feature spaces in Supplementary Table S6 and feature loadings in Supplementary Tables S7–S8), indicating that apparent taxonomic biases reflect systematic preferences along a small number of measurable dimensions rather than complex, high-dimensional patterns.

PC1 (31.6% variance) represents a compactness-size axis. Structural centralization (inverse of average Cβ distance), structural compactness (inverse of radius of gyration), and relative contact order all loaded negatively, while sequence length loaded positively (Figure 3B). Proteins with negative PC1 values are compact, centralized, and enriched in long-range contacts; proteins with positive PC1 values are extended and longer. PC2 (19.8% variance) represents a hydropathy-stability axis. GRAVY (grand average of hydropathy) loaded positively (+0.575) and instability index loaded negatively (-0.557), such that positive PC2 values indicate proteins with higher average hydrophobicity and lower predicted instability, while negative PC2 values indicate more hydrophilic compositions and higher predicted instability.

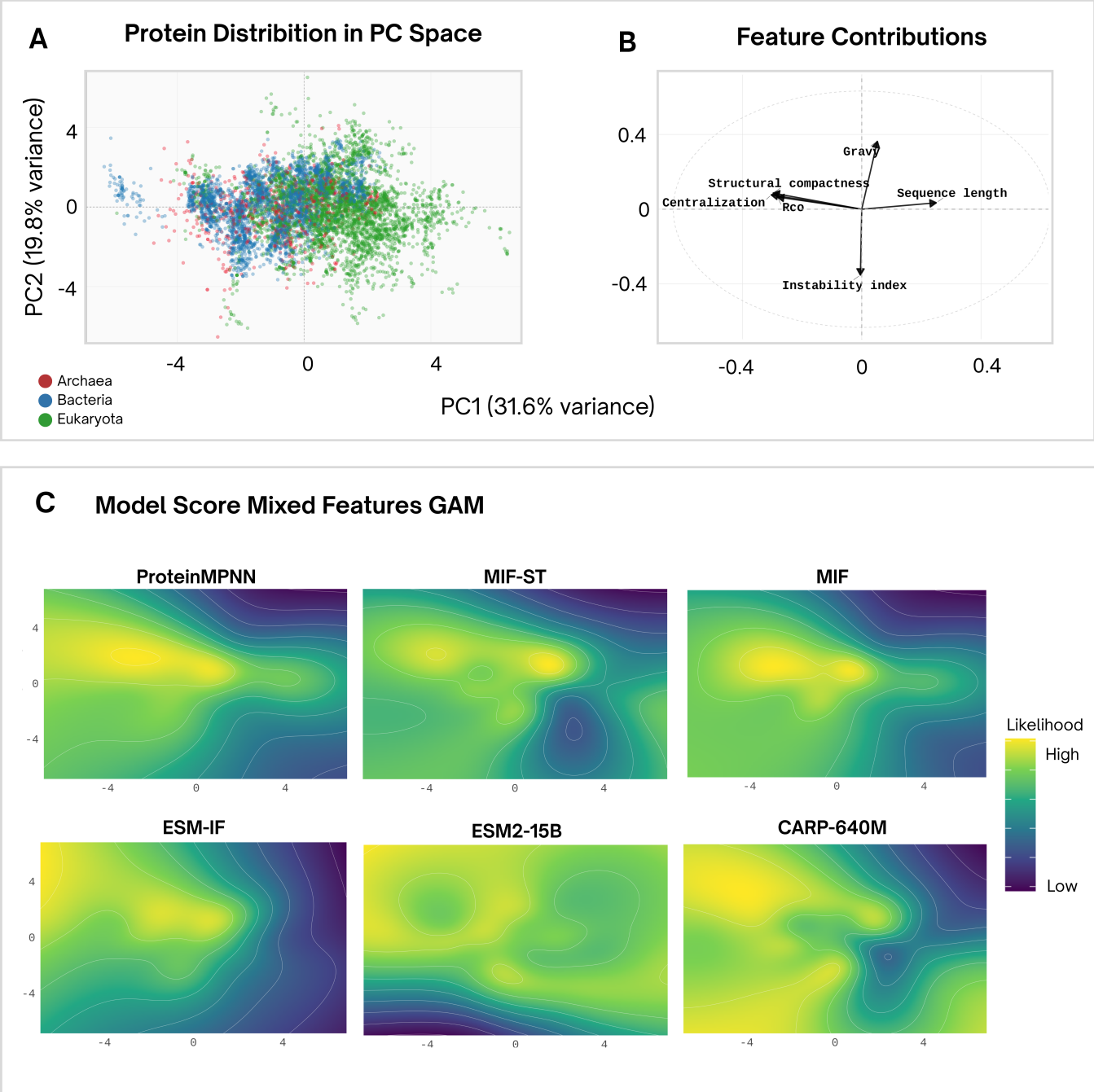

Figure 3: Model preferences mapped onto biophysical principal component space. **A.** Proteins (n=7,843) coloured by taxonomic domain in PC space derived from 11 biophysical features. Shows how proteins from different domains occupy distinct regions of biophysical space.**B.**PC1 (31.6% variance): compactness-size axis; negative values correspond to compact, centralized proteins with high relative contact order, positive values to extended, longer proteins. PC2 (19.8% variance): hydropathy-stability axis; positive values correspond to higher GRAVY (hydrophobicity) and lower instability index, negative values to lower GRAVY and higher instability. Archaeal proteins cluster toward compact regions (negative PC1); eukaryotic proteins extend toward larger, less compact regions (positive PC1). This separation explains how structure-conditioned models' biophysical preferences manifest as apparent taxonomic biases. Feature loadings show how measured properties contribute to each principal component. **C.** GAM surfaces showing likelihood scores across PC space for protein design models; warmer colours indicate higher scores. Demonstrates that models have distinct preference maxima in biophysical space. Structure conditioned ProteinMPNN has a peak in compact-stable quadrant (negative PC1/PC2), favouring proteins with high relative contact order and low instability index. Sequence conditioned ESM2 shows broader, orthogonal preferences extending toward larger, more hydrophobic proteins (positive PC1/PC2). Contour lines mark score deciles. The 180° rotation between preference maxima explains inverted taxonomic rankings in Table 2.

Individual feature contributions were relatively balanced: centralization (7.6%), structural compactness (7.4%), GRAVY (6.8%), RCO (6.5%), and instability index (6.2%). No single property dominated the principal components, indicating that model preferences integrate multiple complementary signals rather than optimising for isolated features.

The ability to explain 51% of preference variance in two dimensions indicates that model biases operate in low-dimensional property spaces despite proteins' high-dimensional sequence complexity. This low dimensionality suggests systematic preferences reflect learned associations between small numbers of biophysical features rather than complex, emergent patterns across many protein properties.

Structure-conditioned models show strong sensitivity to structural confidence and compactness measures. Correlations between model likelihoods and protein properties revealed distinct patterns by training modality. ESM-IF (ρ=0.63) and ProteinMPNN (ρ=0.60) showed the strongest associations with AlphaFold confidence (pLDDT), followed by hybrid models MIF-ST (ρ=0.53) and MIF (ρ=0.52). ProteinMPNN additionally correlated negatively with molecular weight per residue (ρ=-0.40) and instability index (ρ=-0.40). Sequence-only ESM2 exhibited weak pLDDT correlation as expected. Generalised additive model fitting revealed distinct preference landscapes (Figure 3). ProteinMPNN assigned highest scores to the compact, low-instability, high-hydropathy region (negative PC1, positive PC2). ESM2 showed orthogonal preferences, assigning highest scores toward positive PC1 and positive PC2. These preference patterns remained consistent across protein families, indicating universal biophysical filters rather than family-specific criteria. For sequence-only and structure-only PCA analyses, see Supplementary Figures S2–S5.

Taxonomic groups project onto this biophysical space in characteristic patterns. Archaeal proteins cluster toward negative PC1 (compact), bacterial proteins show intermediate positions, and eukaryotic proteins extend toward positive PC1 (larger, extended). This distribution explains how structure-conditioned models' compactness preferences manifest as apparent taxonomic biases: favouring compact proteins implicitly favours Archaea and certain bacteria, while favouring larger proteins implicitly favours Eukaryota. The ability to explain 51% of preference variance in two dimensions indicates that model biases operate in low-dimensional property spaces despite proteins' high-dimensional sequence complexity.

Having established that taxonomic preferences reflect measurable biophysical tendencies, we tested whether these patterns actively constrain design space exploration by analysing sequences generated for 16 diverse template proteins.

Designed Sequences Show Model-Specific Biophysical Tendencies

Generated sequences exhibit model-specific biophysical tendencies that are modulated by template characteristics, demonstrating that design behaviour reflects both biases and template-specific optimisation. To test whether identified biases actively influence sequence generation, we designed sequences for 16 diverse proteins spanning taxonomic and biophysical space. These templates varied in size (5.7–73.5 kDa), compactness (relative contact order 0.05–0.25), and secondary structure content, representing both model-preferred and model-disfavoured species based on ELO ratings (Supplementary Table S11). For each template, we generated approximately 48 designs per model and analysed systematic modification patterns.

We computed a principal component analysis over six sequence-derived features (molecular weight per residue, isoelectric point, instability index, GRAVY, sequence length, aromaticity) for wild-type proteins, then projected designed sequences into this space. For each template, we calculated the wild-type-to-design shift vector and averaged these vectors across templates to obtain model centroid shifts characterising systematic biophysical tendencies (Figure 4, supplementary Table S12).

Structure-conditioned models exhibit distinct directional tendencies in principal component space. ProteinMPNN produces designs shifted toward lighter, less aromatic compositions with modestly increased hydropathy, consistent with its bias towards for compact, efficiently-packed proteins observed. MIF shifts toward increased hydropathy and stability while lowering isoelectric point, suggesting a core-optimising strategy. MIF-ST drives toward increased aromaticity and residue mass, likely reflecting its hybrid sequence-structure training that incorporates signals. ESM-IF tends toward increased charge and hydrophilicity, though with substantial across-protein variability. These directional

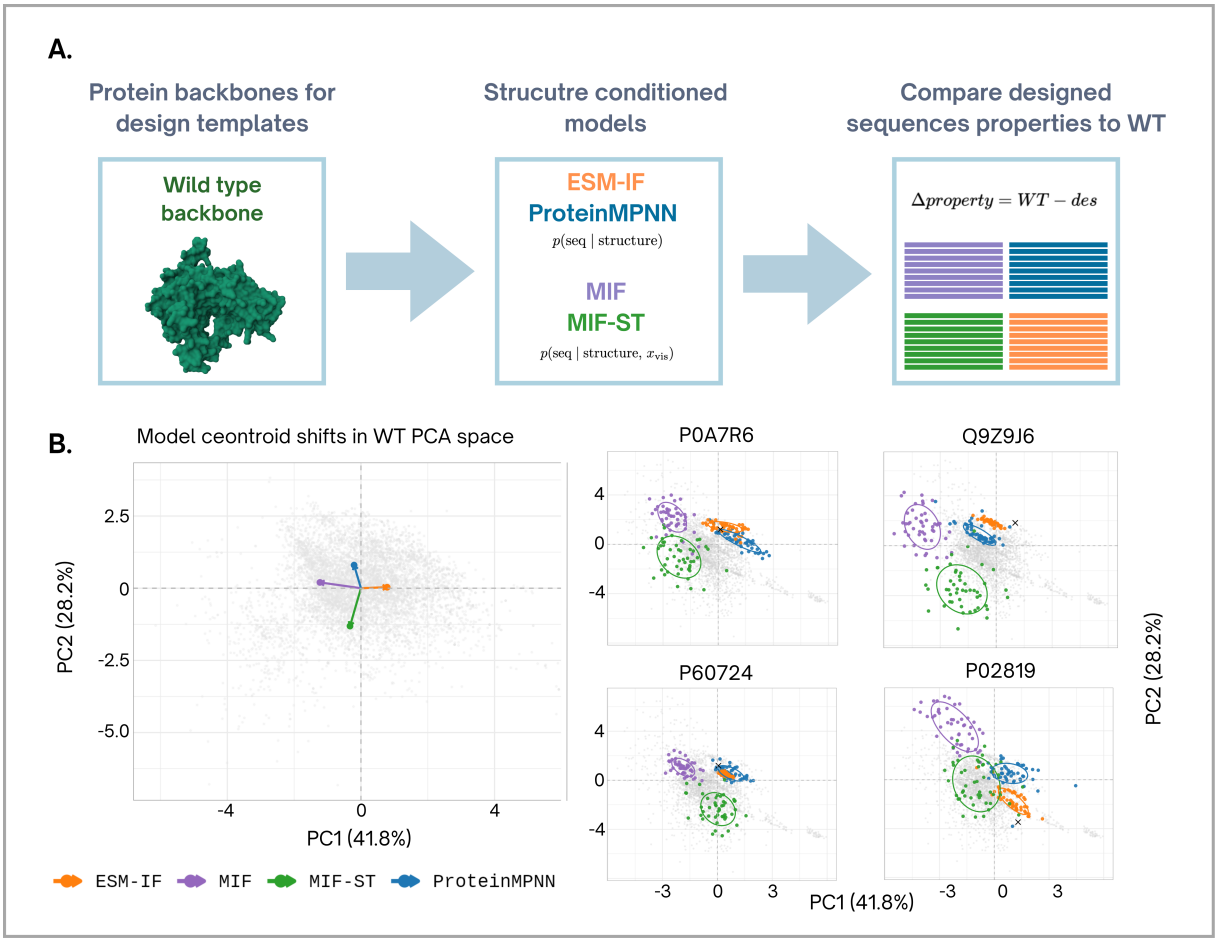

Figure 4: Model-designed sequences occupy distinct regions of wild-type sequence space. **(A)** Experimental workflow: each structure-conditioned generative model (ESM-IF, ProteinMPNN, masked inverse folding models MIF and MIF-ST) receives a wild-type protein backbone as input and generates sequences conditioned on that structure; designed sequences are then compared to the wild-type across multiple properties. **(B)** Principal component analysis of amino acid composition reveals model-specific biases. Left panel: centroid shifts. Arrows originating at WT mean, terminating at each model's mean. Right panels: four representative proteins (PDB entries PDA7R6, Q92536, P60724, P02819) show that designs cluster by model rather than diversifying around the wild-type sequence (black ×), with ellipses highlighting model-specific modes. This clustering pattern indicates that models biases and the input structure guide the design outcomes. Statistical details in supplementary Table S3.

tendencies are statistically significant at the population level for MIF, MIF-ST, and ProteinMPNN, though individual templates show considerable variation around these mean shifts ((Supplementary Table S12).

Template characteristics modulate but do not eliminate model-specific preferences. Cosine similarity analysis of per-protein wild-type-to-design shift vectors reveals weak-to-moderate pairwise agreement between models, with substantial variation across templates. The highest agreement occurs between MIF and MIF-ST, reflecting their shared architectural foundation. Structure-conditioned models show weak pairwise agreement, while ESM-IF shows near-orthogonal shifts relative to MIF-based models. This pattern indicates that while models exhibit systematic directional tendencies at the population level, template-specific factors such as size, topology, and native biophysical properties substantially modulate design directions. Design behaviour therefore emerges from interaction between learned biases and template-specific constraints.

The systematic directional shifts we observe map a predictive framework for anticipating model behaviour on unseen proteins. For applications where model tendencies align with design objectives, biases become advantageous and reduce the search space toward functional sequences. Conversely, when tendencies conflict with objectives, bias characterisation enables informed model selection or targeted interventions. The substantial template-dependent variation we observe suggests that single-protein evaluations may yield misleading conclusions about model behaviour;

systematic characterisation across diverse templates is necessary to separate population-level tendencies from protein-specific responses.

Fine-Tuning Demonstrates Bias Malleability

To test whether identified biases represent trainable parameters amenable to strategic modification versus immutable architectural constraints, we fine-tuned ProteinMPNN on alkaline-adapted proteins from organisms thriving at pH >9.0. Alkaliphilic organisms maintain near-neutral intracellular pH despite high external pH, with adaptations including modified surface charge distributions on proteins [52], particularly elevated isoelectric points and increased basic residue content to buffer against external pH. If fine-tuning can redirect model preferences toward these pH-optimised characteristics while preserving structural compatibility, this would demonstrate that biases represent malleable learned behaviours.

Table 4: Biophysical Property Shifts After Fine-Tuning on Alkaline-Adapted Proteins. Mean property values for sequences designed with vanilla ProteinMPNN versus fine-tuned Alkaline-MPNN across 16 diverse template proteins (48 designs per template). Alkaline adaptation produces systematic shifts toward pH-optimised characteristics, such as elevated isoelectric point, increased basic residue content, and positive net charge—while maintaining structural compatibility. Alkaliphile score is a composite metric incorporating these pH-relevant features (range 0-1, higher values indicate stronger alkaline adaptation). All changes significant at $p < 0.001$ (paired t-test with Bonferroni correction).

Property	Vanilla	Alkaline	Change
Isoelectric Point	7.47	8.95	+1.48
Net Charge at pH 7	-4.12	+19.13	+23.25
Basic Residue %	12.8%	24.2%	+11.4%
Acidic Residue %	12.3%	11.5%	-0.8%
Alkaliphile Score	0.228	0.364	+60%

Fine-tuning preserved structural conditioning's bias reduction while redirecting specific biophysical preferences. The fine-tuned model (Alkaline-MPNN) retained low unexplained species variance (3.43%), comparable to vanilla ProteinMPNN (2.21%), demonstrating that structural conditioning's taxonomic bias mitigation persists after targeted modification (Supplementary Table S15). Partial R^2 for species decreased slightly (16.07% vs 27.46%), indicating that fine-tuning may further reduce apparent taxonomic effects—possibly because alkaliphilic adaptations occur across multiple taxonomic groups. Overall taxonomic ordering remained consistent: Alkaline-MPNN scored archaeal proteins highest, followed by bacteria, then eukaryotes, preserving the structure-conditioned preference pattern. However, bacterial penalties decreased relative to archaea, possibly reflecting alkaliphilic adaptations being particularly common among bacterial extremophiles [52].

Alkaline adaptation fine-tuning significantly altered sequence design toward pH-optimised characteristics. Analysis of designed-versus-wild-type sequence pairs revealed systematic biochemical shifts. The most pronounced changes occurred in charge properties: isoelectric point increased from 7.47 to 8.95 (shifting from near-neutral to basic), basic residue fraction rose from 12.8% to 24.2% by increasing lysine and arginine content, while acidic content remained stable (12.3% to 11.5%). These modifications shifted net charge at pH 7 from -4.12 to +19.13 and increased the alkaliphile score by 60% (0.228 to 0.364, $p < 0.001$) (Table 4; detailed statistical comparisons in Supplementary Table S14).

Quantitative analysis in principal component space demonstrates targeted preference redirection. We performed PCA on pH-specific biophysical features (buffer capacity, charge per residue, basic/acidic/ionisable residue fractions) to create a projection space where principal components directly represent pH-relevant biochemical axes. GAM-based analysis measured how fine-tuning realigned ProteinMPNN's preference vectors with the pH loading vector, which points in the direction of increasing alkaline adaptation (Figure 5).

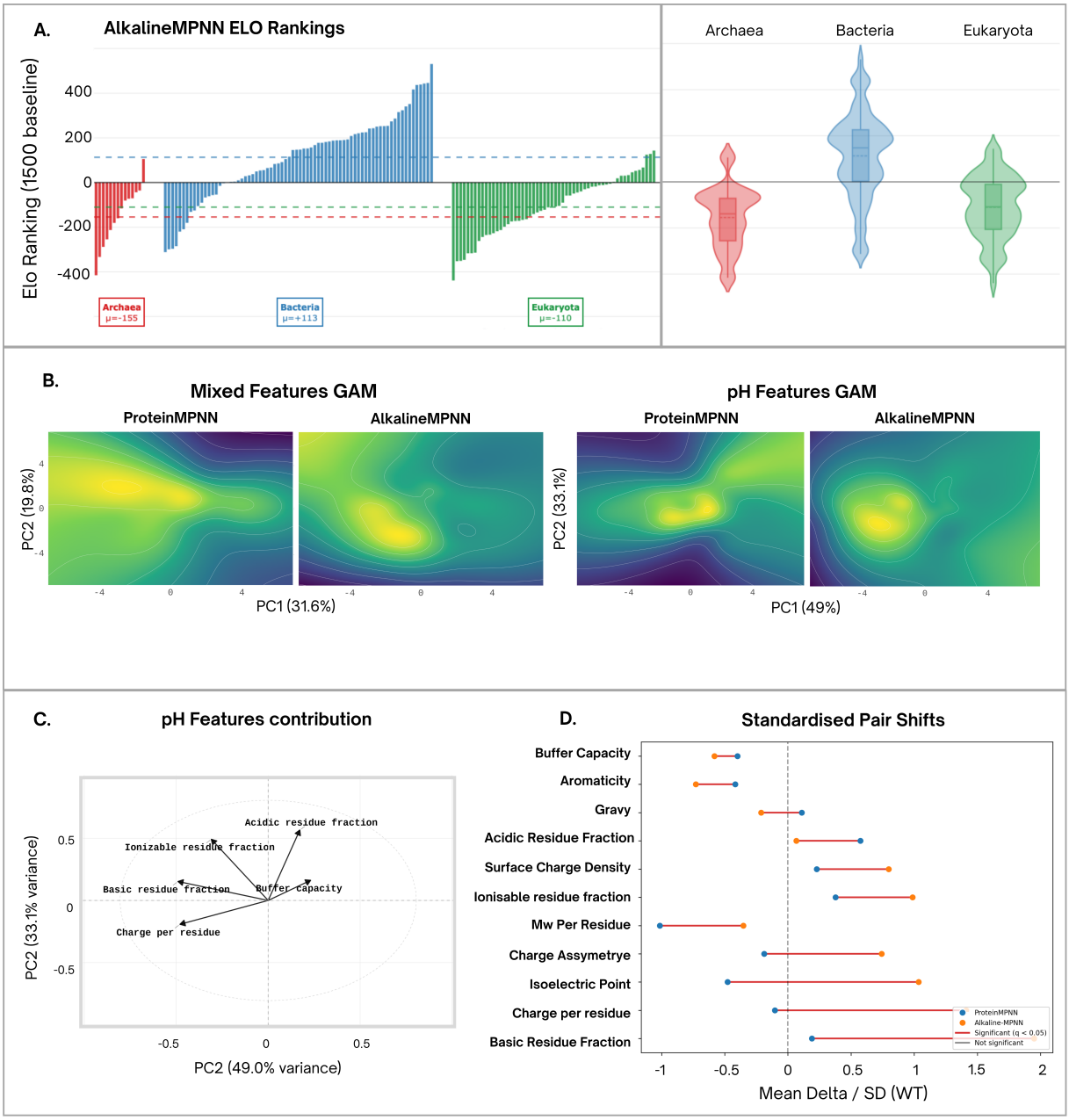

Figure 5: Biophysical Property Shifts After Fine-Tuning on Alkaline-Adapted Proteins. **(A)** ELO rankings for AlkalineMPNN across 144 species (left: individual species ranked by domain; right: domain-level distributions). Alkaline fine-tuning preserves the bacterial preference characteristic of structure-conditioned models while reducing the archaeal bias seen in vanilla ProteinMPNN. **(B)** GAM-fitted preference landscapes in mixed feature space (left panels) and pH-specific feature space (right panels), comparing vanilla ProteinMPNN (left of each pair) with AlkalineMPNN (right of each pair). Warmer colours indicate higher likelihood scores. Fine-tuning shifts the preference maximum toward alkaline-adapted regions in pH space while maintaining structural preferences in mixed feature space. **(C)** Feature loadings on pH-specific principal components, showing how buffer capacity, charge properties, and residue composition contribute to PC1 (49% variance) and PC2 (33% variance). **(D)** Standardised property shifts between vanilla ProteinMPNN (orange) and AlkalineMPNN (blue) designs across 16 template proteins (48 designs per template, $n = 768$ sequences per model). Properties with non-overlapping error bars indicate significant differences. AlkalineMPNN systematically increases isoelectric point, basic residue content, and charge asymmetry while maintaining other biophysical properties. All pH-relevant changes significant at $p < 0.001$ (paired t-test with Bonferroni correction).

Vanilla ProteinMPNN showed negative alignment with the pH loading vector (cosine similarity = -0.48, 95% CI: $[-0.50, -0.47]$), indicating systematic preference against high-pH proteins. AlkalineMPNN reversed this preference entirely (cosine similarity = $+0.46$, 95% CI: $[0.42, 0.50]$), representing a total preference shift of 0.91 units ($p < 0.0001$, posterior simulation with 300 iterations). This reversal actively disfavouring to actively favouring alkaline-adapted proteins demonstrates complete redirection of model preferences along the targeted biochemical axis.

Critically, this pH-specific redirection occurred without disrupting structural preferences. Analysis in structure-only

PC space (RCO, compactness, secondary structure composition) showed no significant difference between vanilla and fine-tuned models (Δcosine = -0.49, $p = 0.80$), supporting that fine-tuning modified sequence-level biochemistry while preserving structural compatibility constraints. This differential effect aligns with alkaline adaptation being primarily a surface chemistry phenomenon—proteins maintain similar three-dimensional folds while adjusting charge distribution, rather than requiring structural reorganisation.

Discussion

We have characterised bias as systematic, reproducible deviation in model scoring that reflects training data properties or architectural constraints rather than biological fitness. This definition differs from aggregate accuracy: models can achieve comparable performance on mutational effect prediction while exhibiting distinct systematic preferences for specific protein classes. Benchmark performance on ProteinGym's mutational scanning assays (Spearman 0.26–0.42) does not predict design space preferences [22]. ProteinMPNN, despite scoring lowest on this benchmark (0.26), may be preferable when its preferences align with application requirements. This disconnect indicates that mutational effect prediction and design space navigation represent distinct capabilities.

We present a comparative analysis of protein design models across training modalities, demonstrating that sequence-only and structure-conditioned architectures develop mechanistically distinct biases. Structure-conditioned models (ProteinMPNN, ESM-IF, MIF) trained on experimentally determined structures develop preferences for biophysical properties enriched in structure datasets, such as compact cores, high hydrophobicity, ordered secondary structure. These preferences produce low unexplained variance (2–3%), meaning most scoring variation can be attributed to measurable properties. Sequence-only models (ESM-2, CARP) trained on broader taxonomic coverage retain higher unexplained variance (19–25%) after accounting for biophysical factors, organising representations around sequence-level patterns that are not captured by the biophysical characteristics we analysed.

This mechanistic distinction has practical implications. Training data modality influences which regions of design space models systematically favour or disfavour, shaping which designs enter experimental pipelines. For applications where model preferences align with design objectives, this steering becomes advantageous. Yet, the same preferences may become obstacles when designing membrane proteins requiring conformational flexibility [53] or intrinsically disordered regions where lack of stable structure is functionally essential [45]. Derry et al. demonstrated this phenomenon, showing how models trained exclusively on X-ray structures systematically disfavoured lysine because its conformational entropy discourages crystallisation, not because lysine destabilises proteins [41]. The model learned experimental tractability as proxy for biological fitness.

Current validation practices may obscure these biases. When researchers synthesise top-scoring designs and validate their function, this confirms the model guided them to something functional but provides limited insight into whether alternatives with different biophysical properties were systematically down-weighted. A design receiving a low score might reflect protein fitness or the model's biases. Without systematic characterisation, distinguishing these possibilities remains challenging.

We demonstrate that these preferences may represent malleable learned behaviours rather than fixed architectural constraints. Fine-tuning ProteinMPNN reversed its pH-relevant preferences while maintaining structural compatibility metrics. This malleability suggests practitioners can strategically redirect specific bias dimensions when default behaviour conflicts with application requirements. However, simultaneously correcting multiple biases through fine-tuning may become impractical, and the extent to which this approach generalises across preference types requires further investigation.

The methodology we present enables characterising model preferences independently of aggregate performance metrics.

Variance decomposition partitions scoring variation into taxonomic, biophysical, and structural components. Principal component analysis combined with generalised additive models reveals how preferences vary across biophysical property gradients, while ELO rankings quantify relative preferences when models disagree on identical proteins. This framework enables profiling new models to build preference maps for informed selection and constructing targeted fine-tuning datasets when default behaviour conflicts with design requirements.

These considerations become increasingly important as model outputs are incorporated into design pipelines in ways that may be difficult to disentangle. Tan et al. showed that AlphaFold predictions occupy different conformational regions than experimental structures [54]. If models train substantially on predictions embedding earlier models' assumptions, and if predictions influence which proteins researchers prioritise for characterisation, feedback dynamics could potentially create reinforcing cycles where preferences appear more accurate than they are. Whether such dynamics meaningfully affect current models remains unclear, but warrants consideration as predicted-to-experimental structure ratios continue to grow.

The practical challenge lies in making model preferences sufficiently interpretable that researchers can account for them when biases might conflict with design objectives. Some preferences may reflect genuine biological principles, others likely reflect methodological filters in training data, most probably reflect complex combinations that cannot be fully decomposed. The value of bias characterisation lies in describing biases well enough to inform when models might guide appropriately for specific applications and when they might steer away from viable but underrepresented protein classes. As models scale in complexity, frameworks for characterising these biases become important for understanding how learned patterns shape which proteins are designed.

Methods

Dataset Construction

Protein Dataset

We constructed a protein dataset following the methodology established by Ding & Steinhardt (2024) [30]: species selection based on SwissProt representation, orthologous grouping by protein name and function, and the requirement for $\geq$ sharing orthologs per protein family to ensure phylogenetic breadth.

To ensure adequate representation across all three domains of life, we augmented the initial species pool with additional archaeal species, prioritising organisms that displayed extremophile characteristics. We retained orthologous protein groupings where functional annotations were consistent across species. For proteins from archaeal extremophiles with inconsistent functional annotations, we assigned proteins to existing functional groups based on sequence similarity while maintaining phylogenetic breadth.

Structure Retrieval. We initially attempted to use experimentally determined structures from the Protein Data Bank, but this reduced the dataset to too few proteins due to limited coverage across our taxonomic breadth, particularly for archaeal and extremophile bacterial species. To maintain phylogenetic diversity while ensuring structural information for all proteins, we used AlphaFold2 predicted structures from the AlphaFold Protein Structure Database [51, 55].

For each protein (identified by UniProt accession), we retrieved the corresponding AlphaFold predicted structure using a custom Python script that fetches PDB files from the AlphaFold database (`https://alphafold.ebi.ac.uk/`), caching structures locally to minimise redundant downloads. This yielded predicted structures for all 7,843 proteins in our dataset.

Dataset Composition. The final dataset comprised 7,843 proteins across 144 species, with domain distribution of 692 Archaea (8.8%), 3,232 Bacteria (41.2%), and 3,919 Eukaryota (50.0%). Proteins were organised into 229 functional families based on orthology. Phylogenetic coverage varied by family: 3,753 proteins (47.8%) belonged to families present in all three domains, 1,471 (18.8%) spanned two domains, and 2,619 (33.4%) were domain-specific. Sequence lengths ranged from 50 to 650 amino acids. Mean predicted Local Distance Difference Test (pLDDT) scores across the dataset were 89.0 ± 8.5, indicating generally high-confidence structural predictions. This fixed cohort was used consistently across all analyses.

Table 5 summarises the final dataset composition. This dataset was used for all analyses presented in this work unless otherwise specified (e.g., alkaliphile fine-tuning dataset, design evaluation subset of 16 proteins).

Table 5: Main Dataset Composition. This dataset was used for all analyses unless otherwise specified.

Metric	Archaea	Bacteria	Eukaryota
Species	14	74	56
Proteins	**692** (8.8%)	**3,232** (41.2%)	**3,919** (50.0%)
Protein families			
Total families			229
Three-domain families			3,753 proteins (47.8%)
Two-domain families			1,471 proteins (18.8%)
Single-domain families			2,619 proteins (33.4%)
Protein characteristics			
Sequence length range			24–2,531 amino acids
Mean pLDDT			89.0 ± 8.5
Mean isoelectric point			8.5 ± 2.8
Mean basic residue fraction			0.16 ± 0.06

Alkaliphile Protein Dataset

We constructed a curated dataset of alkaliphile proteins from the UniProt database [31] to fine-tune ProteinMPNN toward alkaline-adapted sequence characteristics. Starting with 35,218 reviewed SwissProt entries, we selected proteins exhibiting characteristics of alkaliphile adaptation based on isoelectric point, basic residue composition, and net charge.

Proteins were selected if they exceeded the 85th percentile in at least two of three criteria: pI $\geq$ 9.24, basic residue fraction $\geq$ 0.175, and net charge at pH 7.0 was $\geq$ 7.97. This yielded 4,760 candidate proteins.

We filtered candidates based on AlphaFold structure confidence [55], retaining proteins with mean pLDDT $\geq$ 70, high confidence residue fraction (pLDDT $>$ 70) $\geq$ 0.7, and very low confidence residue fraction (pLDDT $\leq$ 50) $\leq$ 0.3. Post-filtering, 4,097 proteins remained.

We partitioned proteins using CD-HIT [56] with 40% sequence identity threshold to group homologous sequences. Clusters were randomly assigned to training (70%), validation (15%), and test (15%) splits, ensuring no sequence pair across splits shared >40% identity. The final dataset comprised 2,822 training proteins, 650 validation proteins, and 625 test proteins. Selected proteins exhibited mean pI 10.1 ± 0.7, basic residue fraction 0.21 ± 0.06, and net charge at pH 7.0 of 15.6 ± 9.0, confirming enrichment for alkaliphile characteristics.

Biophysical Property Calculations

We computed biophysical properties for all proteins using sequence-only and structure-conditioned approaches to capture diverse aspects of protein chemistry and geometry.

Sequence-Derived Properties. We calculated sequence-only features using BioPython's ProteinAnalysis module [57]. Table 6 defines these features and their interpretation.

Table 6: Sequence-Derived Protein Properties.

Property	Definition	Interpretation
Sequence length	Total number of amino acids	Longer = larger protein
Molecular weight per residue	Sum of atomic masses divided by sequence length	Higher = heavier amino acid composition (enriched Trp, Tyr, Phe)
Acidic fraction	Fraction of Asp and Glu residues	Higher = more acidic protein
Basic fraction	Fraction of Lys, Arg, and His residues	Higher = more basic protein
Aromatic fraction	Fraction of Phe, Trp, and Tyr residues	Higher = greater π-electron density
Ionizable fraction	Fraction of Asp, Glu, Lys, Arg, His, Cys, Tyr	Higher = more pH-sensitive residues
Isoelectric point (pI)	pH at which protein carries no net charge	<7 = acidic protein; >7 = basic protein
Net charge at pH 7	Net charge at physiological pH	Positive = excess basic residues; negative = excess acidic residues
Acidic-to-basic ratio	Ratio of acidic to basic residue counts	Higher = acidic; lower = basic
GRAVY	Grand average of hydropathy (Kyte-Doolittle)	Negative = hydrophilic (aqueous); positive = hydrophobic (membrane)
Aromaticity	Aromatic residue fraction	Higher = more stacking interactions
Instability index	Stability predictor from dipeptide frequencies	>40 = unstable; <40 = stable
Buffer capacity	Absolute charge change between pH 6.0 and 8.0, divided by 2	Higher = greater pH sensitivity
Charge asymmetry	|basic count − acidic count| / sequence length	Higher = greater charge imbalance

Structure-Derived Properties. All structure-conditioned calculations used AlphaFold2 predicted structures. We implemented custom Python scripts using BioPython's PDB module to extract geometric and topological features. Table 7 defines these features.

To help interpretability in multivariate analyses, radius-of-gyration and mean Cβ–centroid distances were inverted to yield structural compactness ($1/R_g$) and centralization ($1/d_{C\beta}$), such that higher values uniformly indicate more compact, centrally packed folds.

Protein Likelihood Scoring

We evaluated seven models representing distinct training modalities: structure-conditioned models trained on experimental structures (ProteinMPNN [38]), structure-conditioned models trained on predicted structures (ESM-IF [39]), hybrid models trained on curated sequence–structure pairs (MIF [40]), hybrid models incorporating sequence pretraining transfer (MIF-ST), and sequence-only language models (ESM2 [49], CARP-640M [40]). All structure-conditioned models used AlphaFold-predicted backbones as input and were evaluated in deterministic mode

Table 7: Structure-Derived Protein Properties.

Property	Definition	Interpretation
Relative contact order (RCO)	Average sequence separation of contacting residues (Cα–Cα $\leq$ 8.0 Å, separation $\geq$ 4), normalised by length	Higher = complex topology with long-range contacts; lower = local contact patterns
Surface exposure	Fraction of residues beyond 70th percentile of Cα distances from centroid	Higher = extended conformation, greater solvent access; lower = buried, compact core
Helix percentage	Fraction of residues in α-helix (geometric criteria: consecutive Cα–Cα 3.6–4.0 Å, $i \to i+3$ 4.8–5.8 Å, $i \to i+4$ 5.8–6.8 Å)	Higher = helix-dominant (fibrous); lower = sheet/loop-dominant
Sheet percentage	Fraction of residues in β-sheet (consecutive Cα–Cα 3.7–4.1 Å, zigzag $i \to i+2$ 6.3–7.5 Å)	Higher = sheet-dominant (compact globular); lower = helix/loop-dominant
Loop percentage	Fraction of residues not in helix or sheet	Higher = flexible/disordered regions; lower = well-structured
Average Cβ distance	Mean distance of Cβ atoms (Cα for Gly) from centroid	Higher = extended structure; lower = tightly packed globular
Compactness (radius of gyration)	Root-mean-square Cα distance from centroid	Higher = extended, less compact; lower = tightly folded compact
Helix-sheet contrast	Helix % − Sheet %	Positive = helix-dominant; negative = sheet-dominant
Ordered %	Helix % + Sheet %	Higher = well-structured; lower = flexible/disordered

with dropout disabled. Positions lacking valid coordinates or containing non-canonical residues were excluded from scoring.

Higher log-likelihood values indicate the model assigns greater probability to the wild-type sequence given the structure, interpreted as the model's assessment that the sequence is more "compatible" or "fit" according to its learned distribution.

Structure-Conditioned Models. For structure-conditioned models (ProteinMPNN, ESM-IF, MIF, MIF-ST), we adapted each model's sequence-design interface to compute likelihoods for the wild-type (WT) sequence rather than sample new designs. ProteinMPNN was implemented via the ColabDesign library (version 1.1.1)[1].

At each residue position i, the model outputs a categorical probability distribution $p_i(a \mid \mathbf{X})$ over amino acids conditioned on the backbone $\mathbf{X}$. We summed the log-probabilities of the WT residues and normalised by sequence length L:

$$\log P(\mathbf{s} \mid \mathbf{X}) = \frac{1}{L} \sum_{i=1}^{L} \log p_i(s_i \mid \mathbf{X}), \tag{2}$$

where $\mathbf{s} = (s_1, \ldots, s_L)$ denotes the WT sequence. For models using logits rather than probabilities, log-softmax was applied internally to obtain $p_i(\cdot \mid \mathbf{X})$. MIF-ST differs from MIF only in its pretraining (sequence-transfer from CARP-640M); inference and likelihood computation were identical.

Reported likelihoods are negative log-likelihoods (higher values indicate greater model confidence).

[1] https://github.com/sokrypton/ColabDesign.git@v1.1.1

sequence-only Models. For masked language models (ESM2 and CARP-640M), we estimated pseudo-log-likelihood (PLL) by iteratively masking each residue and evaluating the probability of the true amino acid given the remaining sequence. To improve efficiency, batches of masked positions were processed simultaneously. Per-position probabilities were obtained from model logits via log-softmax, and PLL was computed as:

$$\log P(\mathbf{s}) \approx \frac{1}{L} \sum_{i=1}^{L} \log p_i(s_i \mid \mathbf{s}_{\setminus i}), \tag{3}$$

where $\mathbf{s}_{\setminus i}$ denotes the sequence with residue i masked.

Model checkpoints and default hyperparameters were taken from the original repositories; no fine-tuning or retraining was performed.

Variance Decomposition

To isolate marginal variance attributable to species identity, we fitted a sequence of ordinary least-squares (OLS) models that incrementally introduced confounding factors. Let y denote the model score and let $\mathcal{M}_0$–$\mathcal{M}_5$ denote nested models:

$$\mathcal{M}_0 : y = \beta_0 + \varepsilon \tag{4}$$

$$\mathcal{M}_1 : y = \beta_0 + \mathrm{C}(\text{protein_family}) + \varepsilon \tag{5}$$

$$\mathcal{M}_2 : \mathcal{M}_1 + \mathbf{S}_{\text{structural}} \tag{6}$$

$$\mathcal{M}_3 : \mathcal{M}_2 + \mathbf{Q}_{\text{quality}} \tag{7}$$

$$\mathcal{M}_4 : \mathcal{M}_3 + \mathbf{L}_{\text{sequence}} \tag{8}$$

$$\mathcal{M}_5 : \mathcal{M}_4 + \mathrm{C}(\text{species}) \tag{9}$$

where $\mathbf{S}$, $\mathbf{Q}$, and $\mathbf{L}$ are vectors of structural, quality, and sequence covariates respectively. Structural covariates included secondary structure composition, surface exposure, mean Cβ distance, and relative contact order. Quality covariates consisted of mean pLDDT. Sequence covariates included sequence length, isoelectric point, and net charge at pH 7.

The additional R^2 for species is:

$$\Delta R^2_{\text{species}} = R^2_{\mathcal{M}_5} - R^2_{\mathcal{M}_4} \tag{10}$$

To account for scale dependence, we report the partial R^2:

$$R^2_{\text{partial}} = \frac{\Delta R^2_{\text{species}}}{1 - R^2_{\mathcal{M}_4}} \tag{11}$$

which equals the squared semi-partial correlation between species and residuals after all confounds are removed. We compared $\mathcal{M}_4$ and $\mathcal{M}_5$ with F-tests, reporting results with $p < 0.05$ as significant.

ELO Rating System

We implemented a modified ELO rating system adapted from Boubdir et al. [50] and following the implementation [30] to quantify species-specific preferences of protein design models. Each species began with an initial rating of 1500, with ratings constrained between 100 and 3000. For each protein family, we conducted pairwise comparisons between all species pairs that had predictions for that family.

The expected outcome for species A against species B follows:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{12}$$

where R_A and R_B are current ratings of species A and B.

Before ELO calculations, model scores underwent z-score normalisation within each protein family. To account for structural confidence, we incorporated quality weighting based on pLDDT scores:

$$w = \sqrt{\frac{\text{pLDDT}_A \times \text{pLDDT}_B}{10000}} \tag{13}$$

This geometric mean ensures high-confidence structure comparisons receive full weight, while low-confidence predictions are proportionally down-weighted.

Domain-level differences in ELO ratings were tested using chi-square tests comparing observed versus expected rating distributions across taxonomic domains.

Principal Component Analysis

To identify systematic patterns in how models score proteins differently, we performed PCA on biophysical features and mapped model preferences onto the resulting property space.

Feature Sets and Preprocessing

We conducted separate analyses using three feature sets:

Mixed features (n=10): molecular weight per residue, isoelectric point, instability index, GRAVY, sequence length, aromaticity, helix-sheet contrast, ordered secondary structure percentage, relative contact order, and average Cβ distance. This feature set captures both sequence-level biochemistry and structural geometry.

sequence-only features (n=6): molecular weight per residue, isoelectric point, instability index, GRAVY, sequence length, and aromaticity. These features can be calculated from sequence alone without structural information.

Structure-only features (n=4): helix-sheet contrast, ordered secondary structure percentage, relative contact order, and average Cβ distance. Requires backbone information (e.g PDB) to determine.

For pH-specific analyses related to alkaline adaptation, we used: buffer capacity, charge per residue, acidic residue fraction, basic residue fraction, and ionizable residue fraction.

Features were standardised (zero mean, unit variance) before PCA. We retained proteins with complete data for all features in each analysis (typically >95% of the dataset). Principal components were extracted using singular value decomposition with no anchoring constraints to allow identification of major variation axes.

Feature Contribution Quantification

Individual feature contributions to principal component variance were calculated as:

$$C_f = L^2_{f,\text{PC1}} \cdot \lambda_{\text{PC1}} + L^2_{f,\text{PC2}} \cdot \lambda_{\text{PC2}} \tag{14}$$

where $L_{f,k}$ is the loading of feature f on principal component k, and λ_k is the proportion of total variance explained by PC k.

For instance, if GRAVY loads 0.5 on PC1 (explaining 27% variance) and 0.3 on PC2 (21% variance):

$$C_{\text{GRAVY}} = (0.5)^2 \times 0.27 + (0.3)^2 \times 0.21 = 0.087$$

indicating GRAVY contributes 8.7% to the total variance captured in the PC space.

Generalized Additive Model Preference Landscapes

To visualize how model scores relate to protein properties, we fit generalized additive models (GAMs) relating each model's likelihood scores to principal component coordinates:

$$\text{score} \sim s(\text{PC1}, \text{PC2}, k = k_{\text{basis}}) \tag{15}$$

where the basis dimension k_{basis} was set adaptively based on sample size: $\max(12, \min(50, \lfloor N/300 \rfloor))$. Models were fitted using restricted maximum likelihood (REML) via the `mgcv` package in R.

Predictions were generated on a 120×120 grid spanning symmetric PC ranges with 5% expansion margins. Contour lines were automatically placed at 10 bins to visualise preference gradients. For cosine similarity calculations identifying peak likelihood regions, we used a coarser 60×60 grid spanning observed data ranges.

Cosine Similarity Analysis

To quantify alignment between model preferences and biophysical feature loadings in principal component space, we calculated cosine similarity between preference vectors and feature loading vectors.

For each model, we fitted a generalised additive model (GAM) relating likelihood scores to PC coordinates using a two-dimensional smooth term with basis dimension k=12. We identified the grid point with maximum predicted likelihood on a 60×60 grid spanning the observed PC range, and computed a likelihood-weighted preference vector from the origin to this maximum:

$$\mathbf{v}_{\text{model}} = \frac{\mathbf{p}_{\text{max}}}{\|\mathbf{p}_{\text{max}}\|} \tag{16}$$

where $\mathbf{p}_{\text{max}}$ is the PC coordinate vector at the GAM-predicted maximum likelihood location.

For pH-related features, we calculated a composite pH loading vector $\mathbf{v}_{\text{pH}}$ by averaging the principal component loadings of pH-related features (buffer capacity, charge per residue, basic residue fraction, acidic residue fraction, ionizable residue fraction) and normalising to unit length:

$$\mathbf{v}_{\text{pH}} = \frac{\sum_{f \in \text{pH features}} \mathbf{L}_f}{\left\|\sum_{f \in \text{pH features}} \mathbf{L}_f\right\|} \tag{17}$$

where $\mathbf{L}_f$ is the loading vector for feature f.

Alignment between vectors was quantified using cosine similarity:

$$\cos(\theta) = \mathbf{v}_{\text{model}} \cdot \mathbf{v}_{\text{pH}} \tag{18}$$

where values near +1 indicate aligned preferences (model favors high-pH proteins), values near 0 indicate orthogonal preferences, and values near -1 indicate opposing preferences (model disfavors high-pH proteins).

Statistical significance of cosine similarity changes between vanilla and fine-tuned models was assessed using posterior simulation from the fitted GAM coefficient distributions. For each of 300 bootstrap iterations, we drew

coefficient vectors from the multivariate normal distribution defined by the GAM parameter estimates and their variance-covariance matrix, predicted likelihood surfaces, identified maximum locations, and computed cosine similarities. This posterior-based approach accounts for uncertainty in the GAM fits without requiring model refitting. Significance was determined by the proportion of bootstrap iterations where the difference exceeded zero, with p-values calculated as twice the minimum tail probability.

Protein Design Generation and Analysis

We selected 16 proteins based on model preference scores (ELO ratings), ensuring representation across taxonomic domains (6 Archaea, 4 Bacteria, 6 Eukaryota), size range (5.7–73.5 kDa), compactness (RCO values 0.05–0.25), and secondary structure content (Table 8).

For each protein structure, we generated 48 sequence designs per model using temperature sampling at T=0.1. No iterative refinement or filtering was applied—all 48 initial samples were retained for analysis. Generation used identical random seeds across models to enable direct comparison.

As with the likelihood generation, protein sequences for ProteinMPNN were designed via the ColabDesign library (version 1.1.1)[2]. ColabDesign provides a JAX-compatible interface to ProteinMPNN's pre-trained weights (v_48_002), enabling efficient batched sequence sampling. For each input backbone structure (PDB format), the library computes residue-level amino acid probability distributions conditioned on backbone coordinates.

Biophysical properties were calculated for all designed sequences using BioPython ProteinAnalysis. Changes in biophysical properties between wild-type and designed sequences were assessed using paired t-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.05$). Effect sizes were calculated as Cohen's d_z for paired samples:

$$d_z = \frac{\bar{D}}{s_D} \tag{19}$$

where $\bar{D}$ is the mean difference between paired observations and s_D is the standard deviation of differences. This paired effect size metric accounts for correlation between wild-type and designed sequences from the same template.

For evaluating alkaline adaptation in designed sequences, we used:

$$S_{\text{alk,eval}} = \frac{f_{\text{basic}}}{f_{\text{acidic}} + 0.1} \times \sigma(\text{pI} - 7) \tag{20}$$

where $\sigma(\cdot)$ is the sigmoid function and the 0.1 constant prevents division by zero.

Sequence-Feature Principal Component Analysis

We computed a principal component analysis over six sequence-derived features for wild-type proteins: molecular weight per residue, isoelectric point, instability index, GRAVY, sequence length, and aromaticity. Features were standardised prior to PCA. Designed sequences (approximately 48 per protein×model, temperature $T = 0.1$) were projected into wild-type PC space using wild-type centering, scaling, and loadings, preventing information leakage from designs into the basis.

For each protein×model, we computed the mean designed-sequence position and calculated the wild-type-to-design shift vector. Model centroid shifts were obtained by averaging per-protein shifts across templates. One-sample t-tests assessed whether centroid shifts differed from zero for each PC dimension, with Cohen's d as effect size.

[2] https://github.com/sokrypton/ColabDesign.git@v1.1.1

Table 8: Proteins selected for design experiments.

Entry	Domain	Model Rank	Size (kDa)	RCO	Structure (% α/β)
Highly-ranked proteins (thermophilic bias)					
O67786	Bacteria	Top	25.8	Medium	66.8 / low
P83194	Archaea	Top	21.5	High	low / 65.6
Q58502	Archaea	Top	–	High	low / 60.3
Q980U8	Archaea	Top	39.6	Low	65.6 / low
O58832	Archaea	Top	38.5	Low	Balanced
Q5JI38	Archaea	Top	32.0	Medium	Balanced
Q9Z9J6	Bacteria	Top	7.0	–	Balanced
Poorly-ranked proteins (mesophilic bias)					
P02819	Eukaryota	Bottom	5.7	Medium	63.3 / low
Q4R312	Eukaryota	Bottom	9.2	High	low / 61.0
Q54FJ9	Eukaryota	Bottom	73.5	Medium	Mixed
Q54GT9	Eukaryota	Bottom	46.0	Low	Hydrophobic
Q54JL0	Eukaryota	Bottom	73.5	Low	Mixed
Q69NP0	Eukaryota	Bottom	10.5	High	Balanced
Model organisms (neutral)					
P0A7R6	Bacteria	Neutral	11.7	High	low / 58.3
P60724	Bacteria	Neutral	22.1	Medium	Balanced
Q07938	Eukaryota	Neutral	37.9	Medium	Balanced

To interpret PC shifts biochemically, we linearly back-projected model centroid shifts into feature space via the wild-type loading matrix. This first-order approximation provides effect-size rankings for individual features. Per-feature t-tests were corrected using Benjamini–Hochberg false discovery rate control.

Directional agreement between models was quantified using cosine similarity of per-protein shift vectors. For each pair of models, we computed cosine similarities across all proteins where both models had valid designs, then calculated mean and standard deviation across proteins. Large standard deviations indicate substantial protein-to-protein heterogeneity in directional agreement, reflecting template-specific modulation of design preferences alongside population-level model tendencies.

Fine-Tuning ProteinMPNN

We fine-tuned ProteinMPNN on the alkaliphile dataset to test whether identified biases could be strategically modified while preserving design capability. Each structure was pre-processed to generate PyTorch tensor files containing backbone coordinates and sequences.

Training used: batch size of 1,000 residues, dropout of 0.1, backbone noise of 0.2 Å, learning rate of 1×10^{-4} with standard ProteinMPNN optimisation schedule [38], and gradient clipping disabled. Models trained for 30 epochs using mixed precision training with PyTorch's automatic mixed precision (AMP). The `loss_smoothed` function computed training loss while `loss_nll` provided validation metrics.

ProteinMPNN architecture: 3 encoder/decoder layers, 128-dimensional hidden states, k=48 nearest neighbors. Models initialised from pretrained weights (v_48_002) using the standard ProteinMPNN optimiser implementing learning rate scheduling based on hidden dimensions and training steps. The final model checkpoint was selected from epoch 20 based on validation performance.

The training pipeline used ProteinMPNN's native `featurize` function for batch processing, ensuring proper handling of multi-chain complexes and masking. The `StructureDataset` and `StructureLoader` classes managed dynamic

batching based on total residue count rather than protein count.

To evaluate bias modification, we applied the same variance decomposition, ELO rating, and PCA-based preference analysis to the fine-tuned model. We generated designs for the same 16 test proteins using both vanilla and fine-tuned models, comparing property distributions using Mann-Whitney U tests with Bonferroni correction. Effect sizes were computed using Cohen's d_z.

Computational Implementation

All analyses were implemented in Python 3.11 using `pandas` for data handling, `statsmodels` for statistical modeling, `BioPython` for protein property calculations, and `NumPy` for numerical computations. PCA and visualisation were performed in R 4.3 using `prcomp` for principal component analysis and `ggplot2` for plotting.

Supplementary Information

Additional data and detailed analyses are provided in the Supplementary Information.

References

[1] Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016). URL https://www.nature.com/articles/nature19946. Publisher: Nature Publishing Group.

[2] Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences* **97**, 10383–10388 (2000). URL https://www.pnas.org/doi/full/10.1073/pnas.97.19.10383. Publisher: Proceedings of the National Academy of Sciences.

[3] Holm, L. & Sander, C. Fast and simple monte carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins: Structure, Function, and Bioinformatics* **14**, 213–223 (1992). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340140208. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340140208.

[4] Yue, K. & Dill, K. A. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 4163–4167 (1992). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC525653/.

[5] Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The Protein Folding Problem. *Annual Review of Biophysics* **37**, 289–316 (2008). URL https://www.annualreviews.org/content/journals/10.1146/annurev.biophys.37.092707.153558. Publisher: Annual Reviews.

[6] Woolfson, D. N. A Brief History of *De Novo* Protein Design: Minimal, Rational, and Computational. *Journal of Molecular Biology* **433**, 167160 (2021). URL https://www.sciencedirect.com/science/article/pii/S0022283621003892.

[7] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16**, 1315–1322 (2019). URL https://www.nature.com/articles/s41592-019-0598-1. Publisher: Nature Publishing Group.

[8] Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences* **116**, 8852–8858 (2019). URL https://www.pnas.org/doi/10.1073/pnas.1901979116. Publisher: Proceedings of the National Academy of Sciences.

[9] Dahiyat, B. I. & Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **278**, 82–87 (1997). URL https://www.science.org/doi/10.1126/science.278.5335.82. Publisher: American Association for the Advancement of Science.

[10] Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology* **20**, 681–697 (2019). URL https://www.nature.com/articles/s41580-019-0163-x. Publisher: Nature Publishing Group.

[11] Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024). URL https://www.science.org/doi/full/10.1126/science.adl2528. Publisher: American Association for the Advancement of Science.

[12] Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **13**, 3031–3048 (2017). URL https://doi.org/10.1021/acs.jctc.7b00125. Publisher: American Chemical Society.

[13] Anishchenko, I. *et al.* De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021). URL https://www.nature.com/articles/s41586-021-04184-w. Publisher: Nature Publishing Group.

[14] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021). URL https://www.pnas.org/doi/full/10.1073/pnas.2016239118. Publisher: Proceedings of the National Academy of Sciences.

[15] Notin, P., Rollins, N., Gal, Y., Sander, C. & Marks, D. Machine learning for functional protein design. *Nature Biotechnology* **42**, 216–228 (2024). URL https://www.nature.com/articles/s41587-024-02127-0. Publisher: Nature Publishing Group.

[16] Ovchinnikov, S. & Huang, P.-S. Structure-based protein design with deep learning. *Current Opinion in Chemical Biology* **65**, 136–144 (2021). URL https://www.sciencedirect.com/science/article/pii/S1367593121001125.

[17] Vaswani, A. *et al.* Attention is All you Need. In *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[18] Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023). URL https://www.nature.com/articles/s41586-023-06415-8. Publisher: Nature Publishing Group.

[19] Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems* **15**, 286–294.e2 (2024). URL https://www.cell.com/cell-systems/abstract/S2405-4712(24)00029-2. Publisher: Elsevier.

[20] Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023). URL https://www.nature.com/articles/s41586-023-05696-3. Publisher: Nature Publishing Group.

[21] Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models (2021). URL http://arxiv.org/abs/2006.15222. ArXiv:2006.15222 [cs, q-bio].

[22] Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction (2023). URL https://www.biorxiv.org/content/10.1101/2023.12.07.570727v1. Pages: 2023.12.07.570727 Section: New Results.

[23] Torralba, A. & Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528 (2011). URL https://ieeexplore.ieee.org/document/5995347. ISSN: 1063-6919.

[24] Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623 (Association for Computing Machinery, New York, NY, USA, 2021). URL https://dl.acm.org/doi/10.1145/3442188.3445922.

[25] Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017). URL https://www.science.org/doi/full/10.1126/science.aal4230. Publisher: American Association for the Advancement of Science.

[26] Taori, R. & Hashimoto, T. Data Feedback Loops: Model-driven Amplification of Dataset Biases. In *Proceedings of the 40th International Conference on Machine Learning*, 33883–33920 (PMLR, 2023). URL https://proceedings.mlr.press/v202/taori23a.html. ISSN: 2640-3498.

[27] Chen, I., Johansson, F. D. & Sontag, D. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Inc., 2018). URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/1f1baa5b8edac74eb4eaa329f14a0361-Abstract.html.

[28] Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91 (PMLR, 2018). URL https://proceedings.mlr.press/v81/buolamwini18a.html. ISSN: 2640-3498.

[29] Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019). URL https://www.science.org/doi/10.1126/science.aax2342. Publisher: American Association for the Advancement of Science.

[30] Ding, F. & Steinhardt, J. Protein language models are biased by unequal sequence sampling across the tree of life (2024). URL https://www.biorxiv.org/content/10.1101/2024.03.07.584001v1. Pages: 2024.03.07.584001 Section: New Results.

[31] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, N.J.)* **406**, 89–112 (2007).

[32] Shaw, A. *et al.* Removing bias in sequence models of protein fitness (2023). URL https://www.biorxiv.org/content/10.1101/2023.09.28.560044v1. Pages: 2023.09.28.560044 Section: New Results.

[33] Anand, N. *et al.* Protein sequence design with a learned potential. *Nature Communications* **13**, 746 (2022). URL https://www.nature.com/articles/s41467-022-28313-9. Publisher: Nature Publishing Group.

[34] Huang, H., Lin, Z., He, D., Hong, L. & Li, Y. RiboDiffusion: tertiary structure-based RNA inverse folding with generative diffusion models. *Bioinformatics* **40**, i347–i356 (2024). URL https://doi.org/10.1093/bioinformatics/btae259.

[35] Grishin, N. V. Fold Change in Evolution of Protein Structures. *Journal of Structural Biology* **134**, 167–185 (2001). URL https://www.sciencedirect.com/science/article/pii/S1047847701943357.

[36] Orengo, C. A. & Thornton, J. M. PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL PERSPECTIVE. *Annual Review of Biochemistry* **74**, 867–900 (2005). URL https://www.annualreviews.org/content/journals/10.1146/annurev.biochem.74.082803.133029. Publisher: Annual Reviews.

[37] Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* **77**, 499–508 (2009). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22458. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.22458.

[38] Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022). URL https://www.science.org/doi/10.1126/science.add2187. Publisher: American Association for the Advancement of Science.

[39] Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, 8946–8970 (PMLR, 2022). URL https://proceedings.mlr.press/v162/hsu22a.html. ISSN: 2640-3498.

[40] Yang, K. K., Zanichelli, N. & Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection* **36**, gzad015 (2023). URL https://doi.org/10.1093/protein/gzad015.

[41] Derry, A., Carpenter, K. A. & Altman, R. B. Training data composition affects performance of protein structure analysis algorithms. *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing* **27**, 10–21 (2022). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8669736/.

[42] Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000). URL https://doi.org/10.1093/nar/28.1.235.

[43] Tusnády, G. E., Dosztányi, Z. & Simon, I. PDB_tm: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Research* **33**, D275–D278 (2005). URL https://doi.org/10.1093/nar/gki002.

[44] Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021). URL https://www.nature.com/articles/s41586-021-03828-1. Publisher: Nature Publishing Group.

[45] Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16**, 18–29 (2015). URL https://www.nature.com/articles/nrm3920. Publisher: Nature Publishing Group.

[46] Bondos, S. E., Dunker, A. K. & Uversky, V. N. Intrinsically disordered proteins play diverse roles in cell signaling. *Cell Communication and Signaling* **20**, 20 (2022). URL https://doi.org/10.1186/s12964-022-00821-7.

[47] Moses, D. *et al.* Structural biases in disordered proteins are prevalent in the cell. *Nature Structural & Molecular Biology* **31**, 283–292 (2024). URL https://www.nature.com/articles/s41594-023-01148-8. Publisher: Nature Publishing Group.

[48] Krapp, L. F. *et al.* Context-aware geometric deep learning for protein sequence design. *Nature Communications* **15**, 6273 (2024). URL https://www.nature.com/articles/s41467-024-50571-y. Publisher: Nature Publishing Group.

[49] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). URL https://www.science.org/doi/full/10.1126/science.ade2574. Publisher: American Association for the Advancement of Science.

[50] Boubdir, M., Kim, E., Ermis, B., Hooker, S. & Fadaee, M. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation (2023). URL http://arxiv.org/abs/2311.17295. ArXiv:2311.17295 [cs].

[51] Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* **52**, D368–D375 (2024). URL https://doi.org/10.1093/nar/gkad1011.

[52] Krulwich, T. A., Sachs, G. & Padan, E. Molecular aspects of bacterial pH sensing and homeostasis. *Nature Reviews Microbiology* **9**, 330–343 (2011). URL https://www.nature.com/articles/nrmicro2549. Publisher: Nature Publishing Group.

[53] Teilum, K., Olsen, J. G. & Kragelund, B. B. Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences* **66**, 2231–2247 (2009). URL https://doi.org/10.1007/s00018-009-0014-6.

[54] Tan, C. *et al.* AlphaFold Database Debiasing for Robust Inverse Folding (2025). URL http://arxiv.org/abs/2506.08365. ArXiv:2506.08365 [cs].

[55] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

[56] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). URL https://doi.org/10.1093/bioinformatics/bts565.

[57] Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009). URL https://doi.org/10.1093/bioinformatics/btp163.