MULTI-STATE PROTEIN SEQUENCE DESIGN WITH DYNAMICMPNN

Anonymous authors
Paper under double-blind review

ABSTRACT

Structural biology has long been dominated by the *one sequence, one structure, one function* paradigm, yet many critical biological processes—from enzyme catalysis to membrane transport—depend on proteins that adopt multiple conformational states. Existing multi-state design approaches rely on post-hoc aggregation of single-state predictions, achieving poor experimental success rates compared to single-state design. We introduce DynamicMPNN, an inverse folding model explicitly trained to generate sequences compatible with multiple conformations through joint learning across conformational ensembles. Trained on 46,033 conformational pairs covering 75% of CATH superfamilies and evaluated using Alphafold 3, DynamicMPNN outperforms ProteinMPNN by up to 25% on decoy-normalized RMSD and by 12% on sequence recovery across our challenging multi-state protein benchmark.

1 INTRODUCTION

A commonly derived assumption from Anfinsen's experiment is that proteins adopt only one native 3D structure, leading to the "one sequence, one structure, one function" canon. This view has been indirectly reinforced by the predominant use of X-Ray crystallography in experimental protein structure determination, which requires that proteins form a homogenous, diffractible crystal to be characterised (Dishman & Volkman, 2018). The large collection of static protein structures in the PDB has enabled the development of high-accuracy machine learning models for tasks such as structure prediction (Jumper et al., 2021; Kryshtafovych et al., 2019) and inverse folding (Dauparas et al., 2022; Hsu et al., 2022). Amongst contemporary inverse folding models, ProteinMPNN has been particularly widely adopted in applied protein design projects due to its low inference costs and robust experimental success rates (Dauparas et al., 2022; Watson et al., 2023; Goverde et al., 2024), outperforming traditional physics-based design methods on both fronts (Liu & Kuhlman, 2006).

Although direct experimental characterisation of protein dynamics remains a challenge, the conformational diversity of proteins underlies crucial biological functions such as enzyme catalysis, protein-protein interactions, allostery, and human disease (Monzon et al., 2016). In applied protein design, bio-switches - proteins that switch between two structural states - are of particular importance, with key applications in engineering artificial bio-motors, signalling pathways, biosensors, or drug delivery systems (Stein & Alexandrov, 2015; Praetorius et al., 2023). While most known switches undergo rearrangements in the context of a single fold (Ambroggio & Kuhlman, 2006a), the class of metamorphic proteins undergo changes in both their secondary structure and fold (Fig. 1a) and have been predicted to represent up to 4% of the PDB chains (Porter & Looger, 2018). These proteins typically adopt two main functional states (Dishman & Volkman, 2018) and a finite number of conformations (see Discussion). Beyond the world of switches, other dynamic proteins are characterised by continuous conformational landscapes (e.g. intrinsically disordered proteins (Tompa & Fuxreiter, 2008)).

Multi-state protein design was first achieved through rational design and physics-based methods such as RosettaDesign (Liu & Kuhlman, 2006; Vucinic et al., 2020; Karimi & Shen, 2018). Previous campaigns leveraging these methods have attempted to design metamorphic metal-binding peptides (Ambroggio & Kuhlman, 2006b; Cerasoli et al., 2005), closely related sequences that adopt diverging folds (Wei et al., 2020), and hinge proteins with binder-regulated thermodynamic equilibria, allowing the relative populations of different structural states to be modulated by exogenous proteins

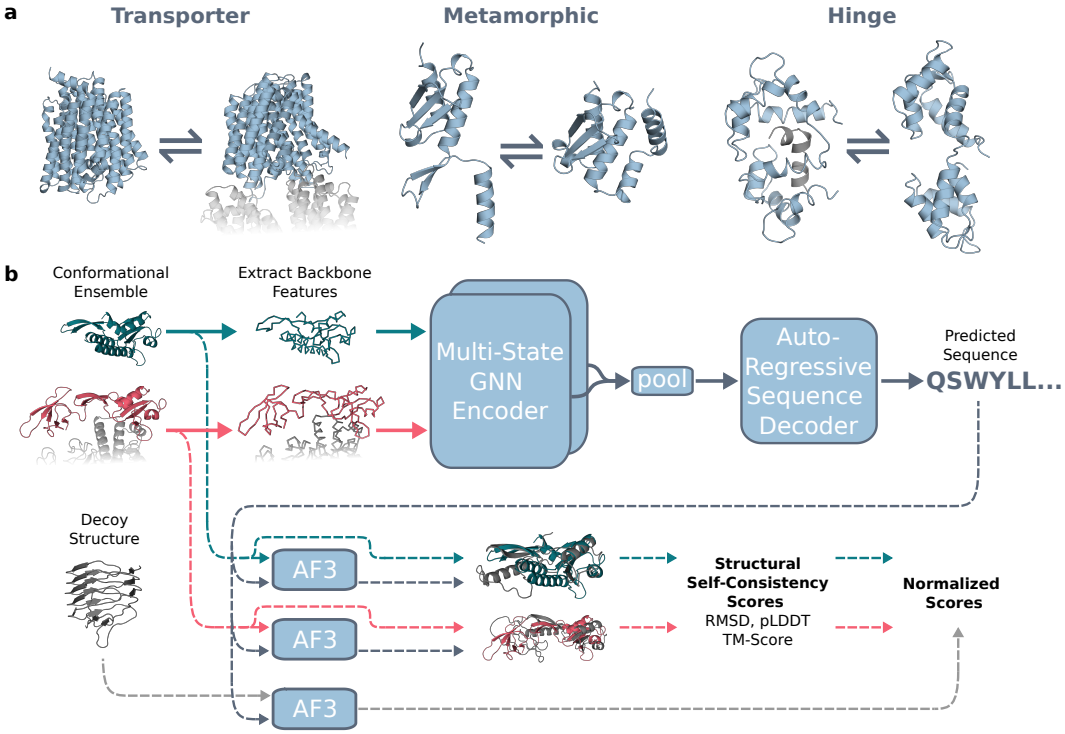


Figure 1: DynamicMPNN for multi-state protein design. (a) Examples of proteins with multiple conformational states: transporters in closed and open states (PDB: 6NC7, 6NC9), metamorphic protein with alternative folds (PDB: 4QHH, 4QHF) and hinges showing domain movement (PDB: 5D0W, 1CFC). (b) Schematic of DynamicMPNN, an inverse folding model trained to generate protein sequences with multiple conformational states. Conformations are encoded with their respective chemical environments (i.e. interaction partners shown in gray). Solid lines show the flow of information in the model, while dashed lines show the evaluation pipeline using AlphaFold 3 (AF3); employing target structures as templates during inference and measuring the deviations between predicted and target structures, with decoy structures serving as negative controls.

(Zhang et al., 2022b; Quijano-Rubio et al., 2021). More recently, Praetorius et al. (2023) leveraged ProteinMPNN Multi-state Design (ProteinMPNN-MSD) (Dauparas et al., 2022) to link independent backbone states via step-wise logit averaging with shared autoregressive context. Multi-state ESM-IF (Hsu et al., 2022) employs an analogous strategy, differing primarily in aggregating probabilities (geometric mean) rather than logits (see Appendix A.3). ProteinGenerator (Lisanza et al., 2024) extends this principle to sequence diffusion, averaging logits across distinct structural conditioning inputs at each denoising step.

Despite these advances, current multi-state design pipelines have shown limited success. In Hsu et al. (2022), sequences designed with the dual-state strategy only showed marginally lower perplexity compared to sequences inverse-folded from one of the states and refoldability was not explored. Praetorius et al. (2023) used ProteinMPNN-MSD to design a de novo hinge protein with one sequence binding to a peptide: From an initial pool of over 2M computational designs conditioned on 28K similar de novo backbone pairs, a rigorous set of computational filters yielded only 9K sequences as likely candidates; experimental validation of a few sampled candidates showed a high success rate (Appendix A.3). Likewise, the authors of ProteinGenerator Lisanza et al. (2024) reported a significantly lower *in silico* design success rate of 0.05% for their multi-state backbone design task compared to rates of 2-10% observed for various single-state sequence design objectives using their framework (Lisanza et al., 2024). These observations - combined with the relative scarcity of published data on ML-driven multi-state *de novo* design campaigns - suggest that current ML methods for multi-state protein design have been significantly less successful than their single-state counterparts. We propose that this gap can be attributed to limited multi-conformational

datasets, weak benchmarks, and the poor performance of folding models in predicting alternative states (Chakravarty et al., 2024), which adversely affects their efficacy as self-consistency filters in protein design workflows.

Our contributions. This paper introduces DynamicMPNN (Fig. 1b), a novel geometric deep learning-based pipeline for multi-state protein sequence design.

- DynamicMPNN is the first explicit multi-state inverse folding model for protein design. To train DynamicMPNN, we create a new ML-ready dataset of proteins with multiple conformations using the PDB and CoDNaS (Monzon et al., 2016) databases, and evaluate the method on 96 biologically relevant metamorphic, hinge, and transporter proteins.
- We introduce a novel data processing pipeline and architecture able to handle sequence-aligned structural ensembles with heterogeneous sequences, enabling us to leverage the conformational diversity across proteins with high sequence similarity.
- We propose a multi-state self-consistency metric and benchmark based on *Alphafold3* (AF3) (Abramson et al., 2024b) using target structures as templates.
- DynamicMPNN improves performance over ProteinMPNN (Dauparas et al., 2022) on AF3 by up to 25% on RMSD and 8% on TM-score decoy normalized self-consistency values.

2 The DynamicMPNN pipeline

2.1 Protein multi-conformational dataset

While over 900K individual protein chains (sequence-structure pairs) are available in the PDB, multi-conformational data is far more scarce with only roughly 12,000 NMR-derived protein ensembles covering just 21% of CATH superfamilies. To overcome this limitation, we exploit the sequence redundancy across the PDB to create build two multi-conformational datasets:

- **CoDNaS** We use the CoDNaS dataset (Monzon et al., 2016), which clusters PDB chains at $\geq$95% sequence similarity with unique UniProt IDs per cluster to prevent homologue leakage, yielding 46,033 clusters with varying numbers of conformations and covering 75% of CATH superfamilies (Fig. 2a,b).
- **PDB80** To capture greater conformational diversity within training ensembles, we also employ $\geq$80% sequence similarity clusters available from the Protein Data Bank (Berman et al., 2000), yielding 46,924 clusters with at least two conformations.

Training-time data sampling. From CoDNaS, we select the maximum-RMSD chain pair from each cluster to maximize conformational signal and reduce alignment artifacts[1] (Fig. 2a). For PDB80, pairs are sampled with probability proportional to their structural dissimilarity ($1 - TM_{score}$), biasing training toward larger conformational changes (Appendix A.4). We also explore extending the latter approach to 3 and 5 states (Section 2.2.2), though most known multi-state proteins switch between just two functional conformations (Dishman & Volkman, 2018; Leaver-Fay et al., 2011; Alberstein et al., 2022), and validating designs beyond two states remains experimentally challenging (Niazi, 2025). While our architecture supports arbitrary numbers of states, for the aforementioned reasons, we focus primarily on two-state design. Molecular dynamics datasets were not used due to limited sequence variability and scarcity of trajectories capturing large conformational changes (Vander Meersche et al., 2023; Mirarchi et al., 2024); disordered protein simulations (Tesei et al., 2024) were similarly excluded given their continuous conformational landscapes and low structural signal-to-noise ratio compared to globular multi-state proteins (Tompa & Fuxreiter, 2008).

Dataset splitting. We curate a benchmark composed of four previous studies of proteins with large 2-state conformational changes: (1) 92 metamorphic proteins (Porter & Looger, 2018), (2) 91 apo-holo proteins (Saldaño et al., 2022), (3) the OC23 and OC85 open-closed datasets (Kalakoti & Wallner, 2025), and (4) 20 transporter proteins (Kalakoti & Wallner, 2025). The proteins with the highest inter-state RMSD were assigned to the test set (96 samples), while the rest were assigned to the validation set (100 samples). Training clusters were filtered to exclude any with TM-score > 0.4

[1]Note that aligning more sequences introduces additional gap tokens (i.e. reducing sequence-structure overlap) diluting the conformational signal.

(Zhang & Skolnick, 2004; Xu & Zhang, 2010) and $> 30\%$ sequence similarity to test/validation structures, preventing structural similarity leakage and yielding a final training set of 44,243 conformer pairs. To evaluate designs beyond two states, we additionally curate a set of six proteins with well-characterized intermediate or flexible conformations: MBP (Wang et al., 2012), α-hemolysin (Chatterjee et al., 2025), and Selecase (López-Pelegrín et al., 2013) (stable intermediates; 3-states), and Calmodulin, α-synuclein (Chen et al., 2021), and Aβ-42 (flexible/disordered; 5-states).

We additionally curate a set of single-state sequence-structure pairs from the 30% sequence similarity clusters not represented in our multi-conformational training dataset ($n = 27,394$). This augmentation strategy maximizes coverage of protein fold space while preserving the multi-conformational learning signal. See Appendix A.4 for further details on dataset composition.

2.2 DynamicMPNN for multi-state inverse folding

Single-state inverse folding methods seek to model the conditional distribution $p(Y|X)$ where $X \in \mathbb{R}^{n\times 3\times 3}$ represents a protein backbone with n residues, and $Y = (y_1, \ldots, y_n)$ is the amino acid sequence. Extensions of these methods to multi-state design have thus far been limited to post-hoc/decoder-level aggregation (A.3) of independent single-state predictions. We believe such methods favour logits highly biased towards one conformation, whose average over the states is higher than the one of moderately valued logits across both states (Joshi et al., 2025). Instead, DynamicMPNN learns the joint conditional distribution of $p(Y|X_1, \ldots, X_m)$ directly through autoregressive sequence generation, where $\{X_1, \ldots, X_m\}$ represent distinct protein conformations encoded into a shared latent space; thus learning a sequence distribution that simultaneously satisfies multiple structural constraints. We decompose this joint conditional probability using the autoregressive factorization:

$$p(Y|X_1, ..., X_m) = \prod_{i=1}^{n} p(y_i|y_{i-1}, ..., y_1; X_1, ..., X_m) \tag{1}$$

where each factor represents the probability of selecting residue y_i given the sequence prefix and the complete structural ensemble.

Overall architecture. DynamicMPNN independently encodes each of the functional states of a protein, together with their binding partners, into a shared latent feature space (Fig. 1b). Embeddings of the chains-to-be-designed are then pooled across conformations to obtain a single embedding from which a sequence is auto-regressively generated.

Our architecture builds upon gRNAde (Joshi et al., 2025), a multi-state GNN model for RNA inverse folding. For both encoder and decoder, we employ SE(3)-equivariant Geometric Vector Perceptron (Jing et al., 2021) layers which maintain computational efficiency through edge sparsity (k-NN edges with k=32). Within the GVP, scalar features $\boldsymbol{s}_i \in \mathbb{R}^{k\times f}$ and vector features $\vec{\boldsymbol{v}}_i \in \mathbb{R}^{k\times f'\times 3}$ are defined for each node i (Duval et al., 2024):

$$\boldsymbol{m}_i, \vec{\boldsymbol{m}}_i := \sum_{j\in\mathcal{N}_i} \textsc{Msg}\big((\boldsymbol{s}_i, \vec{\boldsymbol{v}}_i), (\boldsymbol{s}_j, \vec{\boldsymbol{v}}_j), \boldsymbol{e}_{ij}\big) \tag{2}$$

$$\boldsymbol{s}'_i, \vec{\boldsymbol{v}}'_i := \textsc{Upd}\big((\boldsymbol{s}_i, \vec{\boldsymbol{v}}_i)\ ,\ (\boldsymbol{m}_i, \vec{\boldsymbol{m}}_i)\big) \tag{3}$$

where Msg, Upd are Geometric Vector Perceptrons - a generalization of Multi Layer Perceptrons that takes as input and updates scalar and vector features along separate channels in order to achieve $O(3)$-equivariant message passing. The overall GNN encoder is $SO(3)$-equivariant due to the use of reflection-sensitive input features (dihedral angles) combined with $O(3)$-equivariant GVP-GNN layers Joshi et al. (2025). Both the encoder and decoder are assigned 8 GVP-GNN layers, following findings in Hsu et al. (2022) (2022). See Appendix A.5 for further details.

We present two different encoder architectures:

- **DynamicMPNN** uses independent encoder channels for each conformation, followed by Deep Set pooling (Zaheer et al., 2017) - it is invariant to conformation order and does not add extra parameters to the model. We note that while some more expressive pooling

strategies have been shown to provide marginal performance improvements, they usually come at a great cost in efficiency (Joshi et al., 2025). Only node features are updated.

- **DynamicMPNN + DSS** implements cross-attention between the encoder channels after each layer using a Deep Symmetric Set (DSS) (Maron et al., 2020) module, which allows for richer inter-conformation interactions at the cost of increased computational cost. In the scatter/gather DSS strategy, node embeddings of all design chains are averaged, passed through GVP layers, and added back to features of each channel through a residual connection. Both node and edge features are updated.

Heterogeneous sequence processing. A key architectural contribution is our handling of non-identical sequences across conformations, necessary for exploiting the full conformational diversity in the PDB. Aligning non-identical sequences introduces gap tokens, and X-ray structures often contain unresolved residues; together resulting in ensembles of heterogeneous composition (i.e. varying lengths, missing residues, alignment gaps). We address this with the following protocol: (1) cluster chain members are sequence-aligned prior to featurization; (2) paired PDB complexes are featurized and encoded independently; (3) gap positions are masked and excluded from message passing; (4) during pooling, embeddings from all cluster members are extracted, stacked, and pooled, with gap-node embeddings zeroed out. This preserves all available structural information while incorporating the context of the surrounding chemical environment (i.e. binding partners). To prevent leakage, during training we mask sequence information only for chains with $> 70\%$ similarity to the ground truth.

2.2.1 MULTI-CHAIN STRATEGY

Unlike previous work (Joshi et al., 2025), DynamicMPNN processes full PDB entries containing sampled cluster chains, enabling the model to condition conformational changes on binding partners and oligomeric states. This opens possibilities for engineering controllable conformational switches by tuning free energy differences between folds and their binding interactions (Alberstein et al., 2022). Note that while explicitly a 2-state approach, multi-chain training implicitly exposes the model to additional states when PDB entries contain multiple cluster conformations (Fig. 6). As discussed in Section 2.1, most multi-state proteins switch between only two conformations (Dishman & Volkman, 2018; Leaver-Fay et al., 2011; Alberstein et al., 2022), making 2-state design the most biologically relevant regime.

2.2.2 SINGLE-CHAIN STRATEGY

To efficiently run ablations for $k > 2$ states—training time and memory scale linearly with the number of encoded states—we implement a single-chain strategy encoding only cluster chains without binding partners. Single-chain training is performed on PDB80, which provides sufficient cluster diversity for sampling $k \in \{2, 3, 5\}$ states via TM-score weighted sampling. We evaluate these models on 6 proteins with well-characterized multi-state behavior (Tables 6, 7) and directly compare k=2 single-chain versus multi-chain performance (Table 3). Extending multi-chain training to $k > 2$ states is straightforward and left for future work.

2.3 MULTI-STATE DESIGN EVALUATION

Given the high degeneracy of the sequence-to-structure mapping, where divergent sequences can adopt identical folds (Rost, 1999; Sander & Schneider, 1991), sequence recovery is often an insufficient metric for design success. Therefore, following previous work Wang et al. (2023), we also evaluate the refoldability of generated sequences (Appendix A.2). Existing refoldability methods compare target structures to single conformations predicted by folding models such as AlphaFold2 (Jumper et al., 2021). We argue that this unconstrained approach is unsuitable for multi-state design since folding models typically predict one dominant state or an unphysical interpolation, failing to sample the full conformational ensemble Lane (2023); Chakravarty et al. (2024); Saldaño et al. (2022).

Template-based Alphafold 3 refoldability. To address these sampling limitations, we propose a template-based AlphaFold 3 (AF3) framework Abramson et al. (2024b) to explicitly verify structural compatibility - we adapt the findings of Roney & Ovchinnikov (2022) to our multi-state setting

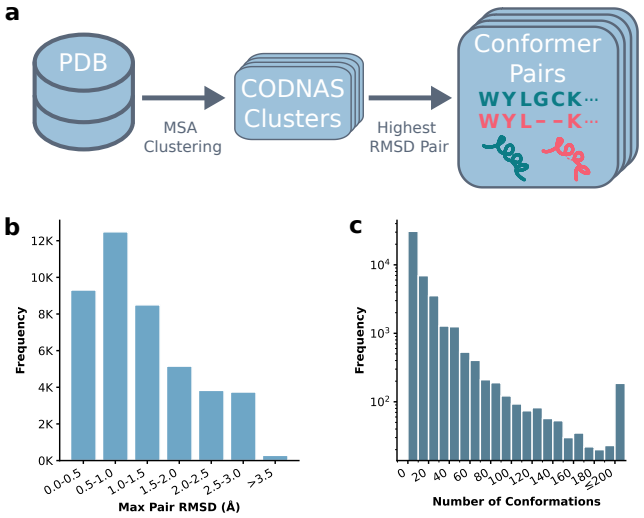


Figure 2: Multi-state protein dataset. (a) Data processing pipeline used to construct sequence-aligned structure pairs. (b) Distribution of the maximum Cα-RMSD between pairs of structures in each CoDNaS cluster. (c) Distribution of the number of conformations per CoDNaS cluster.

(see Appendix A.6). While unconstrained DL models may fail to spontaneously find alternative states, we leverage AF3's template mechanism to moderately bias the model towards specific target conformations. For each de novo sequence, we perform 2 AF3 runs, one with each conformational state provided as template, shifting the evaluation from structural prediction to compatibility assessment. The similarity - Cα-RMSD or TM-score (Zhang & Skolnick, 2004) - between predicted structures and the ground-truth templates, along with AF3 confidence scores, serve as a proxy for the likelihood that the designed sequence can adopt the target states.

Formally, for a protein with conformational states $X = \{X_1, X_2, \ldots, X_m\}$ and designed sequence Y, we define the AF3 RMSD for each target conformation X_k as:

$$\text{AF3}_{template}(Y, X_k) = \text{RMSD}(\text{AF3}(Y, X_k), X_k) \tag{4}$$

where $\text{AF3}_{template}(Y, X_k)$ is the structure predicted by AlphaFold3 for sequence Y when a template of X_k is provided. To account for the structural bias induced by the template, we define a normalization strategy to contextualize observed deviations:

Decoy normalization (Decoy Norm): We provide AF3 structurally dissimilar decoy structures as templates (TM-score < 0.4) using the same sequences designed and measure the resulting deviations. This control assesses whether sequences fold specifically into their targets or may fold equally well into arbitrary structures:

$$\text{RMSD}_{\text{decoy}}(Y, X_k; D) = \frac{\text{AF3}_{template}(Y, X_k)}{\text{AF3}_{template}(Y, D)} \tag{5}$$

where D is a decoy that is structurally dissimilar to X_k. Additionally, we measure pLDDT confidence scores to evaluate AF3 fold uncertainty. High RMSD with low pLDDT indicates poor template matching, while low RMSD with high pLDDT suggests a successful design. The same decoy normalization strategy is also applied to TM-score and pLDDT metrics.

Template-free baseline. To quantify the impact of template conditioning, we additionally evaluate refoldability without structural guidance by omitting the template input in AF3 (i.e. standard AF3 pipeline). Comparing these predictions against both target conformations reveals: (1) whether templates overly bias predictions toward target conformations, and (2) whether AF3 can recover alternative states or collapses to a single dominant structure.

Generative ensemble evaluation. As an orthogonal evaluation approach, we employ BioEmu (Lewis et al., 2025), a generative model trained to sample from protein conformational equilibrium distributions. BioEmu represents a class of emerging generative models specifically trained

to approximate conformational ensembles. We selected BioEmu over alternative ensemble generators such as AlphaFlow (Jing et al., 2024) or PepFlow (Abdin & Kim, 2023) as it represents the current state-of-the-art (Jing et al., 2025), trained on orders of magnitude more data than competing methods, including molecular dynamics equilibrium distributions and conformational clusters from AFDB (Varadi et al., 2024). However, current ensemble generation methods like BioEmu and AlphaFlow are restricted to single protein chains, precluding evaluation of refoldability in the presence of binding partners that often drive conformational changes and likely underestimating our designs' true multi-state capacity. For each designed sequence, we sample 50 structures (Fig. 5) and measure the maximum TM-score and minimum RMSD achieved against each target conformation.

Table 1: Refoldability performance comparison of DynamicMPNN model variants on multi-state protein design benchmark ($n = 96$). Raw metrics show absolute performance values, while normalized metrics show performance relative to random decoy structures.

	Raw Metrics			Decoy-Normalized Metrics		
Model Variant	**pLDDT ↑**	**RMSD (Å) ↓**	**TM-score ↑**	**pLDDT ↑**	**RMSD ↓**	**TM-score ↑**
Combined Training	82.08 (7.62)	2.35 (2.51)	0.870 (0.162)	1.354 (0.410)	**0.124** (0.130)	**6.684** (1.485)
Combined Training + DSS	81.61 (7.23)	2.56 (2.36)	0.862 (0.158)	1.398 (0.442)	0.131 (0.125)	6.627 (1.514)
Single Training	68.35 (15.46)	8.16 (9.96)	0.652 (0.322)	1.383 (0.467)	0.348 (0.336)	4.830 (2.285)
Sampled Pair Training	**82.26** (7.61)	**2.29** (1.92)	**0.872** (0.161)	**1.470** (0.466)	0.125 (0.129)	6.630 (1.506)
Sampled Pair Training + DSS	81.88 (7.85)	2.45 (2.33)	0.865 (0.158)	1.436 (0.470)	0.127 (0.121)	6.668 (1.466)

Table 2: Refoldability performance comparison of DynamicMPNN model variants on template-less multi-state protein design benchmark ($n = 96$).

	Template-less both states			Template-less worst state			With Template worst state		
Model Variant	**pLDDT ↑**	**RMSD (Å) ↓**	**TM-score ↑**	**pLDDT ↑**	**RMSD ↓**	**TM-score ↑**	**pLDDT ↑**	**RMSD ↓**	**TM-score ↑**
Combined Training	81.87 (7.99)	4.38 (4.05)	0.78 (0.20)	79.72 (9.89)	6.18 (5.92)	0.69 (0.27)	80.43 (9.50)	3.05 (3.41)	0.82 (0.22)
Single Training	70.11 (12.41)	10.87 (10.48)	0.54 (0.31)	67.65 (13.67)	12.56 (11.30)	0.47 (0.32)	66.49 (16.09)	9.34 (10.52)	0.59 (0.35)

3 Results and Discussion

Table 3: Sequence recovery performance comparison across DynamicMPNN model variants and ProteinMPNN baseline on multi-state protein design benchmark.

Model Variant	**Sequence Recovery (%) ↑**
Combined Pretraining + Multi Finetuning	**42.7** (8.8)
Single Pretraining + Multi Finetuning	42.1 (8.3)
Combined Pretraining + Multi Finetuning + DSS	41.9 (8.6)
Combined Training	41.0 (8.5)
Single Pretraining + Multi Finetuning + DSS	40.3 (8.2)
Combined Training + DSS	38.8 (7.8)
ProteinMPNN MSD*	38.0 (11.0)
Single chain 2-state	37.35 (9.04)
Single Training	27.1 (9.4)
Single Training + DSS	26.2 (8.5)

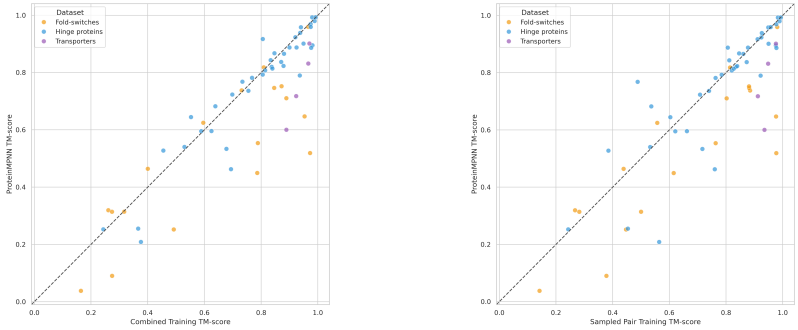
Setup. We evaluate how exposure to multi-state training data affects design performance. Our primary comparison is between *Single Training* (single-state pairs only, analogous to standard inverse-folding models) and *Combined Training* (mixing single-state and multi-state samples). We additionally explore multi-staged training where single-state pretraining precedes multi-state finetuning (*Multi-Finetuning*.

We train on two multi-conformational datasets and explore several architectural variations:

- **CoDNaS:** Multi-chain models encoding full PDB complexes including binding partners. Maximum RMSD conformer pairs are pre-selected per cluster prior to training.

Table 4: Performance comparison of DynamicMPNN variants and ProteinMPNN baseline on subset ($n = 61$) of multi-state protein design benchmark. Standard deviations shown in parentheses. Note that ProteinMPNN MSD's handling of gap tokens and missing residues (i.e., X tokens) limited the number of designs that could be refolded using AF3, necessitating separate comparison.

	Raw Metrics			Decoy-Normalized Metrics		
Model	**pLDDT ↑**	**RMSD (Å) ↓**	**TM-score ↑**	**pLDDT ↑**	**RMSD ↓**	**TM-score ↑**
Combined Training	82.22 (6.76)	**2.27** (1.59)	0.849 (0.166)	1.266 (0.332)	**0.129** (0.109)	6.482 (1.340)
Combined Training + DSS	82.00 (6.66)	2.65 (2.40)	0.836 (0.176)	1.286 (0.336)	0.144 (0.126)	6.513 (1.586)
Single Training	69.35 (14.45)	8.00 (9.44)	0.623 (0.324)	1.286 (0.396)	0.376 (0.338)	4.455 (2.186)
Sampled Pair Training	81.89 (7.36)	2.28 (1.75)	**0.850** (0.179)	**1.370** (0.405)	0.134 (0.127)	6.449 (1.521)
Sampled Pair Training + DSS	**82.45** (6.96)	2.58 (2.47)	0.839 (0.177)	1.341 (0.431)	0.139 (0.120)	**6.529** (1.570)
ProteinMPNN MSD*	79.55 (9.75)	3.31 (2.88)	0.806 (0.207)	1.326 (0.348)	0.187 (0.191)	6.054 (1.771)



(a) Combined Training vs ProteinMPNN MSD (n=61)

(b) Sampled Pair Training vs ProteinMPNN MSD (n=61)

Figure 3: Comparison between 2 DynamicMPNN versions and ProteinMPNN MSD on protein class-stratified TM-score AF3 refoldability. The classes are Fold-switches (yellow), Hinge proteins (blue; combining the APO-HOLO, OC23, and OC85 datasets), and Transporters (purple).

Table 5: BioEmu refoldability comparison of DynamicMPNN model variants on test set ($n = 96$). Standard deviations in parentheses.

	Performance Metrics		Success Rate (%)		
Model Variant	**TM-score ↑**	**RMSD (Å) ↓**	**$TMS_{0.7}$ ↑**	**$TMS_{0.8}$ ↑**	**$TMS_{0.9}$ ↑**
Combined Training	**0.623** (0.230)	**5.66** (4.82)	**37.5**	**22.9**	**11.5**
Single Training	0.394 (0.252)	13.28 (10.27)	17.7	7.3	3.1

Table 6: Comparison between 2-state and 3-state single chain models with average pooling or DSS for 3 curated 3-state proteins.

	Sequence Recovery (%) ↑				TM-score ↑			
Protein	**2-state**	**2-state + DSS**	**3-state**	**3-state + DSS**	**2-state**	**2-state + DSS**	**3-state**	**3-state + DSS**
MPD	50.3 (1.01)	62.3 (1.03)	51.3 (0.98)	56.1 (1.07)	0.93 (0.04)	0.90 (0.04)	0.88 (0.02)	0.89 (0.04)
Selecase	42.9 (2.81)	51.3 (2.21)	45.4 (1.22)	45.0 (1.90)	0.77 (0.03)	0.81 (0.01)	0.81 (0.03)	0.81 (0.01)
α-hemolysin	32.3 (0.87)	31.4 (1.24)	33.0 (1.10)	32.4 (0.92)	0.93 (0.03)	0.91 (0.03)	0.94 (0.02)	0.91 (0.03)

- **PDB80 (Sampled Pair Training):** Multi-chain models with TM-score weighted sampling to dynamically select conformer pairs during training, capturing greater conformational diversity.
- **Single-chain ablations:** To efficiently ablate designs beyond two states, we train single-chain models (encoding only target chains without binding partners) on PDB80, dynami-

Table 7: Comparison between 2-state and 5-state single chain models with average pooling or DSS for 3 curated highly flexible proteins.

	Sequence Recovery (%) ↑				TM-score ↑			
Protein	**2-state**	**2-state + DSS**	**5-state**	**5-state + DSS**	**2-state**	**2-state + DSS**	**5-state**	**5-state + DSS**
Aβ-42	27.8 (2.25)	9.8 (1.21)	17.4 (2.00)	16.8 (1.74)	0.08 (0.05)	0.16 (0.01)	0.08 (0.04)	0.07 (0.05)
α-synuclein	15.7 (1.20)	15.8 (1.47)	17.6 (2.80)	24.2 (1.71)	0.07 (0.02)	0.10 (0.03)	0.15 (0.01)	0.12 (0.08)
Calmodulin	34.2 (1.10)	34.0 (1.48)	43.4 (0.80)	43.0 (1.09)	0.47 (0.15)	0.46 (0.11)	0.51 (0.05)	0.58 (0.05)

cally sampling up to $k \in \{2, 3, 5\}$ conformational states per cluster using the same TM-score weighted scheme.

We also compare architectures with and without DSS cross-attention modules across these configurations. All models were trained on either 8 A100-80GB or 8 H100-80GB GPUs using a batch size of 32 and Adam (Kingma, 2014) optimizer with learning rate 10^{-3}. Training for each stage was run until convergence of performance on the validation set, typically after 20-40 epochs (≈17-34 hours).

Then, DynamicMPNN models and ProteinMPNN (using Multi-state Design inference strategy) were used to sample 16 sequences for each of the 96 benchmark test proteins (see Table 8 for DynamicMPNN inference costs and model size).

Evaluation protocol. For each designed sequence, we evaluate refoldability using the template-based pipeline described above, sampling 5 structures per sequence. For each target conformation, we select the best-performing sample (i.e., lowest RMSD, highest TM-score, highest pLDDT), then aggregate metrics by averaging across the 16 designed sequences. The same protocol is applied with decoy templates for normalization.

Additionally, for sequences designed using *Combined Training* and *Single Training* models, we conduct refoldability evaluations with BioEmu and template-free AF3. For BioEmu (Lewis et al., 2025), we sample 50 structures per sequence and report the best-performing sample per target conformation averaged across designs. We define success rates (TMS_τ) as the percentage of targets for which at least one designed sequence has both conformational states recovered with maximum TM-score (across the sampled ensemble) exceeding threshold $\tau \in \{0.7, 0.8, 0.9\}$.

For template-free AF3, we apply the same evaluation protocol as template-based AF3. To assess whether designs consistently recover both states, we additionally report worst-state metrics—the performance on whichever conformational state is harder to recover for each design—for both template-based and template-free evaluations (Table 2).

DynamicMPNN outperforms existing benchmarks across multiple evaluation metrics. Our best-performing model achieves substantial improvements over baseline methods: a 25% reduction in decoy-normalized RMSD (Tab. 4) and a 12% improvement in sequence recovery (Fig.4; Tab. 3) compared to ProteinMPNN Multi-State Design (MSD). This performance gain is particularly noteworthy considering that ProteinMPNN's training dataset contains proteins within 30% sequence similarity clusters of 91 out of 96 benchmark proteins, making our improvement on this established baseline significant. We leave the retraining of ProteinMPNN on a rigorous train-test split for future work.

Combined and multi-state finetuning training strategies far outperform the single-state only training. A more rigorous comparison involves our single-state trained model, which adheres to the same training-test split and thus eliminates potential data leakage concerns. Models trained only on single-state data (Single Training) perform poorly on multi-state design as they are unable to create and decode a meaningful latent representation of the conformational changes. The Combined Training and approaches, which exposes models to multi-state examples during training, achieve the optimal balance and consistently outperform Single Training as well as other models across all metrics. (Tab. 3, 1). Some likely reasons to explain the under performance of our Single Training checkpoint with respect to ProteinMPNN is the aforementioned unfair train splitting of ProteinMPNN as well as the multi-state latent space embeddings being out-of-distribution for the decoder parameters that have only been trained on single state data. Additionally, versions trained on PDB80 (Sampled Pair Training) with TM-weighted sampling slightly outperform those trained on CoDNaS

with MAX TM-score sampling, likely due to the increased diversity of conformational pairs seen during training 1.

Fold-switch designs show best improvements over ProteinMPNN. Analyzing performance stratified by the protein class in the test set (i.e. metamorphic, hinge, and transporter), (Fig. 3a, b), we observe that DynamicMPNN's advantage over ProteinMPNN is most pronounced in fold-switching proteins (metamorphic). This shows that DynamicMPNN can effectively fit the most complex conformational changes that undergo complex rearrangements of both tertiary and secondary structure (Dishman & Volkman, 2018). The distributions of AF3-predicted TM-scores over the inter-state TM-scores are plotted in Fig. 7.

Template-free AF3 is not suitable for evaluating multi-state designs. As has been previously shown for AlphaFold 2 (AF2) Chakravarty & Porter (2022), without structural guidance, AF3 collapses to a single dominant conformation, with worst-state TM-scores degrading from 0.82 (template-based) to 0.69 for Combined Training designs (Table 2). This inability to recover alternative states motivates our template-based evaluation framework. Notably, Combined Training outperforms Single Training across all conditions—template-free average (0.78 vs 0.54), worst-state (0.69 vs 0.47), and template-based (0.82 vs 0.59)—confirming multi-state training benefits persist regardless of evaluation protocol.

BioEMU mirrors AF3-template results. Both AF3 and BioEmu evaluation frameworks demonstrate consistent model rankings and highly correlated refoldability scores (Pearson $r = 0.71$ for TM-score, $r = 0.79$ for RMSD; decoy-normalized $r = 0.58$, $r = 0.61$), with Combined Training consistently outperforming Single Training. This further emphasizes that regardless of the in silico evaluation metric that we is used, models trained with multi-state data consistently outperform those without.

The optimal number of states is highly dependent on the protein system. While 2-state modelling is the most biologically relevant approach for most multi-state proteins, a question worth exploring is whether DynamicMPNN can be generalised to proteins with metastable states, intermediates or even proteins with flatter energy landscapes. We compared the performance of the 2-state single chain model to the 3-state and 5-state respectively on 3 proteins each: (1) 3 proteins with 2 main states with a stable intermediate (MPD, α-hemolysin, Selecase - Table 6); (2) 3 highly flexible proteins, either with disordered linkers (calmodulin) or fully disordered and which fold upon binding (α-synuclein, Aβ-42) - Table 7. Results show how including extra states can be both beneficial (calmodulin) as well as detrimental (MPD), as the model has to learn to reconcile more diverse structural information. The inclusion of DSS is equally inconclusive. This is probably due to the lack of such proteins in the training set. To assess whether the single chain nature of these models impacts the metrics, we also compare the 2-state single chain to the 2-state multi chain model, observing a moderate albeit expected drop in performance (Table 3).

4 Conclusion

We present DynamicMPNN, the first explicit multi-state inverse folding model, achieving up to 25% improvement over ProteinMPNN on our multi-state benchmark. By jointly learning across conformational ensembles rather than aggregating single-state predictions, DynamicMPNN captures sequence constraints required for multiple functional conformations. Critically, models trained with multi-state data consistently outperform single-state trained models across all evaluation metrics (template guided AF3, AF3 template-free, and BioEmu) demonstrating robust benefits of explicit multi-state training regardless of evaluation framework. This opens possibilities for engineering synthetic bioswitches, allosteric regulators, and molecular machines. It is unclear if the presented one-size-fits-all approach to multi-state design will be effective experimentally, or if specialized models for different classes of conformational changes depending on their thermodynamic complexity will be beneficial.

References

Osama Abdin and Philip M Kim. Pepflow: direct conformational sampling from peptide energy landscapes through hypernetwork-conditioned diffusion. *bioRxiv*, pp. 2023–06, 2023.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Josh Bambrick, Sebastian W Bodenstein, David A Evans, Chia-Chun Hung, Megan O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Elisabet Arvaniti, Charles Beattie, Octavian Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B Fuchs, Hannah Gladman, Rishub Jain, Yousuf A Khan, Caroline M R Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Cheng Tong, Stig Yakneen, Ellen D Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024a. doi: 10.1038/s41586-024-07487-w.

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, 2024b.

Robert G. Alberstein, Aidan B. Guo, and Tanja Kortemme. Design principles of protein switches. *Current Opinion in Structural Biology*, 72:71–78, Feb 2022. doi: 10.1016/j.sbi.2021.08.004. URL https://doi.org/10.1016/j.sbi.2021.08.004. PMID: 34537489; PMCID: PMC8860883.

Xavier I Ambroggio and Brian Kuhlman. Design of protein conformational switches. *Current Opinion in Structural Biology*, 16(4):525–530, 2006a. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2006.05.014. URL https://www.sciencedirect.com/science/article/pii/S0959440X06000984. Membranes / Engineering and design.

Xavier I. Ambroggio and Brian Kuhlman. Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society*, 128(4):1154–1161, 2006b. doi: 10.1021/ja054718w. URL https://doi.org/10.1021/ja054718w. PMID: 16433531.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Robert B Best, Kresten Lindorff-Larsen, Mark A DePristo, and Michele Vendruscolo. Relation between native ensembles and experimental structures of proteins. *Proceedings of the National Academy of Sciences*, 103(29):10901–10906, 2006.

N. Bordin, I. Sillitoe, V. Nallapareddy, C. Rauer, S. D. Lam, V. P. Waman, N. Sen, M. Heinzinger, M. Littmann, S. Kim, S. Velankar, M. Steinegger, B. Rost, and C. Orengo. Alphafold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications Biology*, 6(1):160, 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04488-9. URL https://doi.org/10.1038/s42003-023-04488-9.

Eleonora Cerasoli, Belinda K. Sharpe, and Derek N. Woolfson. Zico: A peptide designed to switch folded state upon binding zinc. *Journal of the American Chemical Society*, 127(43):15008–15009, 2005. doi: 10.1021/ja0543604. URL https://doi.org/10.1021/ja0543604. PMID: 16248623.

Debayan Chakravarty, Jonah W. Schafer, Elizabeth A. Chen, Daniel A. Keedy, Maximillian D. Smith, and Lauren L. Porter. Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nature Communications*, 15:7296, 2024. doi: 10.1038/s41467-024-51801-z. URL https://doi.org/10.1038/s41467-024-51801-z.

Devlina Chakravarty and Lauren L Porter. Alphafold2 fails to predict protein fold switching. *Protein Science*, 31(6):e4353, 2022.

A. Chatterjee, A. Roy, T. Satheesh, et al. Structural insights into pre-pore intermediates of alpha-hemolysin in the lipidic environment. *Nature Communications*, 16:6348, 2025. doi: 10.1038/s41467-025-61741-x. URL https://doi.org/10.1038/s41467-025-61741-x.

Jiaxing Chen, Sofia Zaer, Paz Drori, Joanna Zamel, Khalil Joron, Nir Kalisman, Eitan Lerner, and Nikolay V. Dokholyan. The structural heterogeneity of α-synuclein is governed by several distinct subpopulations with interconversion times slower than milliseconds. *Structure*, 29 (9):1048–1064.e6, 2021. ISSN 0969-2126. doi: 10.1016/j.str.2021.05.002. URL https://www.sciencedirect.com/science/article/pii/S096921262100160X.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

Acacia F. Dishman and Brian F. Volkman. Unfolding the mysteries of protein metamorphosis. *ACS Chemical Biology*, 13(6):1438–1446, 2018. doi: 10.1021/acschembio.8b00276. URL https://doi.org/10.1021/acschembio.8b00276. PMID: 29787234.

Alexandre Duval, Simon V. Mathis, Chaitanya K. Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D. Malliaros, Taco Cohen, Pietro Liò, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems, 2024. URL https://arxiv.org/abs/2312.07511.

Casper A. Goverde, Martin Pacesa, Nicolas Goldbach, Lars J. Dornfeld, Petra E. M. Balbi, Sandrine Georgeon, Stéphane Rosset, Srajan Kapoor, Jagrity Choudhury, Justas Dauparas, Christian Schellhaas, Simon Kozlov, David Baker, Sergey Ovchinnikov, Alex J. Vecchio, and Bruno E. Correia. Computational design of soluble and functional membrane protein analogues. *Nature*, 631(8020):449–458, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07601-y. URL http://dx.doi.org/10.1038/s41586-024-07601-y.

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. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8946–8970. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Arian R. Jamasb, Alex Morehead, Chaitanya K. Joshi, Zuobai Zhang, Kieran Didi, Simon V. Mathis, Charles Harris, Jian Tang, Jianlin Cheng, Pietro Lio, and Tom L. Blundell. Evaluating representation learning on the protein structure universe. In *The Twelfth International Conference on Learning Representations*, 2024.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons, 2021. URL https://arxiv.org/abs/2009.01411.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles, 2024. URL https://arxiv.org/abs/2402.04845.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Ai-based methods for simulating, sampling, and predicting protein ensembles. *arXiv preprint arXiv:2509.17224*, 2025.

Chaitanya K. Joshi, Arian R. Jamasb, Ramon Viñas, Charles Harris, Simon V. Mathis, Alex Morehead, Rishabh Anand, and Pietro Liò. grnade: Geometric deep learning for 3d rna inverse design, 2025. URL https://arxiv.org/abs/2305.14749.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Y. Kalakoti and B. Wallner. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications Biology*, 8(1):373, 3 2025. doi: 10.1038/s42003-025-07791-9. URL https://doi.org/10.1038/s42003-025-07791-9. PMID: 40045015; PMCID: PMC11882827.

Mostafa Karimi and Yang Shen. icfn: an efficient exact algorithm for multistate protein design. *Bioinformatics*, 34(17):i811–i820, 2018.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)-round xiii. *Proteins*, 87(12): 1011–1020, 12 2019. doi: 10.1002/prot.25823. URL https://doi.org/10.1002/prot.25823. PMID: 31589781; PMCID: PMC6927249.

Thomas J Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, 20(2):170–173, 2023.

Andrew Leaver-Fay, Robert Jacak, P Brian Stranges, and Brian Kuhlman. A generic program for multistate protein design. *PLoS One*, 6(7):e20937, Jul 2011. doi: 10.1371/journal.pone.0020937. URL https://doi.org/10.1371/journal.pone.0020937. PMID: 21754981; PMCID: PMC3130737.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E. Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025. doi: 10.1126/science.adv9817.

Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel W. K. Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J. Hendel, Miriam K. Simma, Ge Liu, Muna Yase, Hongwei Wu, Claire D. Tharp, Xinting Li, Alex Kang, Evans Brackenbrough, Asim K. Bera, Stacey Gerben, Bruce J. Wittmann, Andrew C. McShan, and David Baker. Multistate and functional protein design using rosettafold sequence space diffusion. *Nature Biotechnology*, 43(8):1288–1298, September 2024. ISSN 1546-1696. doi: 10.1038/s41587-024-02395-w. URL http://dx.doi.org/10.1038/s41587-024-02395-w.

Yang Liu and Brian Kuhlman. Rosettadesign server for protein design. *Nucleic Acids Research*, 34(Web Server issue):W235–W238, 7 2006. doi: 10.1093/nar/gkl163. URL https://doi.org/10.1093/nar/gkl163. PMID: 16845000; PMCID: PMC1538902.

Mar López-Pelegrín, Núria Cerdà-Costa, Anna Cintas-Pedrola, Fátima Herranz-Trillo, Pau Bernadó, Juan R. Peinado, Joan L. Arolas, and F. Xavier Gomis-Rüth. Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metallopeptidase. *Angewandte Chemie International Edition*, 52(20):5212–5217, 2013. doi: 10.1002/anie.201300923.

Haggai Maron, Or Litany, Gal Chechik, and Ethan Fetaya. On learning sets of symmetric elements, 2020. URL https://arxiv.org/abs/2002.08599.

Alessandro Mirarchi, Toni Giorgino, and Gianni De Fabritiis. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11:1299, 2024. doi: 10.1038/s41597-024-04140-z. URL https://doi.org/10.1038/s41597-024-04140-z.

Alexander M. Monzon, Christian O. Rohr, Maria Silvina Fornasari, and Gustavo Parisi. Codnas 2.0: a comprehensive database of protein conformational diversity in the native state. *Database*, 2016:baw038, 2016. doi: 10.1093/database/baw038. URL https://doi.org/10.1093/database/baw038. PMID: 27022160; PMCID: PMC4809262.

Sarfaraz K. Niazi. Critical assessment of ai-based protein structure prediction: Fundamental challenges and future directions in drug discovery. *Computational and Structural Biotechnology Reports*, 2:100064, 2025. ISSN 2950-3639. doi: 10.1016/j.csbr.2025.100064. URL https://www.sciencedirect.com/science/article/pii/S2950363925000353.

Lauren L. Porter and Loren L. Looger. Extant fold-switching proteins are widespread. *Proceedings of the National Academy of Sciences*, 115(23):5968–5973, 2018. doi: 10.1073/pnas.1800168115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1800168115.

Florian Praetorius, Philip J. Y. Leung, Maxx H. Tessmer, Adam Broerman, Cullen Demakis, Acacia F. Dishman, Arvind Pillai, Abbas Idris, David Juergens, Justas Dauparas, Xinting Li, Paul M. Levine, Mila Lamb, Ryanne K. Ballard, Stacey R. Gerben, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Brian F. Volkman, Jeff Nivala, Stefan Stoll, and David Baker. Design of stimulus-responsive two-state hinge proteins. *Science*, 381(6659):754–760, 2023. doi: 10.1126/science.adg7731. URL https://www.science.org/doi/abs/10.1126/science.adg7731.

Alfredo Quijano-Rubio, Hsien-Wei Yeh, Jooyoung Park, et al. De novo design of modular and tunable protein biosensors. *Nature*, 591:482–487, 2021. doi: 10.1038/s41586-021-03258-z. URL https://doi.org/10.1038/s41586-021-03258-z.

James P. Roney and Sergey Ovchinnikov. State-of-the-art estimation of protein model accuracy using alphafold. *Phys. Rev. Lett.*, 129:238101, Nov 2022. doi: 10.1103/PhysRevLett.129.238101. URL https://link.aps.org/doi/10.1103/PhysRevLett.129.238101.

Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2):85–94, February 1999. doi: 10.1093/protein/12.2.85. URL https://doi.org/10.1093/protein/12.2.85.

Tadeo Saldaño, Nahuel Escobedo, Julia Marchetti, Diego Javier Zea, Juan Mac Donagh, Ana Julia Velez Rueda, Eduardo Gonik, Agustina García Melani, Julieta Novomisky Nechcoff, Martín N Salas, Tomás Peters, Nicolás Demitroff, Sebastian Fernandez Alberti, Nicolas Palopoli, Maria Silvina Fornasari, and Gustavo Parisi. Impact of protein conformational diversity on alphafold predictions. *Bioinformatics*, 38(10):2742–2748, 04 2022.

C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991. doi: 10.1002/prot.340090107. URL https://doi.org/10.1002/prot.340090107.

Viktor Stein and Kirill Alexandrov. Synthetic protein switches: design principles and applications. *Trends in Biotechnology*, 33(2):101–110, 2015. ISSN 0167-7799. doi: https://doi.org/10.1016/j.tibtech.2014.11.010. URL https://www.sciencedirect.com/science/article/pii/S016777991400239X. Special Issue: Manifesting Synthetic Biology.

Giulio Tesei, Anders Irbäck Trolle, Niels Jonsson, Johannes Betz, Frederik Elberling Knudsen, Francesco Pesce, Kristoffer E. Johansson, and Kresten Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626(8000):897–904, Feb 2024. doi: 10.1038/s41586-023-07004-5. URL https://doi.org/10.1038/s41586-023-07004-5. PMID: 38297118.

Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends in Biochemical Sciences*, 33(1):2–8, 2008. ISSN 0968-0004. doi: https://doi.org/10.1016/j.tibs.2007.10.003. URL https://www.sciencedirect.com/science/article/pii/S096800040700285X.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1084. URL https://doi.org/10.1093/nar/gkad1084.

Mihaly Varadi, Damiano Bertoni, Pauline Magana, Urmila Paramval, Ivanna Pidruchna, Mandar Radhakrishnan, Mihail Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Andrew Laydon, Alexander Žídek, Harry Tomlinson, Dhanu Hariharan, Joel Abrahamson, Tristan Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, January 2024. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

Jelena Vucinic, David Simoncini, Manon Ruffini, Sophie Barbe, and Thomas Schiex. Positive multistate protein design. *Bioinformatics*, 36(1):122–130, 2020.

Chuanrui Wang, Bozitao Zhong, Zuobai Zhang, Narendra Chaudhary, Sanchit Misra, and Jian Tang. Pdb-struct: A comprehensive benchmark for structure-based protein design. *arXiv preprint arXiv:2312.00080*, 2023.

Yong Wang, Chun Tang, Erkang Wang, and Jin Wang. Exploration of multi-state conformational dynamics and underlying global functional landscape of maltose binding protein. *PLOS Computational Biology*, 8(4):e1002471, Apr 2012. doi: 10.1371/journal.pcbi.1002471. URL https://doi.org/10.1371/journal.pcbi.1002471.

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, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL http://dx.doi.org/10.1038/s41586-023-06415-8.

Kathy Y. Wei, Dimitar Moschidi, Matthew J. Bick, Santrupti Nerli, Andrew C. McShan, Lauren P. Carter, Po-Ssu Huang, Daniel A. Fletcher, Nikolaos G. Sgourakis, Scott E. Boyken, and David Baker. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proceedings of the National Academy of Sciences*, 117(13):7208–7215, 2020. doi: 10.1073/pnas.1914808117. URL https://doi.org/10.1073/pnas.1914808117. PMID: 32188784; PMCID: PMC7132107.

Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with tm-score = 0.5? *Bioinformatics*, 26(7):889–895, Apr 2010. doi: 10.1093/bioinformatics/btq066. URL https://doi.org/10.1093/bioinformatics/btq066.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nature Methods*, 19(9):1109–1115, September 2022a. doi: 10.1038/s41592-022-01585-1. URL https://doi.org/10.1038/s41592-022-01585-1.

J. Z. Zhang, H. W. Yeh, A. C. Walls, et al. Thermodynamically coupled biosensors for detecting neutralizing antibodies against sars-cov-2 variants. *Nature Biotechnology*, 40:1336–1340, 2022b. doi: 10.1038/s41587-022-01280-8. URL https://doi.org/10.1038/s41587-022-01280-8.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33(7):2302–2309, Apr 2005. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524.

A Appendix

A.1 LLM acknowledgement

The authors acknowledge that they have used LLMs in the process of paper writing for style suggestions and proofreading.

A.2 Supplementary Results

Table 8: Inference cost for sampling 16 sequences using one NVIDIA A100-80GB GPU and 32 AMD EPYC 7763 CPU cores. Residue count includes target and binding partner chains. DynamicMPPN model sizes: 4.86M parameters (+DSS) and 4.23M parameters (no DSS).

Total Residues	Time (s) +DSS	Time (s) No DSS	Peak GPU Memory (MB) +DSS	Peak GPU Memory (MB) No DSS
250	22.5	13.8	755	373
500	34.1	20.8	1012	514
1000	51.5	31.3	1358	710
2000	77.9	47.2	1824	984
4000	117.7	71.1	2453	1365
8000	177.9	107.0	3302	1896

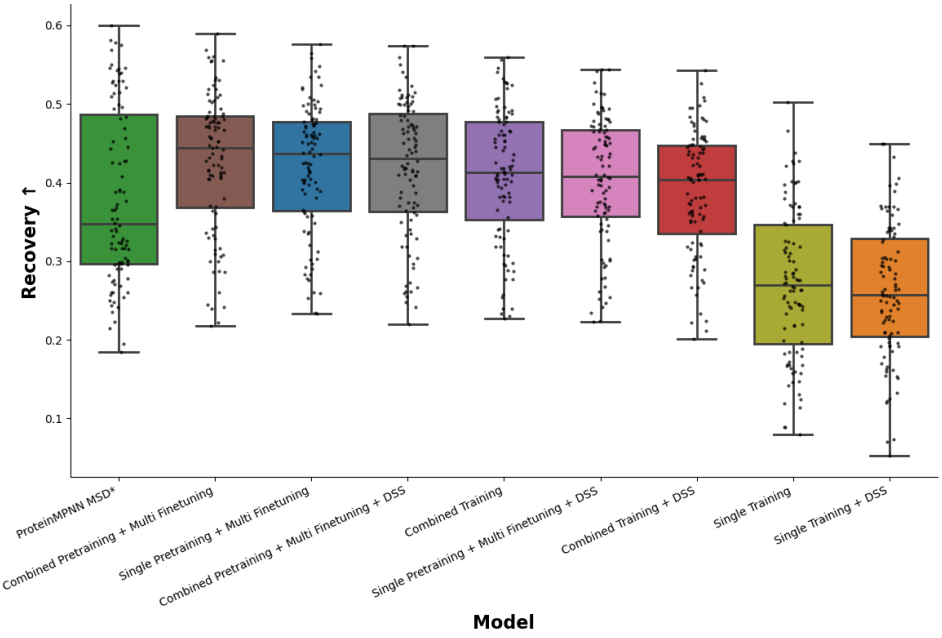


Figure 4: Sequence recovery performance across DynamicMPNN model variants and ProteinMPNN baseline on multi-state protein benchmark ($n = 96$). Combined training approaches achieve highest performance, with models that only incorporate single state training data performing poorly.

A.3 Details on Previous Work

A.3.1 Comparison of Multi-State Inference Strategies

Both ProteinMPNN-MSD (Dauparas et al., 2022; Praetorius et al., 2023) and Multi-state ESM-IF (Hsu et al., 2022) extend single-state inverse folding models to multi-state design tasks using an identical underlying mathematical strategy: step-wise logit aggregation with shared autoregressive context.

Hsu et al. (2022) formulate the objective as maximizing the geometric average of the conditional likelihoods for two states A and B. Since sampling only depends on relative logit values (softmax normalizes), averaging logits before sampling yields equivalent results to the geometric mean of probabilities:

$$\underbrace{\sqrt{P(y_t|X_A) \cdot P(y_t|X_B)}}_{\text{Geometric Mean (ESM-IF)}} \propto \exp \underbrace{\left(\frac{\text{Logits}(y_t|X_A) + \text{Logits}(y_t|X_B)}{2} \right)}_{\text{Arithmetic Mean of Logits (ProteinMPNN-MSD)}} \tag{6}$$

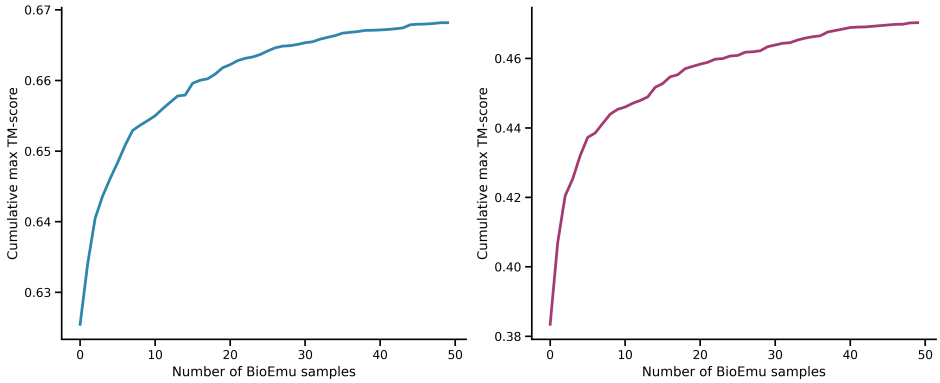


Figure 5: Convergence of cumulative maximum TM-score for BioEmu samples of Combined Training (left) and Single Training (right) model sequences. The cumulative maximum across 16 sequence designs is averaged over all test set targets.

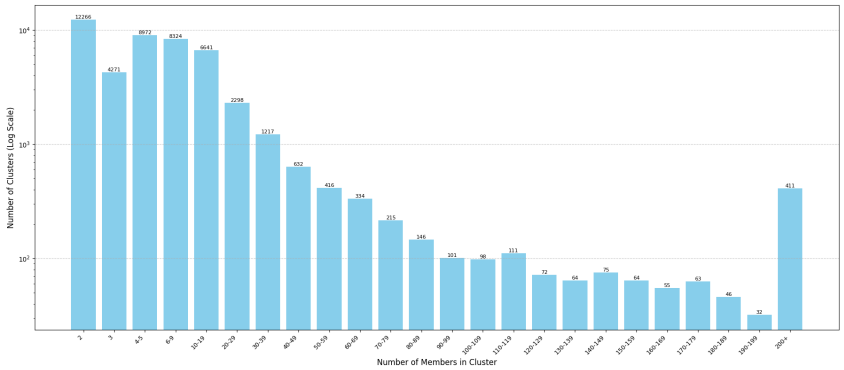


Figure 6: Distribution of cluster sizes in the PDB80 multi-conformational dataset. The histogram shows bins of cluster population sizes, illustrating the unbalanced nature of the dataset with most clusters containing few conformations while a small number of clusters contain thousands of chains.

Consequently, both methods perform the same operation: averaging the logits from independent structural encoders at step t, sampling a token, and feeding that single consensus token back into the autoregressive context for all states at step $t + 1$.

Weighted aggregation. While Hsu et al. (2022) used unweighted logit averaging, Praetorius et al. (2023) applied in their hinge protein design a 40%-60% weighting scheme to bias designs towards the effector-bound (holo) state.

A.3.2 Experimental validation of hinge protein designs by Praetorius et al. (2023)

Out of the 9K designed sequences using ProteinMPNN-MSD and Rosetta (Liu & Kuhlman, 2006), 76 were randomly selected for experimental validation based on RosettaDesign (Liu & Kuhlman, 2006) and AlphaFold2 (pLDDT ¿ 92, RMSD ¡ 1.5, PAE ¡ 5) (Jumper et al., 2021) filters. 46 out of 79 designs were expressed solubly and predominantly monomeric. Because many of the corresponding effector peptides were insoluble, only 20 out the 46 designs were tested for binding. 9 out of 20 designs showed binding via SEC experiments; advanced structural characterization with DEER and FRET confirmed the presence of the desired conformational change and peptide binding for all 8 and 3 respectively tested designs. in 3 out of these 9 designs.

(a) ProteinMPNN (n=61)

(b) Combined Training (n=96)

(c) Combined Training + DSS (n=96)

(d) Sample Pair Training + DSS (n=96)

(e) Sample Pair Training (n=96)

(f) Single Training (n=96)

Figure 7: Distribution of Predicted template-AF3 TM-score of samples against the inter-state ground-truth TM-score. All multi-state trained model prediction accuracy correlate well with the inter-state TM-scores.

A.4 Dataset Details

To construct the dataset, we obtained 46,033 Multiple Sequence Alignment (MSA) clusters at $\geq 95\%$ local sequence similarity from the latest version of CoDNaS (`v2025`) (Monzon et al., 2016), including NMR model structures. Importantly, the CoDNaS dataset prevents potential errors created by mixing different homologues in the same cluster by enforcing the same UniProt ID for all custer members. All available conformations of a protein are included - as different experimental conditions and sequence variations can reveal distinct thermodynamic states of the same protein (Best et al., 2006). While clusters contain varying numbers of conformations (Fig. 2b) we constructed our dataset using only pairs of chains from one or two PDB entries that have the largest RMSD from each cluster (Fig. 2c). Chosen pairs represent the most distinct conformational states.

While other inverse folding models (Hsu et al., 2022) saw improved performance when trained on orders of magnitude of more protein structures from AFDB, previous studies have found that the majority of high-confidence structures in AFDB map to known CATH superfamilies (Bordin et al., 2023), and that AlphaFold struggles in predicting alternative states (Chakravarty et al., 2024). We therefore decided against including AFDB structures in our training set.

PDB80 dataset, on the other hand, is meant to include more structural redundancy via its lower sequence similarity threshold (80%) at the expense of homology leakage and introduction of mutational fold-switches.

To calculate the pairwise TM-scores between the cluster members, we use the TM-align algorithm (Zhang & Skolnick, 2005). To reduce the computational cost, for clusters larger than 10 members, we perform structural clustering within each sequence cluster with a threshold of TM-score=0.9

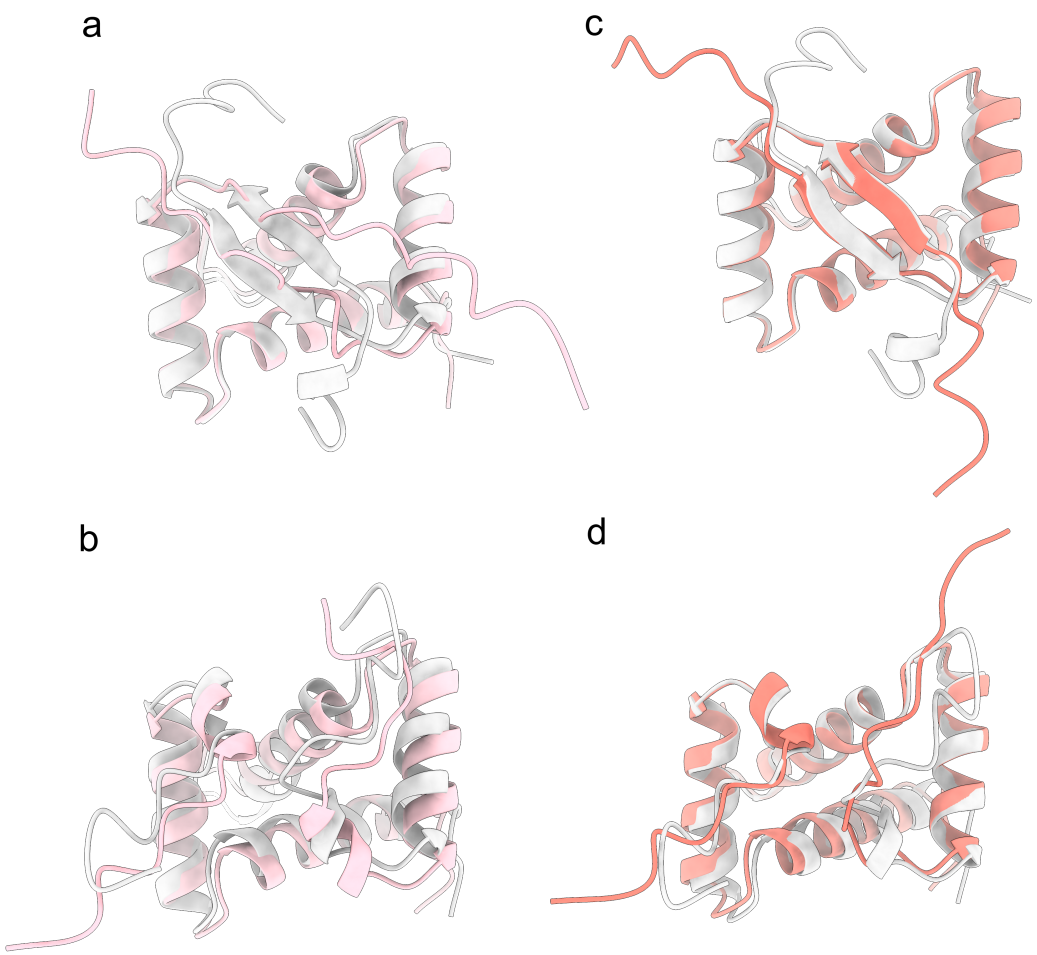


Figure 8: Switch Arc protein case study. (a, b) ProteinMPNN and (c,d) DynamicMPNN best design structure prediction (pink and salmon, respectively) against both Arc states from PDB ID: 1BDT and 1QTG respectively (grey). The DynamicMPNN design recapitulates the beta sheet fold (c), but the ProteinMPNN design0 does not (a).

using qTMclust (part of USalign suite - Zhang et al. (2022a)). We select one representative from each cluster, if the number of clusters is less than 10, extra representatives are randomly sampled to top up the selections to 10 members, and only then the pairwise TM-scores are computed as described before.

During pair sampling in Sample Pair Training, a pair is selected with a probability as a function of the TM-score: for sequential selection of k conformations, we sample each structure with probability proportional to its dissimilarity from already-selected structures:

$$P(c_i) = \frac{\exp\left(\frac{1-\overline{\text{TM}}_i}{\tau}\right)}{\sum_{j\in\mathcal{C}} \exp\left(\frac{1-\overline{\text{TM}}_j}{\tau}\right)} \tag{7}$$

where $\overline{\text{TM}}_i = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \text{TM}(c_i, s)$ is the average TM-score between candidate c_i and the set of already-selected structures $\mathcal{S}$, $\mathcal{C}$ is the set of remaining candidates, and $\tau = 2.0$ is a temperature parameter.

A.5 Model Details

A.5.1 Featurisation scheme

We use a similar featurisation scheme as in (Jamasb et al., 2024). **Node scalar features** are transformer-like positional encoding in a 16-dimensional array; backbone dihedral angles $\phi, \psi, \omega \in \mathbb{R}^6$; the virtual torsion and virtual bond angle $\kappa, \alpha \in \mathbb{R}^4$. **Node vector features** are position vectors of C_α, $\tilde{x_i} \in \mathbb{R}^3$. **Edge scalar features** are established via k-NN (k=16) and the edge length expressed in 32 Radial Basis Functions, $e_{RBF} \in \mathbb{R}^{32}$, as well as the length of the edge itself. **Edge vector features** are edge directional unit vectors for both directions $v_{\tilde{e^{ij}}} = \tilde{x_i} - \tilde{x_j}$. To further prevent overfitting on crystallisation artifacts, random Gaussian noise ($\bar{x} = 0, \sigma = 0.1$Å) was added to the coordinates (Dauparas et al., 2022).

A.5.2 Multi-state GNN

DynamicMPNN processes one or multiple protein backbone graphs via a multi-state GNN encoder (Joshi et al., 2025). Overall, DynamicMPNN's encoder is equivariant to 3D roto-translation of coordinates as well as ordering of the states in its input. Encoding is followed by pooling node features across states, which is invariant to the ordering of the states, and autoregressive sequence decoding.

When representing conformational ensembles as a multi-graph, each node feature tensor contains three axes: (#nodes, #conformations, feature channels). Multi-state GNN's encode multi-graphs by performing message passing on the multi-graph adjacency to *independently* process each conformer, while maintaining permutation equivariance of the updated feature tensors along both the first (#nodes) and second (#conformations) axes.

A.5.3 Geometric Vector Perceptron layers

Geometric Vector Perceptrons (GVPs) (Jing et al., 2021) are a generalization of MLPs to take tuples of scalar and vector features as input and apply $O(3)$-equivariant non-linear updates. GVP GNN layers process scalar and vector features on separate channels to maintain equivariance. The node scalars $\mathbf{s}_i \in \mathbb{R}^{k \times m}$, node vectors $\tilde{\mathbf{v}}_i \in \mathbb{R}^{k \times m' \times 3}$, and edge scalars $\mathbf{e}_{ij}$ and vectors $\tilde{\mathbf{e}}_{ij}$ communicate through a message passing operation:

$$\mathbf{m}_i, \tilde{\mathbf{m}}_i := \sum_{j \in N_i} \text{GVP}\left((\mathbf{s}_i, \tilde{\mathbf{v}}_i), (\mathbf{s}_j, \tilde{\mathbf{v}}_j), \mathbf{e}_{ij}, \tilde{\mathbf{e}}_{ij}\right), \qquad \text{(Message \& aggregate steps)} \qquad (8)$$

$$\mathbf{s'}_i, \tilde{\mathbf{v}'}_i := \text{GVP}\left((\mathbf{s}_i, \tilde{\mathbf{v}}_i), (\mathbf{m}_i, \tilde{\mathbf{m}}_i)\right). \qquad \text{(Update step)} \qquad (9)$$

The overall GNN encoder is $SO(3)$-equivariant due to the use of reflection-sensitive input features (dihedral angles) combined with $O(3)$-equivariant GVP-GNN layers.

A.5.4 Conformation order-invariant pooling

After using message passing layers that are conformation order-equivariant, we add a conformation order-invariant head, which performs average pooling across the conformation channel of the scalar and vector feature tensors, similar to Joshi et al. (2025) (2025): $\mathbf{S} \in \mathbb{R}^{n \times k \times m}$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{n \times k \times m' \times 3}$ to $\mathbf{S} \in \mathbb{R}^{n \times m}$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{n \times m' \times 3}$, where n is the sequence length, k is the number of backbones, m is the number of scalar features, and m' is the number of vector features. The only pooling strategy used in this work is the pooling of the maximum RMSD pair of chains - therefore $k = 2$ - although more pooling strategies for homo-oligomers can be used, such as equal averaging of all chains to be inverse folded in the selected PDB entries.

A.6 Metrics

A.6.1 Decoy-normalised AF3 self-consistency evaluation

AlphaFold as a Biophysical Energy Function. While AlphaFold (AF) excels at predicting static structures from evolutionary data, it often struggles to spontaneously sample alternative conformational states for a single sequence, effectively failing at the global *search* problem for multi-state

proteins. However, recent work by Roney & Ovchinnikov (2022) established that AF has learned a robust, coevolution-independent biophysical energy function that can accurately score sequence-structure compatibility when the search space is constrained (Roney & Ovchinnikov, 2022). Specifically, they demonstrated that when a candidate structure is provided as a template, AF's output confidence metrics (pLDDT, pTM) and structural consistency (TM-score between input template and output) correlate strongly with the actual accuracy of the model, effectively acting as a state-of-the-art energy scoring function.

Leveraging this finding, we utilize AlphaFold 3 (AF3) (Abramson et al., 2024a) not as a search engine to find the conformations, but as a scoring function to *evaluate* the compatibility of our designed sequences with the specific target geometries. By explicitly providing the target conformational state X_k as a template, we direct the model to the relevant basin of the energy landscape.

Addressing Template Bias via Decoy Normalization. A critical challenge in template-guided evaluation is the potential for "template bias," where the model might simply copy the input geometry regardless of the sequence's actual propensity to fold into that state. To distinguish between true sequence-structure compatibility and template memorization, we introduce a **decoy-normalization** strategy.

We define a set of decoy structures D that are structurally dissimilar to the target (TM-score < 0.4) but represent valid globular protein folds. For a designed sequence Y and a target conformational state X_k, we calculate a normalized score that compares the structural self-consistency on the target against the self-consistency on a decoy:

$$\text{Normalized Score} = \frac{\text{Self-Consistency}(Y, X_k)}{\text{Self-Consistency}(Y, D)} \tag{10}$$

In the context of RMSD (where lower is better), this is formulated as:

$$\text{RMSD}_{\text{decoy}}(Y, X_k; D) = \frac{\text{RMSD}(\text{AF3}(Y, X_k), X_k)}{\text{RMSD}(\text{AF3}(Y, D), D)} \tag{11}$$

Physical Interpretation. This metric serves as a proxy for the **specificity gap** or energy gap (ΔE) between the target fold and competing misfolded states.

- **High Specificity (Successful Design):** The sequence is highly compatible with the target (low RMSD / high pLDDT when prompted with X_k) but incompatible with the decoy (high RMSD / low pLDDT when prompted with D). This results in a favorable normalized score, indicating the sequence "accepts" the target fold and "rejects" the decoy.
- **Low Specificity (Hallucination/Promiscuity):** If the sequence creates a low-energy structure for both the target *and* the decoy (or high error for both), the normalized score approaches 1.0. This identifies sequences that are either generically "sticky" or for which the model is over-relying on template inputs without sequence support.

By requiring our designs to outperform decoys, we extend the "AF-as-energy-function" paradigm from simple ranking (as proposed by Roney & Ovchinnikov (Roney & Ovchinnikov, 2022)) to a rigorous **specificity filter** for *de novo* multi-state design.