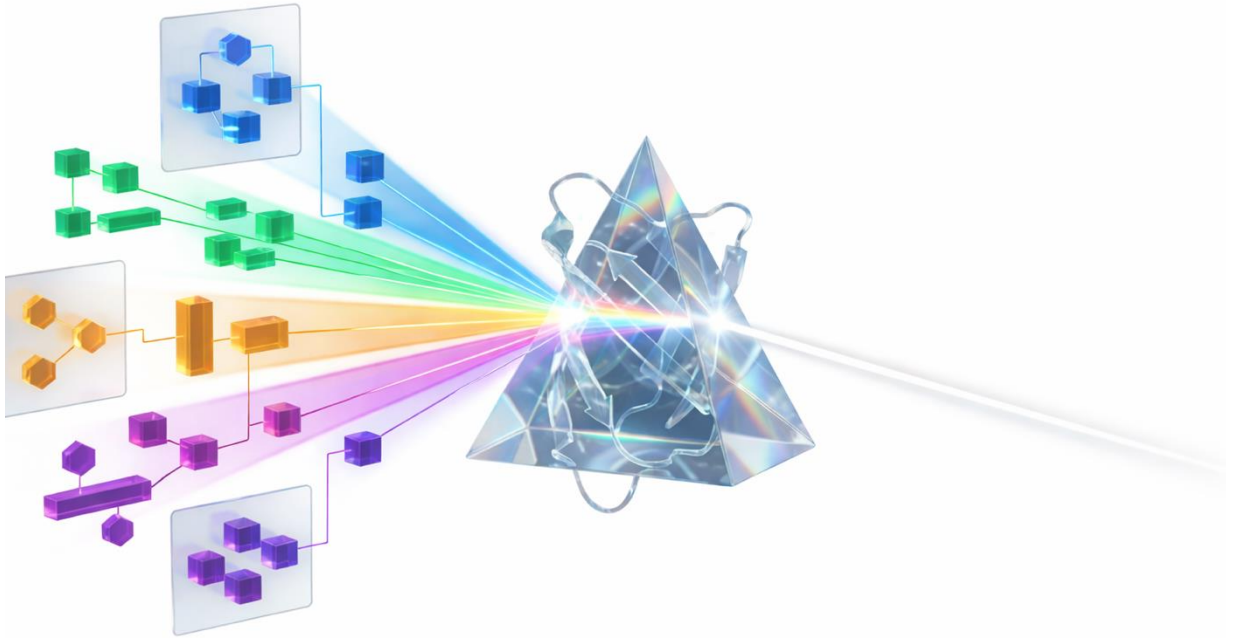
PRISM: Enhancing **PR**otein **I**nverse Folding Through Fine-Grained Retrieval on **S**tructure-Sequence **M**ultimodal Representations

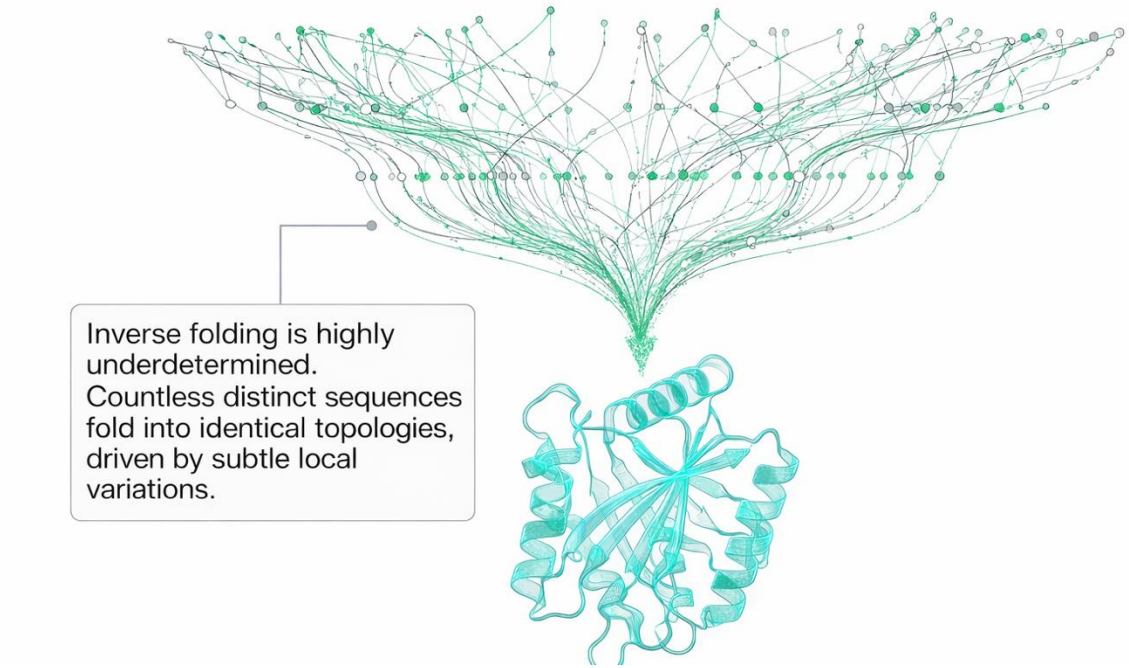
Sazan Mahbub[c], Souvik Kundu[i], Eric Xing[c,m,g]

[c]Carnegie Mellon University, [m]Mohamed bin Zayed University of AI, [g]GenBio AI, [i]Intel



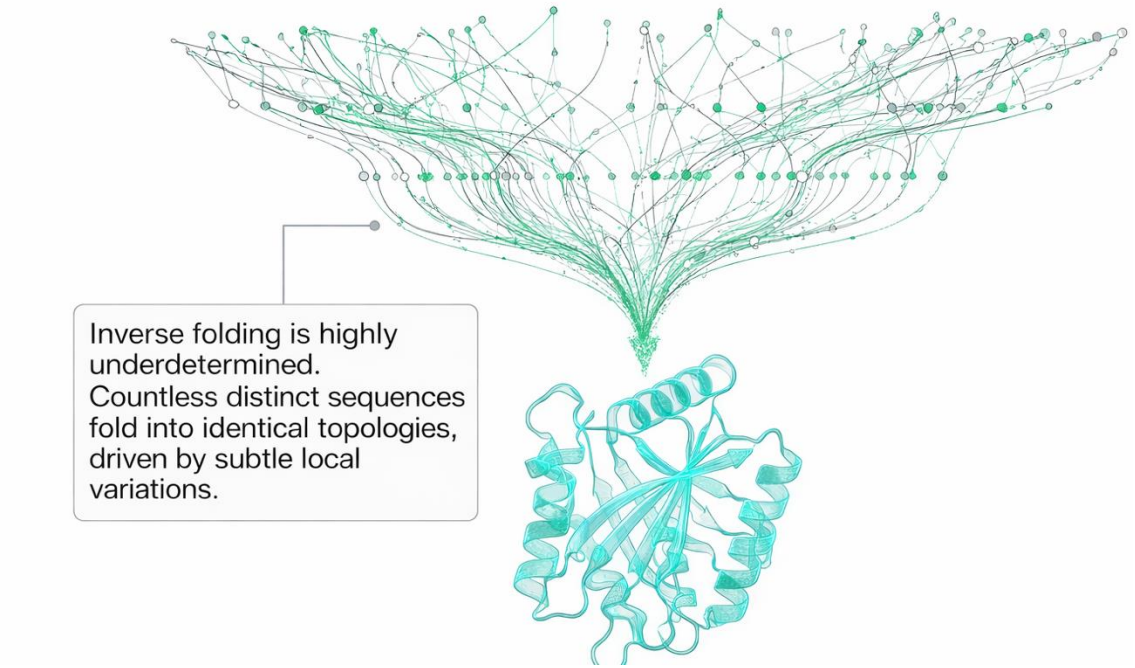
The Inverse Folding Bottleneck: A Lack of Explicit Pattern Reuse

The Problem Space:

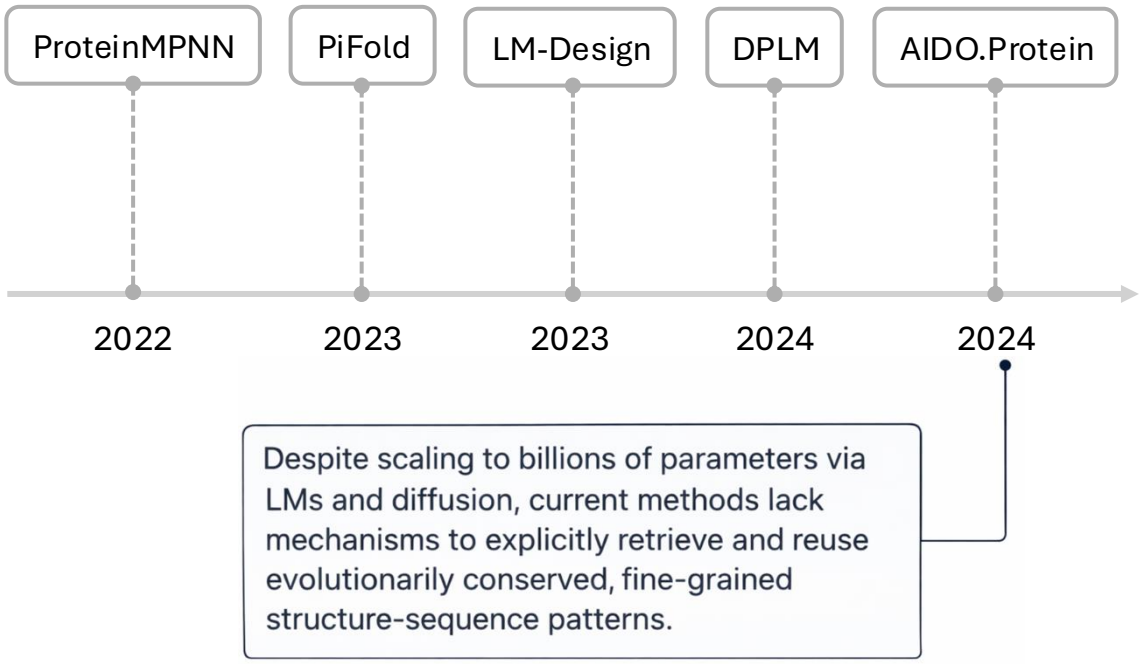


The Inverse Folding Bottleneck: A Lack of Explicit Pattern Reuse

The Problem Space:

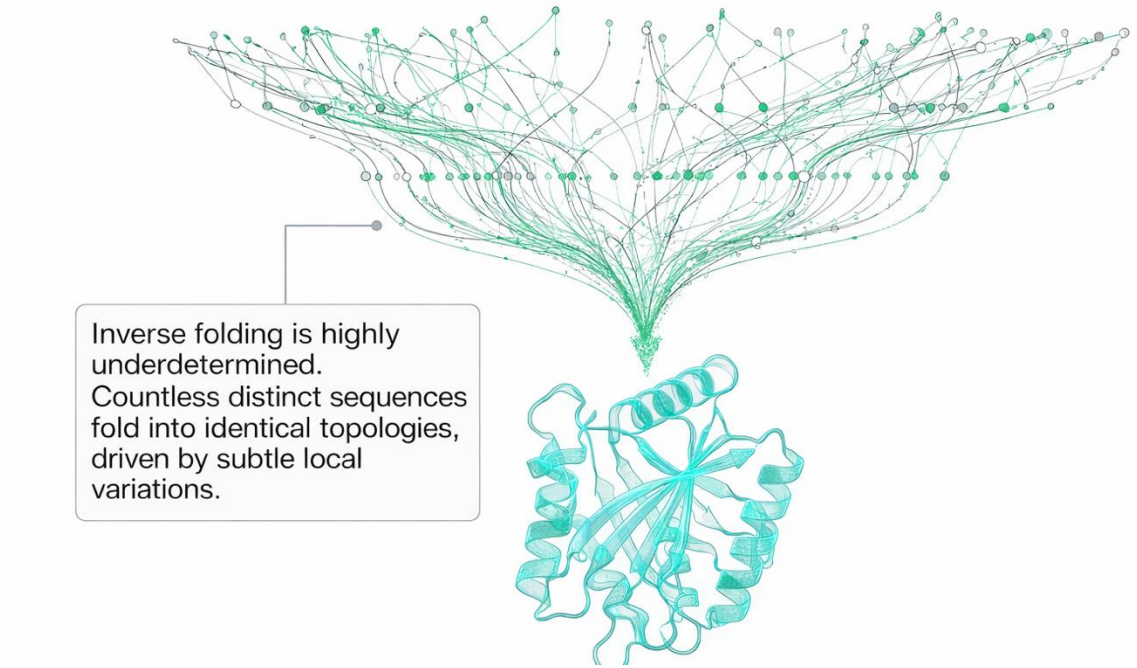


The Architectural Gap:

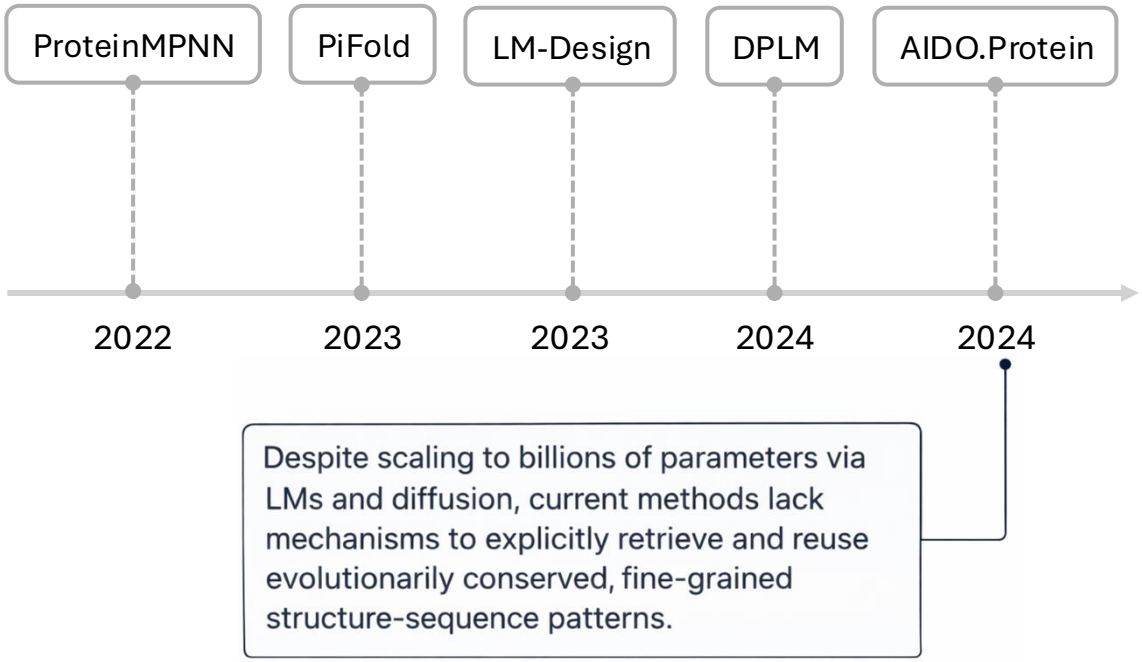


The Inverse Folding Bottleneck: A Lack of Explicit Pattern Reuse

The Problem Space:



The Architectural Gap:

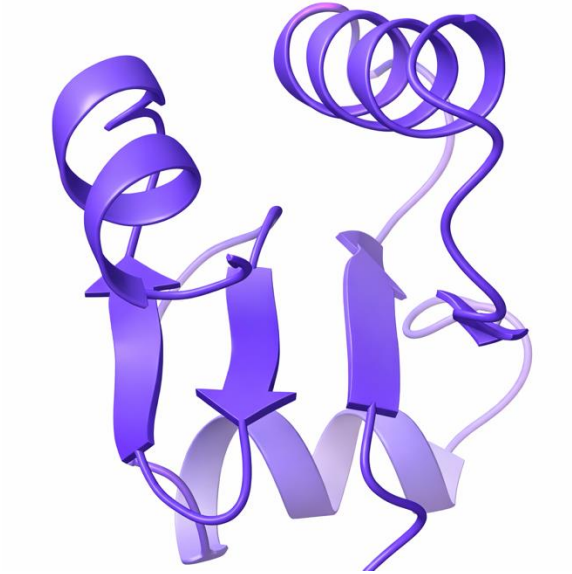


The PRISM Insight: Inverse folding can benefit from treating local structure-sequence neighborhoods as explicit, memory-retrievable building blocks.

Defining the Fundamental Unit: From Canonical to Potential Motifs

Definition 3.1: Protein Motif (Canonical)

Definition 3.1 (Protein Motif)**.** *It is a recurring local structural–sequential pattern of residues that is evolutionarily conserved and often functionally significant. Formally, it can be described as a short stretch of amino acids together with its surrounding 3-D conformation, capturing local folding rules and biochemical properties independent of the global protein context.*

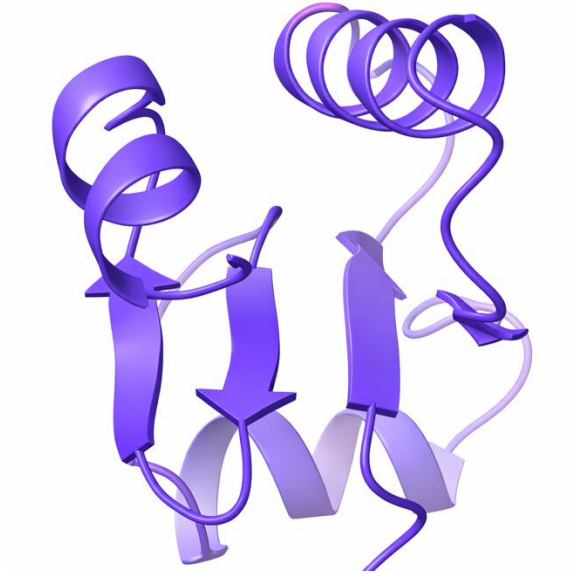


Key Trait: Evolutionarily conserved, functionally significant stretch of amino-acids with surrounding 3D conformation.

Defining the Fundamental Unit: From Canonical to Potential Motifs

Definition 3.1: Protein Motif (Canonical)

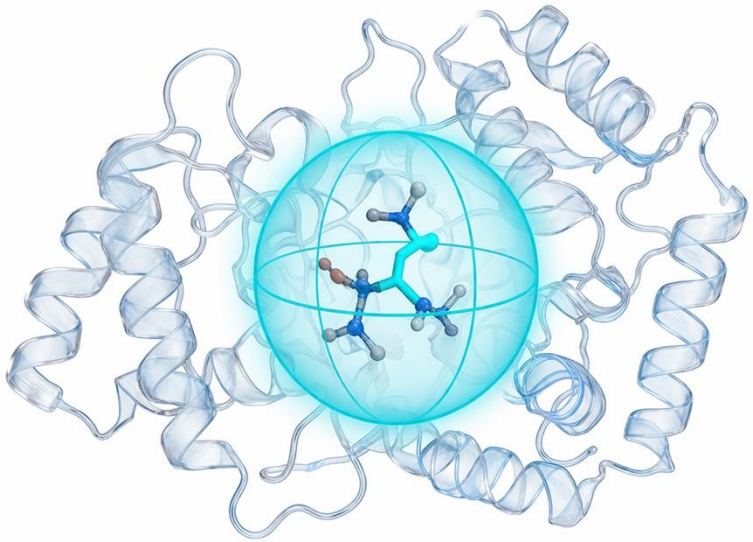
Definition 3.1 (Protein Motif)**.** *It is a recurring local structural–sequential pattern of residues that is evolutionarily conserved and often functionally significant. Formally, it can be described as a short stretch of amino acids together with its surrounding 3-D conformation, capturing local folding rules and biochemical properties independent of the global protein context.*



Key Trait: Evolutionarily conserved, functionally significant stretch of amino-acids with surrounding 3D conformation.

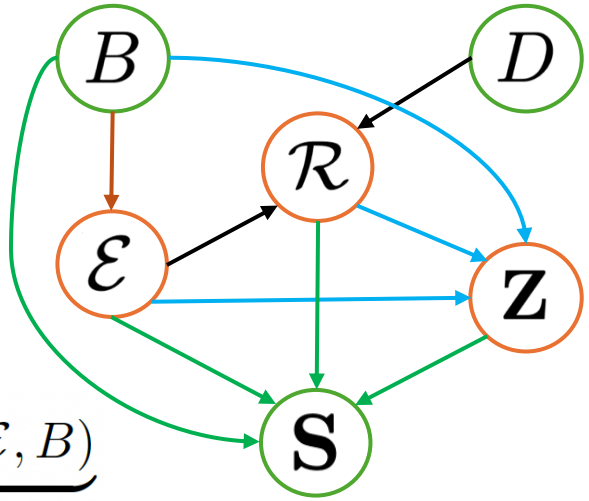
Definition 3.2: Potential Motif (PRISM's Generalization)

Definition 3.2 (Potential Motif)**.** *We generalize motifs by treating each residue together with its local 3-D neighborhood as a* potential motif*. A potential motif may or may not align with a canonical structural motif, but serves as a fine-grained* motif-like unit *that encodes transferable structure–sequence information. These representations are the building blocks for retrieval and sequence emission in our RAG framework.*



Key Trait: A fine-grained, transferable *motif-like* unit acting as the atomic building block for retrieval and sequence emission.

Formalizing as a Latent-Variable Probabilistic Model



Basic Generative Factorization,

$$p(\mathbf{S}, \mathcal{E}, \mathcal{R}, \mathbf{Z} \mid B, D) = \underbrace{p(\mathcal{E} \mid B)}_{\text{representation}} \underbrace{p(\mathcal{R} \mid \mathcal{E}, D)}_{\text{retrieval kernel}} \underbrace{p(\mathbf{Z} \mid \mathcal{R}, \mathcal{E}, B)}_{\text{attribution}} \underbrace{p(\mathbf{S} \mid \mathbf{Z}, \mathcal{R}, \mathcal{E}, B)}_{\text{sequence emission}}$$

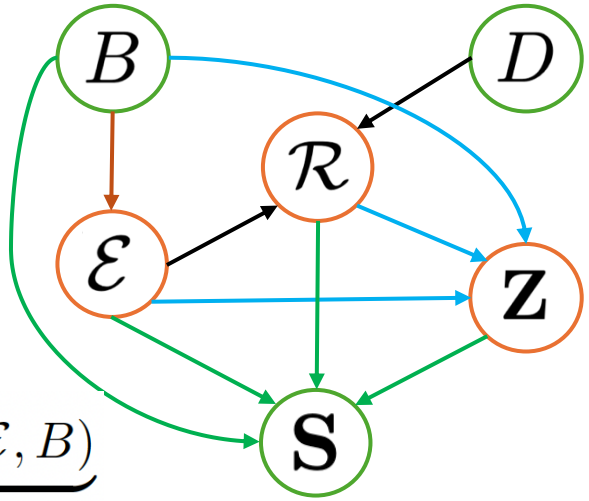
Marginalizing over the latents yields the conditional distribution we aim to model,

$$p(\mathbf{S} \mid B, D) = \mathbb{E}_{p(\mathcal{E}|B)\, p(\mathcal{R}|\mathcal{E},D)\, p(\mathbf{Z}|\mathcal{R},\mathcal{E},B)}\big[p(\mathbf{S} \mid \mathbf{Z}, \mathcal{R}, \mathcal{E}, B)\big]$$

$\mathbf{S}$: Emitted Amino-Acid Sequence	B : Target 3D Backbone	D : Prior-Knowledge Vector Database
$\mathcal{E}$: Potential-Motif Representation	$\mathcal{R}$: Latent Retrieval Hypothesis	$\mathbf{Z}$: Attribution Variables

Key Detail: This formulation induces a *family of valid objectives* arising under different approximations or parameterizations of the latents $\mathcal{E}$, $\mathcal{R}$, and $\mathbf{Z}$.

The PRISM Blueprint: Factorizing the Generative Process

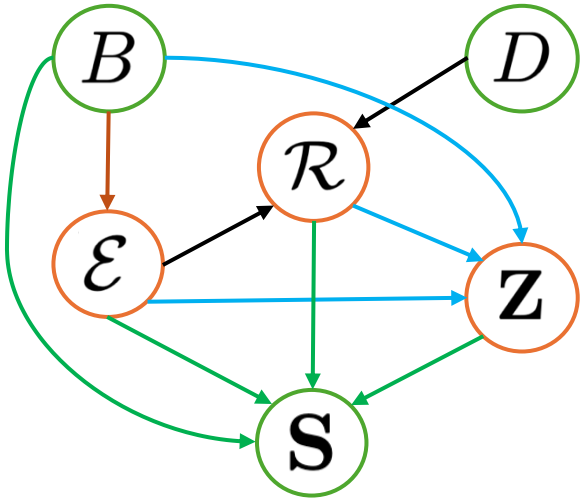


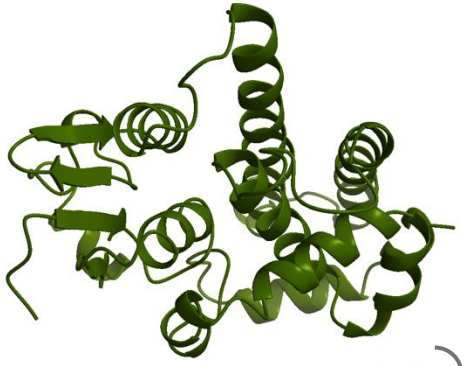
Basic Generative Factorization,

$$p(\mathbf{S}, \mathcal{E}, \mathcal{R}, \mathbf{Z} \mid B, D) = \underbrace{p(\mathcal{E} \mid B)}_{\text{representation}} \underbrace{p(\mathcal{R} \mid \mathcal{E}, D)}_{\text{retrieval kernel}} \underbrace{p(\mathbf{Z} \mid \mathcal{R}, \mathcal{E}, B)}_{\text{attribution}} \underbrace{p(\mathbf{S} \mid \mathbf{Z}, \mathcal{R}, \mathcal{E}, B)}_{\text{sequence emission}}$$

Probabilistic Term	Latent Meaning	Architectural Instantiation
$p(\mathcal{E} \mid B)$	Representation	Joint encoder $\mathcal{G}(B, \mathbf{S}) \in \mathbb{R}^{L \times d}$
$p(\mathcal{R} \mid \mathcal{E}, D)$	Retrieval Kernel	Stochastic retrieval from D and its deterministic approximation
$p(\mathbf{Z} \mid \mathcal{R}, \mathcal{E}, B)$	Attribution	Attention weights of the MHSCA decoder
$p(\mathbf{S} \mid \mathbf{Z}, \mathcal{R}, \mathcal{E}, B)$	Sequence Emission	Categorical distribution parameterized by the decoder logits.

Multimodal Representation of Potential-Motifs ($\mathcal{E}$)



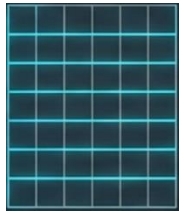


True Backbone B

…RLTRQQ…

True Sequence $\mathbf{S}$

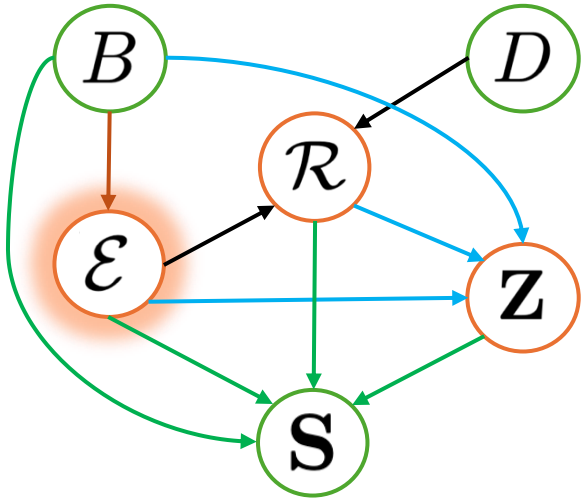
Joint Emb.



Potential-Motif Representation

$$\mathcal{E} = \mathcal{G}(P) = \mathcal{G}(B, \mathbf{S}) \in \mathbb{R}^{L \times d}$$

Multimodal Representation of Potential-Motifs ($\mathcal{E}$): What Happens During inference?



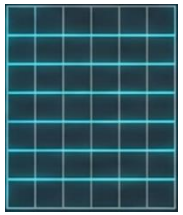
Query Backbone B^q

An off-the-shelf inverse folding model (*Base Estimator*), e.g., AIDO.Protein-IF, PiFold, etc.

…KIIEQT…

Initial Sequence Estimate $\hat{S}^q$

Joint Emb.

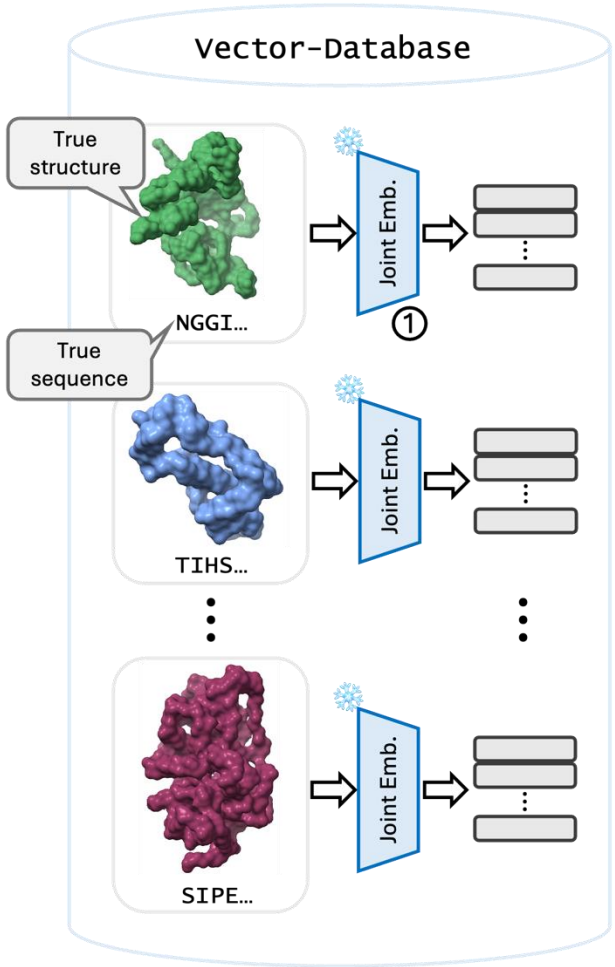


Potential-Motif Representation

$$\hat{\mathcal{E}}^q = \mathcal{G}(B^q, \hat{S}^q) \in \mathbb{R}^{L \times d}$$

Key Detail: Each vector $\hat{\mathcal{E}}_i^q \in \mathbb{R}^d$ contextualizes residue $i \in [L]$ by its local 3D neighborhood and its global placement, acting as a sample from the marginal, i.e., $\hat{\mathcal{E}}^q \sim p(\mathcal{E} \mid B = B^q)$.

Constructing the Vector Database as a Prior-Knowledge Memory (D)



Source

CATH-4.2 Training Split
(Strictly separated to prevent data leakage)

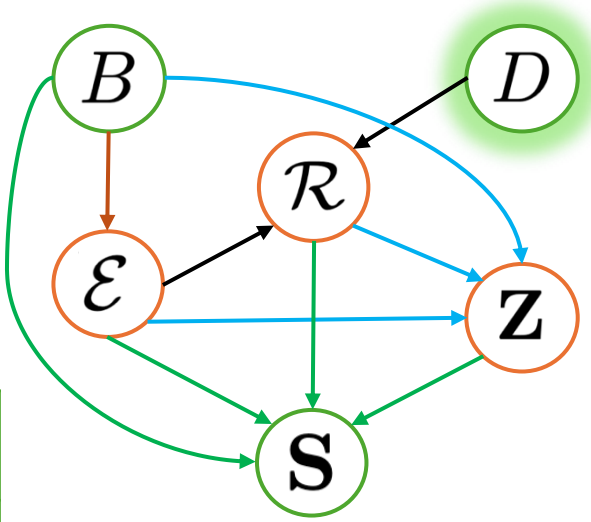
Scale

3,941,775 discrete potential motif embeddings

Format

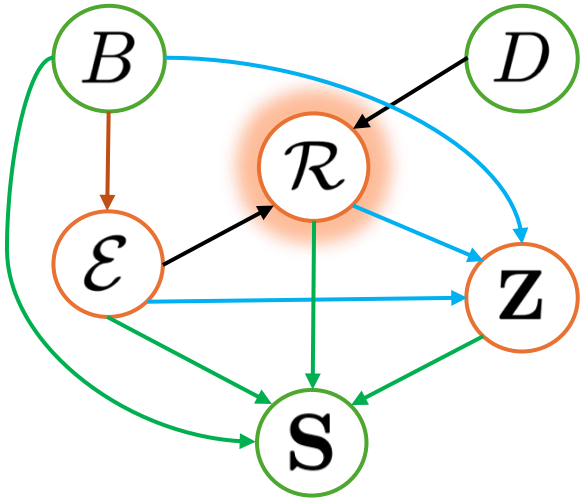
$$D = \{ d = (\mathcal{E}_r^p, r, p) \ : \ p \in [M], \ r \in [|P^p|] \}$$

Each residue embedding $\mathcal{E}_r^p$ summarizes the locality around residue r in protein p.



The entire index is stored and searched entirely on the GPU, guaranteeing minimal latency during inference

Latent Retrieval Hypothesis ($\mathcal{R}$)



$\mathcal{R} = \{\mathcal{R}_i\}_{i=1}^{L}$ denote a *latent retrieval hypothesis*, where $\mathcal{R}_i$ are neighbors of i retrieved from D. We define the retrieval kernel as $p(\mathcal{R} \mid \mathcal{E}, D)$.

The latent-variable formulation *enables training a retrieval kernel* by introducing an amortized posterior q and maximizing the ELBO,

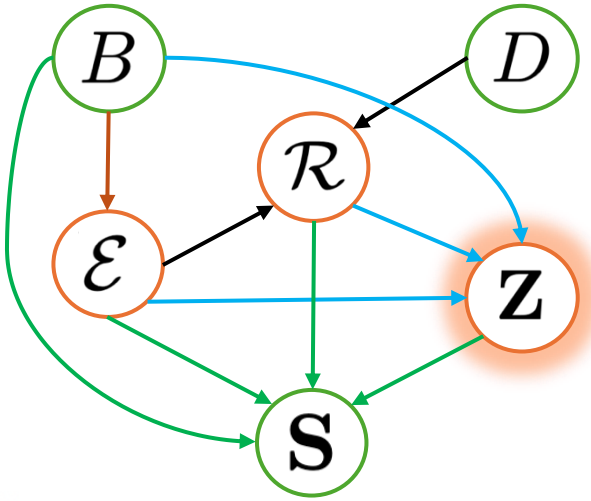
$$\mathcal{L}_{\mathrm{ELBO}}(q) = \mathbb{E}_q\big[\log p(\mathbf{S} \mid \mathbf{Z}, \mathcal{R}, \mathcal{E}, B)\big] - \mathbb{E}_q\big[\mathrm{KL}\big(q(\mathcal{R} \mid \mathbf{S}, \mathcal{E}, D) \,\|\, p(\mathcal{R} \mid \mathcal{E}, D)\big)\big]$$

Deterministic Approximation as an Efficient Alternative:

$$p(\mathcal{R}_i \mid \mathcal{E}_i, D) = \delta\big(\mathcal{R}_i - \mathrm{Top}K(\mathcal{E}_i; D)\big)$$

With $\mathrm{Top}K$ retrieval, the distribution collapses to a Dirac point mass, yielding a highly efficient $\mathrm{O}(1)$ approximation.

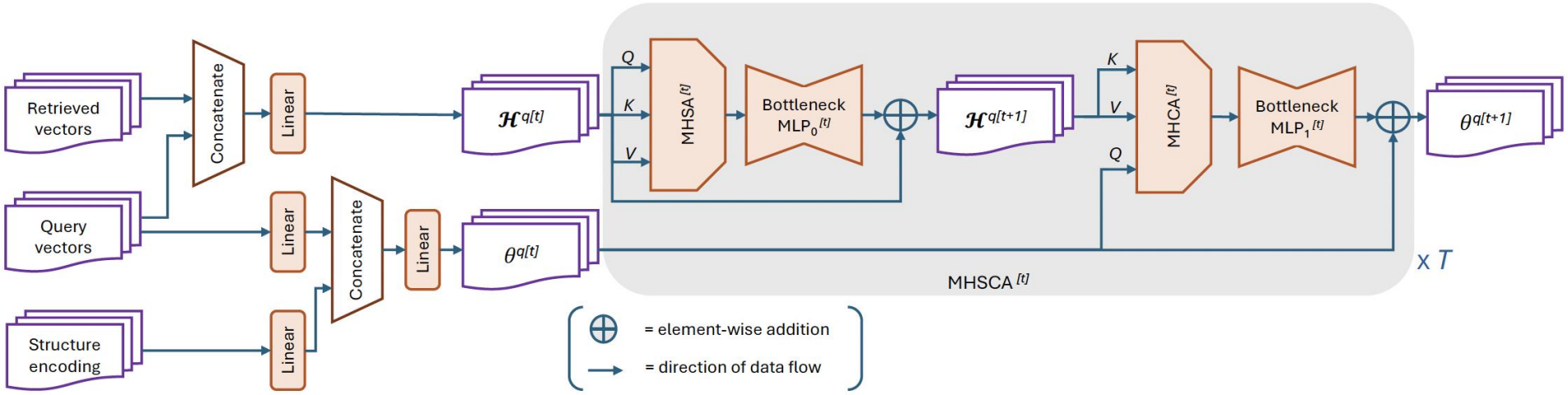
Attention weights of MHSCA as Attribution ($\mathbf{Z}$)



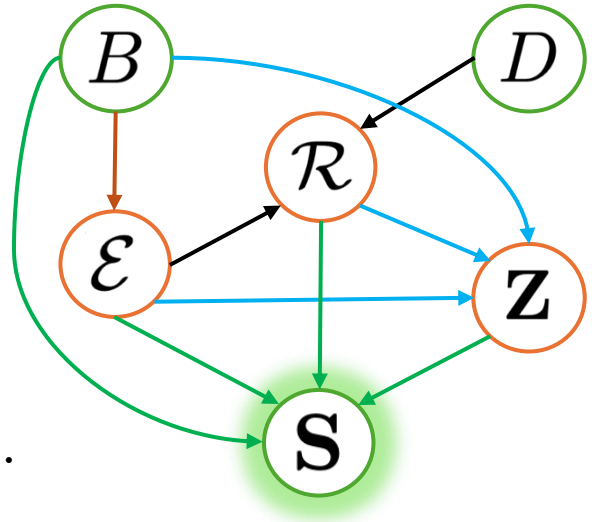
Attribution latent $\mathbf{Z}$ specifies how retrieved neighbors $\mathcal{R}$ contribute to emissions $\mathbf{S}$. We approximate the attribution $\mathbf{Z}$ as the attention weights of the hybrid MHSCA layers, i.e.,

$$\{\alpha_{ik}^{(t,h)}\}_{i,k,t,h} = \mathcal{A}(\mathcal{R}, \mathcal{E}, B), \qquad p(\mathbf{Z} \mid \mathcal{R}, \mathcal{E}, B) = \delta\big(\mathbf{Z} - \mathcal{A}(\mathcal{R}, \mathcal{E}, B)\big)$$

Hybrid Multi-Head Self-Cross Attention (MHSCA) in our decoder:



Sequence Emission ($\mathbf{S}$) and Training Objective



The emission distribution factorizes,

$$p(\mathbf{S} \mid \mathcal{E}, B, \mathcal{R}, \mathbf{Z}) = \prod_{i=1}^{L} \mathrm{Cat}\big(S_i;\, \mathrm{softmax}(\mathbf{Y}(\mathcal{E}, B, \mathcal{R}))_i\big) \,.$$

Here per-residue logits $\mathbf{Y}(\mathcal{E}, B, \mathcal{R}) = F_{\theta_{\mathbf{Z}}}\big(F_{\theta_B}(B),\ \mathcal{E},\ \mathcal{R}\big) \in \mathbb{R}^{L \times 20}$ are produced by the decoder $F_{\theta_{\mathbf{Z}}}$, where F_{θ_B} is a 3D structure encoder.

Training objective:

When the retrieval prior is trainable our training objective becomes,

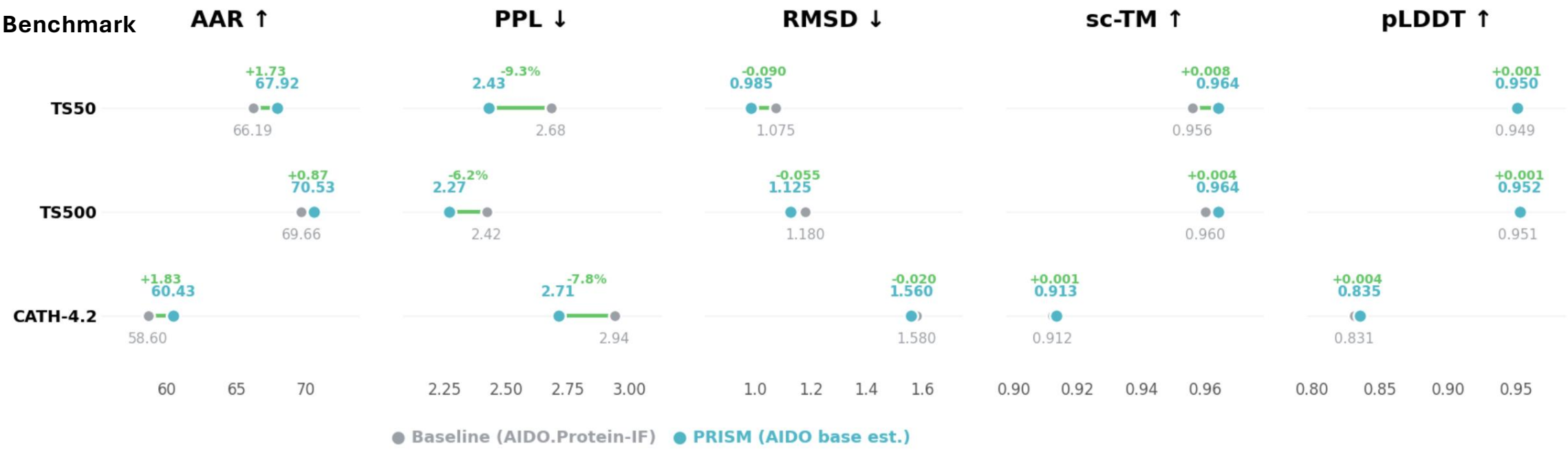
$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta}} \mathbb{E}_q\big[\log p_{\underline{\theta}}(\mathbf{S} \mid \cdot)\big] - \mathbb{E}_q\big[\mathrm{KL}\big(q(\mathcal{R} \mid \cdot) \,\|\, p_{\underline{\theta}}(\mathcal{R} \mid \cdot)\big)\big], \text{ with } \underline{\theta} = \{\theta_{\mathbf{Z}},\, \theta_B,\, \theta_{\mathcal{G}}\}$$

Under our deterministic retrieval, the objective collapses to maximizing the standard log-likelihood via per-residue cross-entropy,

$$\hat{\theta} = \arg\max_{\theta} \mathbb{E}_p\big[\log p_\theta(\mathbf{S} \mid \cdot)\big]$$

Experiments and Results: SoTA across both Sequence- and Structure-Level Evaluations

PRISM improves upon SoTA across multiple benchmarks on both sequence-level (AAR, PPL) and structure-level metrics (RMSD, sc-TM, pLDDT). For instance, comparing against a recent method AIDO.Protein-IF:



Experiments and Results: Robustness Against Severe Distributional Shift

Temporal Shift
(PDB Date split)

Strictly temporally disjoint from training data.

Comparing to the base estimator AIDO.Protein-IF PRISM boosts AAR from 66.27% to 67.47%, providing forward-looking generalization.

Real-World Targets
(CAMEO 2022)

Proteins outside standard CATH classifications.

Perplexity drops from 2.68 to 2.53 compared to the base estimator AIDO.Protein-IF; strong foldability maintained.

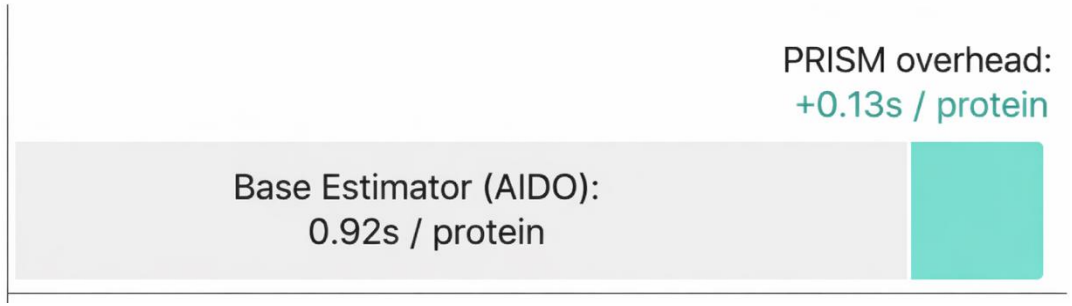
The Ultimate Test
(Orphan Proteins)

11 highly challenging proteins with *zero detectable homologs*.

PRISM achieves foldability score close to native sequences. Model reasons structurally, rather than reciting homologs.

Negligible Computational Overhead for High Recovery and Diversity

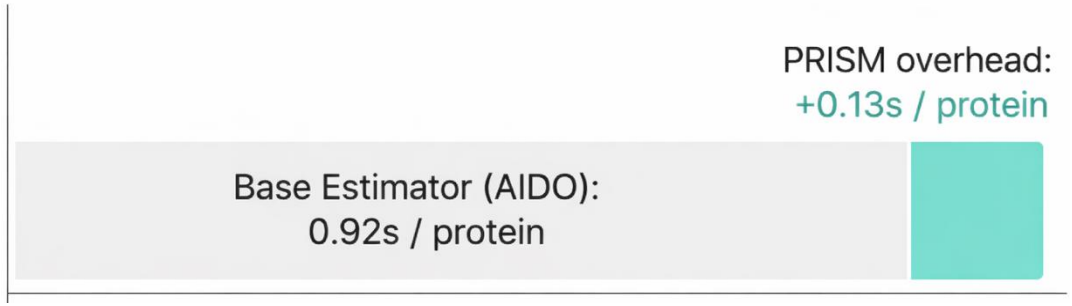
Cost vs. Benefit



A massive +3.9% absolute average AAR gain for only a ~14% runtime overhead (1.05 seconds total). Vector DB runs entirely on GPU.

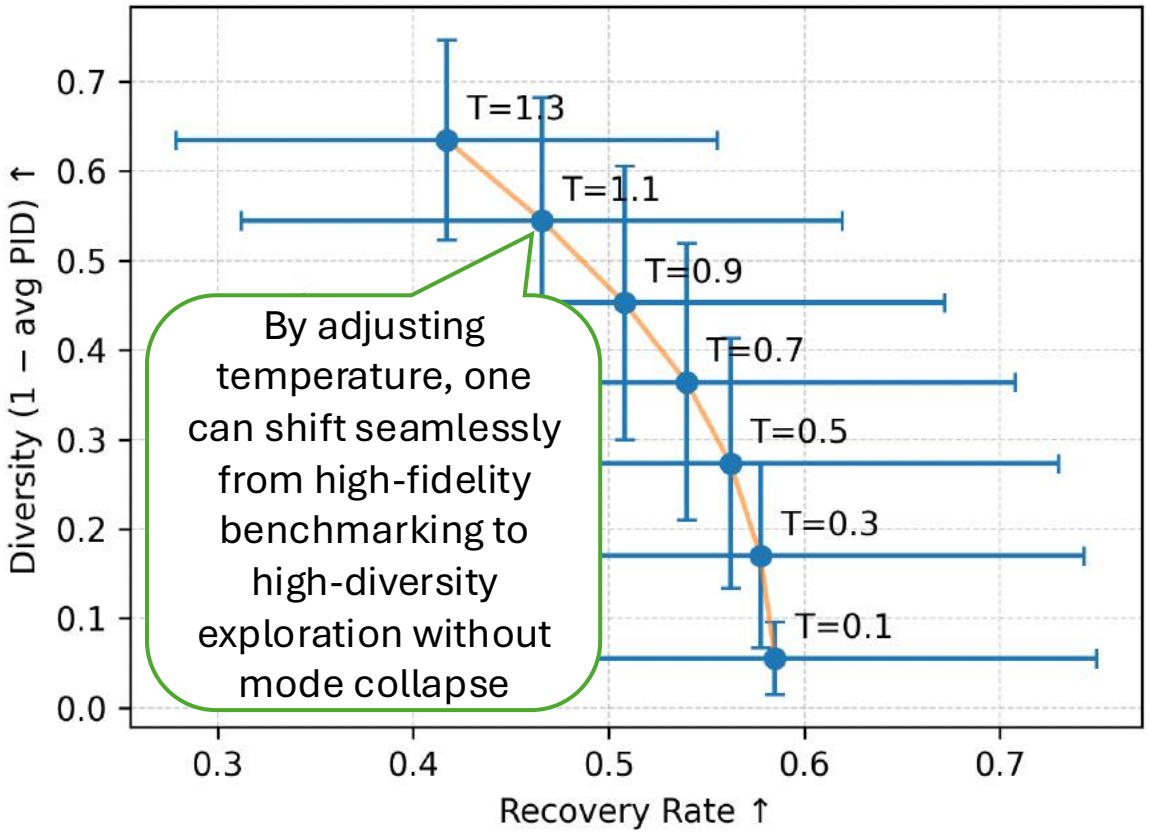
Negligible Computational Overhead for High Recovery and Diversity

Cost vs. Benefit



A massive +3.9% absolute average AAR gain for only a ~14% runtime overhead (1.05 seconds total). Vector DB runs entirely on GPU.

The Recovery-Diversity Frontier



Shifting Inverse Folding from Monolithic to Multimodal Retrieval-Augmented Generation

Residue-Level Granularity

First Multimodal RAG framework to explicitly reuse fine-grained evolutionary patterns

Theoretical Grounding

A latent-variable probabilistic model and an efficient approximation for implementation

Uncompromised Performance

New SoTA across **6 major benchmarks** with negligible computational overhead.

PRISM establish fine-grained retrieval as a principled and scalable approach for advancing protein sequence design.

Thank you!

Paper