Elucidating the Design Space of Multimodal Protein Language Models (DPLM-2.1)

ICML 2025 Spotlight

ByteDance Research

Presenter: Angxiao Yue

May 26, 2025

Outline

Background

Task: **MultiModal Protein Language Models (PLMs).**

Existing foundation models often treat sequence and structure as separate modalities (✗, ✓).

For PLMs family:

- **ESM2**. Science 2023. Citations: 3054. ✗
- **ESM3**. Science 2025. Citations: 274. ✓
- **DPLM**. ICML 2024. Citations: 42. ✗
- **DPLM-2**. ICLR 2025. Citations: 13. ✓
- **DPLM-2.1**. ICML 2025. ✓

For DPLM and DPLM-2:

- Tokenization loss.
- Inaccurate structure prediction.

Contributions

Build upon **DPLM-2** to advance the design space spanning **improved generative modeling**, **structure-aware architectures**, **representation learning**, and **data exploration**.

PS. In this paper, generative modeling = structure prediction = Folding.

Outline

Notation

- L: Number of protein residues.
- $\mathbf{s} \in \mathbb{R}^{L \times |S|}$: Amino acid sequence, $s_i \in \{0,1\}^{|S|}$, $S = \{1, 2, \ldots, 20\}$.
- $\mathbf{x} \in \mathbb{R}^{L \times N_{\text{atoms}} \times 3}$: Cartesian coordinates (structure), $x_i \in \mathbb{R}^{N_{\text{atoms}} \times 3}$
- $\mathbf{z} \in \mathbb{R}^{L \times D}$: Latent representation of structure.

DPLM

The family of DPLMs is grounded in the absorbing discrete diffusion framework.

- Protein sequence $\mathbf{s}$ is a categorical distribution $\mathrm{Cat}(\mathbf{s}; \mathbf{p})$, parameterized by a vector $\mathbf{p}$ on $(|\mathcal{V}|-1)$-dimensional probability simplex.
- **Forward Process** (Noise Perturbation): Perturbs data $\mathbf{s} \sim q(\mathbf{s})$ to a stationary noise distribution $\mathbf{s}^{(T)} \sim q_{\text{noise}}$.

$$q(\mathbf{s}^{(t)}|\mathbf{s}^{(t-1)}) = \mathrm{Cat}(\mathbf{s}^{(t)}; \beta_t \mathbf{s}^{(t-1)} + (1-\beta_t) q_{\text{noise}}) \tag{1}$$

- **Backward Process** (Learned Denoising): A learned process $p_\theta(\mathbf{s}^{(t-1)}|\mathbf{s}^{(t)})$ reverses the noise, aiming to recover the original data distribution $\mathbf{s}$.
- **Learning Objective**:

$$\begin{aligned} \mathcal{J}_t &= \mathbb{E}_{q(\mathbf{s})} - \mathrm{KL}[q(\mathbf{s}^{(t-1)}|\mathbf{s}^{(t)}, \mathbf{s})||p_\theta(\mathbf{s}^{(t-1)}|\mathbf{s}^{(t)})] \\ &= \mathbb{E}_{q(\mathbf{s})}[\lambda^{(t)} \Sigma_{1 \le i \le L} b_i(t) \cdot \log p_\theta(s_i|\mathbf{s}^{(t)})]. \end{aligned} \tag{2}$$

DPLM-2

Extends DPLM by introducing a token-based latent representation for protein structure.

- **Learning Objective** For structure token sequence $\mathbf{z}$ and amino acid sequence $\mathbf{s}$:

$$\mathcal{J}_t = \mathbb{E}_{q(\mathbf{x},\mathbf{z})}\left[\lambda^{(t)} \sum_i b_i(t)\left(\log p_\theta(s_i|\cdot) + \log p_\theta(z_i|\cdot)\right)\right] \tag{3}$$

- **Structure tokenization**

$$\mathbf{x} \xrightarrow{\text{encoder}} \mathbf{z}_{\text{cont}} \xrightarrow[\text{quantize}]{\text{dimension-wise}} \mathbf{z}_{\text{quant}} \Leftrightarrow \mathbf{z}_{\text{index}} \xrightarrow{\text{decoder}} \tilde{\mathbf{x}},$$

DPLM-2

- $\mathbf{x} \in \mathbb{R}^{L \times N_{\text{atom}} \times 3} \xrightarrow{\text{encoder}} \mathbf{z}_{\text{cont}} \in \mathbb{R}^{L \times D}$
- $\mathbf{z}_{\text{cont}} \xrightarrow[\text{quantize}]{\text{dimension-wise}} \mathbf{z}_{\text{quant}} \in \{-1, +1\}^{L \times D}$

  For $\mathbf{z}^{[k]}_{\text{cont},i}$, $k = 1, 2, \cdots, D$:
  - If $\mathbf{z}^{[k]}_{\text{cont},i} > 0$, $\mathbf{z}^{[k]}_{\text{quant},i} = +1$;
  - If $\mathbf{z}^{[k]}_{\text{cont},i} \leq 0$, $\mathbf{z}^{[k]}_{\text{quant},i} = -1$.
- $\mathbf{z}_{\text{quant}} \Leftrightarrow \mathbf{z}_{\text{index}} \in \mathbb{R}^{L} \xrightarrow{\text{decoder}} \tilde{\mathbf{x}} \in \mathbb{R}^{L \times N_{\text{atom}} \times 3}$

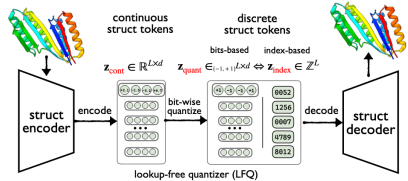
$$\mathbf{z}_{\text{index},i} = \sum_{k=1}^{D} \mathbf{1}(\mathbf{z}^{[k]}_{\text{quant},i} > 0) \cdot 2^{k-1} \in \{0, \dots, 2^D - 1\}^L$$

e.g., if $D = 3$, $\mathbf{z}_{\text{quant},i} = [+1, -1, +1]$:

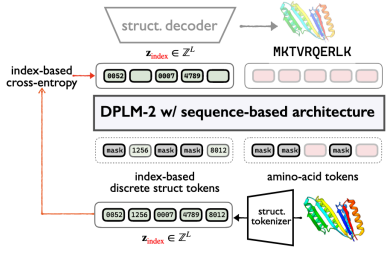
- $k = 1 : \mathbf{1}(+1 > 0) \cdot 2^{1-1} = 1 \cdot 1 = 1$
- $k = 2 : \mathbf{1}(-1 > 0) \cdot 2^{2-1} = 0 \cdot 2 = 0$
- $k = 3 : \mathbf{1}(+1 > 0) \cdot 2^{3-1} = 1 \cdot 4 = 4$
- $\mathbf{z}_{\text{index},i} = 1 + 0 + 4 = 5.$

(A1) structure tokenizer

tokenization converts structure into (latent) struct tokens



(A2) sequence-based language modeling w/ tokenized structure



Outline

Observation 1

(O1): Structure tokenization results in information loss.

Table 1. **Effects of feature quantization on structure tokenizer reconstruction.**

Latent feature	Struct token type	Reconstruction	
		RMSD↓	TMscore↑
$\mathbf{z}_{\text{cont}}$	(pre-quantized) continuous token	**1.3127**	**0.9733**
$\mathbf{z}_{\text{index}} \Leftrightarrow \mathbf{z}_{\text{quant}}$	(quantized) discrete token	1.9806	0.9385

This suggests that learning to **recover the lost residuals**—particularly as a refinement step—could enhance structure prediction accuracy.

Observation 2

(O2): High reconstruction accuracy does not guarantee better structure generative performance in language models, while a significant gap remains in between.

Table 2. **Tokenizer reconstruction *vs*. language model generation.** Evaluation of folding on CAMEO 2022.

Tokenizer	Reconstruction		Generation	
	rRMSD↓	rTMscore↑	RMSD↓	TMscore↑
DPLM-2	1.9806	0.9385	**7.7025**	**0.7936**
ESM3	**0.7248**	**0.9912**	8.4424	0.7924

This suggests that, given that mild improvement in reconstruction do not necessarily translate into better generation, greater emphasis should be placed on improving **structure-aware** generative modeling and architectural design.

Observation 3

(O3): Index-based structure tokens? Multimodal PLM gets them miserably wrong in structure prediction.

Table 3. **Language model structure token prediction accuracy.** Index-based *vs*. bits-based evaluation on structure folding.

Model	Testset	Struct Token Acc↑		Struct Eval Metric	
		index	bit	RMSD↓	TMscore↑
DPLM-2 index-based	CAMEO 2022	0.0864	0.7720	7.7025	0.7936
	PDB date split	0.1188	0.7932	5.3071	0.8306
DPLM-2 BIT-based	CAMEO 2022	**0.1258**	**0.7958**	**6.4028**	**0.8380**
	PDB date split	**0.2641**	**0.8648**	**3.2213**	**0.9043**

This suggests that while the model struggles to recover exact indices, it effectively captures structural patterns at the bit level.

Outline

Overview

Recall the Contributions:

Build upon **DPLM-2** to advance the design space spanning **improved generative modeling (O1&O3)**, **structure-aware architectures(O2)**, **representation learning (O2)**, and **data exploration(Φ)**.

Improved Generative Modeling(O1&O3)

1. Recovering Tokenization Loss by Learning Quantization Residuals with RESDIFF (O1).

They introduce a light-weight diffusion module, i.e., RESDIFF, to predict the residual information: $\mathbf{r} = \mathbf{z}_{\text{cont}} - \mathbf{z}_{\text{quant}}$. The training loss is to minimize:

$$\mathcal{L}_\phi = \mathbb{E}_{q(\mathbf{r}),\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\left\|\epsilon - \epsilon_\phi(\mathbf{r}_t, t, \mathbf{h}, \mathbf{z}_{\text{quant}})\right\|_2^2\right]. \tag{4}$$

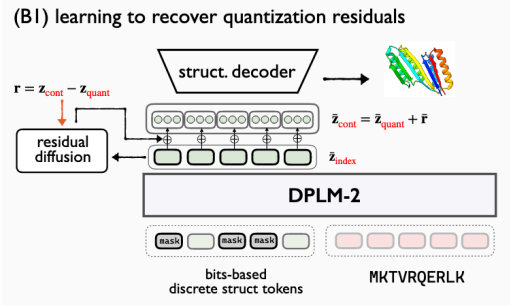


Table 4. **Evaluation of improved approaches for structure prediction based upon DPLM-2.** Folding SFT: supervised fine-tuning with folding objective.

Models	CAMEO 2022		PDB date split	
	RMSD↓	TMscore↑	RMSD↓	TMscore↑
ESMFold (3B) (Lin et al., 2022)	3.9900	0.8500	2.8400	0.9300
MultiFlow (Campbell et al., 2024a)	17.8400	0.5000	15.6400	0.5300
ESM3 (1.4B) (Hayes et al., 2024)	6.3300	0.8400	4.9003	0.8653
DPLM-2 (650M)	7.7025	0.7936	5.3071	0.8306
DPLM-2 + RESDIFF	7.2881	0.8087	5.1072	0.8430
DPLM-2 (BIT-based)	6.4028	0.8380	3.2213	0.9043
DPLM-2 (BIT-based) + RESDIFF	6.1781	0.8428	3.0168	0.9076
DPLM-2 (BIT-based) + FM	6.1825	0.8414	2.8697	0.9099
DPLM-2 (BIT-based) + FM + RESDIFF	6.0765	0.8456	2.7884	0.9146
w/ folding SFT	5.8472	0.8442	2.3698	0.9270
DPLM-2 (3B) *w/* folding SFT	5.9832	0.8443	3.1502	0.9012

Improved Generative Modeling(O1&O3)

2. Bridging Discrete and Continuous Structure Tokens with BIT-based Language Modeling (O3).

To bridge this gap, they aim to perform language modeling of the bit-based feature of structure tokens instead of their indices:

$$\mathcal{J}_t^{\text{bit}} = \mathbb{E}_{q(\mathbf{x},\mathbf{z})}\left[\lambda^{(t)} \sum_i b_i(t) \left(\log p_\theta(s_i|\cdot) + \sum_k \log p_\theta(z_{i,\text{quant}}^{[k]}|\cdot) \right)\right] \tag{5}$$

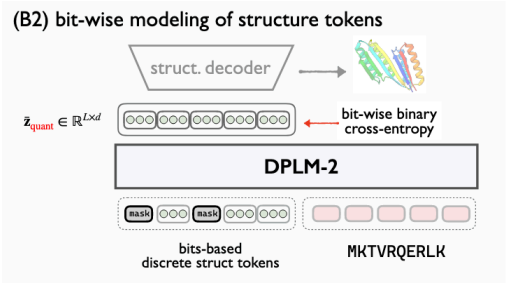


Table 4. **Evaluation of improved approaches for structure prediction based upon DPLM-2.** Folding SFT: supervised fine-tuning with folding objective.

Models	CAMEO 2022		PDB date split	
	RMSD↓	TMscore↑	RMSD↓	TMscore↑
ESMFold (3B) (Lin et al., 2022)	3.9900	0.8500	2.8400	0.9300
MultiFlow (Campbell et al., 2024a)	17.8400	0.5000	15.6400	0.5300
ESM3 (1.4B) (Hayes et al., 2024)	6.3300	0.8400	4.9003	0.8653
DPLM-2 (650M)	7.7025	0.7936	5.3071	0.8306
DPLM-2 + ResDiff	7.2881	0.8087	5.1072	0.8430
DPLM-2 (Bit-based)	6.4028	0.8380	3.2213	0.9043
DPLM-2 (Bit-based) + ResDiff	6.1781	0.8428	3.0168	0.9076
DPLM-2 (Bit-based) + FM	6.1825	0.8414	2.8697	0.9099
DPLM-2 (Bit-based) + FM + ResDiff	6.0765	0.8456	2.7884	0.9146
w/ folding SFT	5.8472	0.8442	2.3698	0.9270
DPLM-2 (3B) *w/* folding SFT	5.9832	0.8443	3.1502	0.9012

Improved Generative Modeling(O1&O3)

3. Hybrid Generative Approach Enables Direct Data-space Modeling (Φ).

The combination of the structure encoder, language model, and decoder as a whole effectively functions as a denoising model, capable of refining structure in atomic coordinates.

$$\mathbf{x}_\theta(\mathbf{x}_t, t) : \mathbf{x}_t \mapsto \bar{\mathbf{x}}_{\text{denoised}} \triangleq \text{decoder} \circ \text{PLM} \circ \text{encoder}(\mathbf{x}_t).$$

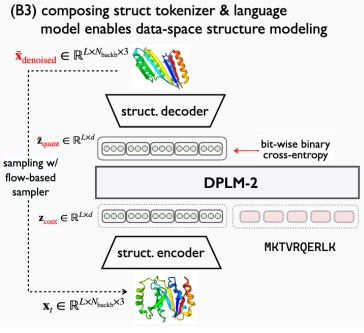
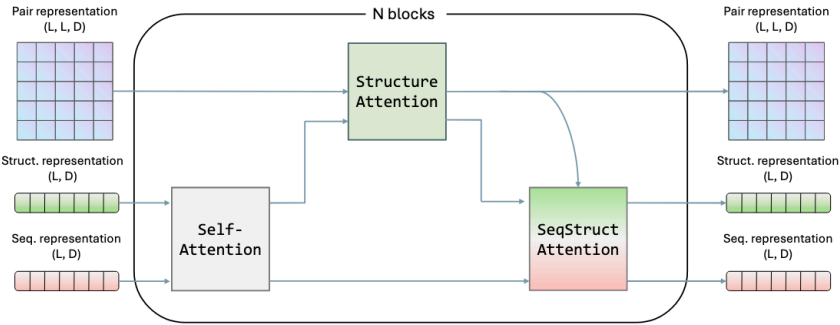


Table 4. **Evaluation of improved approaches for structure prediction based upon DPLM-2.** Folding SFT: supervised fine-tuning with folding objective.

Models	CAMEO 2022		PDB date split	
	RMSD↓	TMscore↑	RMSD↓	TMscore↑
ESMFold (3B) (Lin et al., 2022)	3.9900	0.8500	2.8400	0.9300
MultiFlow (Campbell et al., 2024a)	17.8400	0.5000	15.6400	0.5300
ESM3 (1.4B) (Hayes et al., 2024)	6.3300	0.8400	4.9003	0.8653
DPLM-2 (650M)	7.7025	0.7936	5.3071	0.8306
DPLM-2 + ResDiff	7.2881	0.8087	5.1072	0.8430
DPLM-2 (Bit-based)	6.4028	0.8380	3.2213	0.9043
DPLM-2 (Bit-based) + ResDiff	6.1781	0.8428	3.0168	0.9076
DPLM-2 (Bit-based) + FM	6.1825	0.8414	2.8697	0.9099
DPLM-2 (Bit-based) + FM + ResDiff	6.0765	0.8456	2.7884	0.9146
w/ folding SFT	5.8472	0.8442	2.3698	0.9270
DPLM-2 (3B) *w/* folding SFT	5.9832	0.8443	3.1502	0.9012

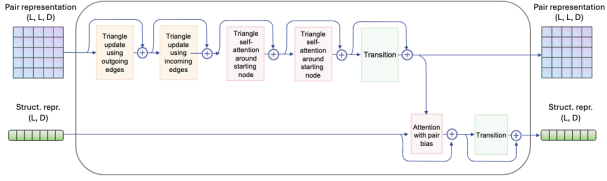
Structure-aware Architectures(O2)

1. GeoDPLM: Geometry-aware Model Architecture (O2).
Operates on compact 2D pair representations to capture pairwise spatial dependencies of residues: **Structure attention module** + **Seqstruct attention module**.

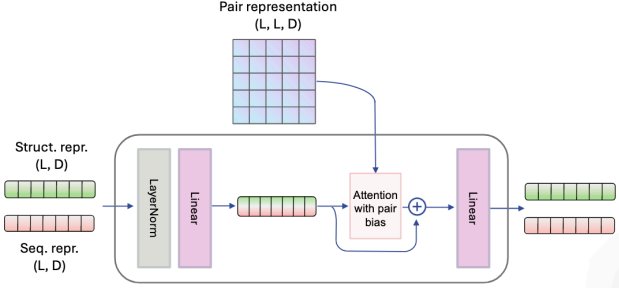


Structure-aware Architectures(O2)

Structure attention module.



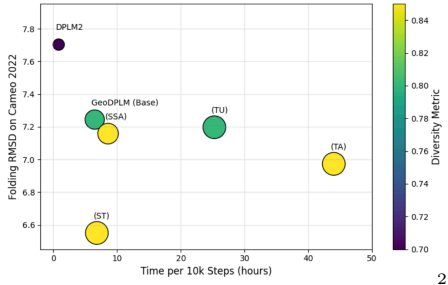
Seqstruct attention module.



Structure-aware Architectures(O2)

Methods	Structure Attention				Seqstruct Attn.	SFT	PDB date split		CAMEO 2022	
	P. Bias & Tran.	S. Tran.	Tri. Up.	Tri. Attn.			RMSD↓	TMscore↑	RMSD↓	TMscore↑
DPLM-2	×	×	×	×	×	×	5.521	0.8287	7.703	0.7936
GeoDPLM (Base)	√						4.823	0.8521	7.244	0.8128
(ST)	√	√					**3.883**	**0.8857**	**6.550**	**0.8339**
(TU)	√	√	√				4.837	0.8598	7.197	0.8255
(TA)	√	√		√			4.415	0.8690	6.973	0.8210
(SSA)	√				√		4.040	0.8841	7.158	0.829
DPLM-2	×	×	×	×	×	√	3.347	0.9008	6.612	0.8233
GeoDPLM (Base)	√					√	3.165	0.9046	**6.227**	**0.8414**
(ST)	√	√				√	**3.021**	**0.9062**	6.288	0.8393
(TU)	√	√	√			√	3.639	0.8903	6.877	0.8322
(TA)	√	√		√		√	3.863	0.8790	6.393	0.8340
(SSA)	√				√	√	3.134	0.9054	6.329	0.8379

- **P.Bias & Tran**: pair bias and transition layer for pair representation.
- **S. Tran**: transition layer for structure representation.
- **Tri. Up. & Attn.**: triangle update and attention layers.

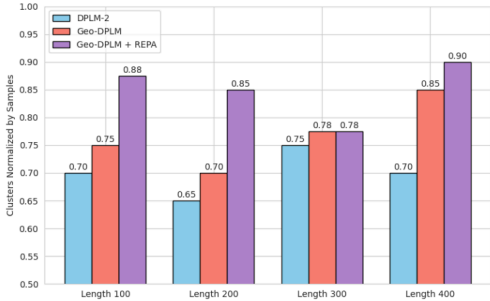


Representation Learning (O2)

1. Representation Alignments to Folding Model (O2).
The primary challenge of training diffusion models is learning high-quality representations. To address this, they adopt **REPA** by aligning the representations of the protein language model to transfer meaningful structural semantics from specialized folding model.

Table 6. **Representation alignment improves structure prediction.** REPA is compatible with both language model-based architectures (DPLM-2) and geometric designs (GeoDPLM).

Methods	PDB date split		CAMEO 2022	
	RMSD↓	TMscore↑	RMSD↓	TMscore↑
DPLM-2	5.521	0.8287	7.703	0.7936
w REPA	4.919	0.8508	7.344	0.8046
GeoDPLM	4.823	0.8521	7.244	0.8128
w REPA	**4.340**	**0.8671**	**7.058**	**0.8217**



On the Orthogonality of Design Methods

Examine the interactions of these designs by combining them in a unified setting.

The recommended setting: **Geo + Bit-based modeling**.

Table 7. **Analysis of orthogonality**. We analyze the compatibility of design methods when combined, with the recommended setting highlighted. *All** denotes the combination of all methods: Geo, Bit, FM, REPA, ResiDiff, and SFT.

Models	PDB date split		CAMEO 2022		Uncond. Gen.
	RMSD↓	TMscore↑	RMSD↓	TMscore↑	Diversity↑
DPLM-2 (650M)	5.307	.8306	7.703	.7936	0.700
Bit	3.221	.9043	6.403	.8380	0.825
Bit + FM	2.870	.9099	6.183	.8418	0.525
Bit + FM + ResDiff	2.788	.9146	6.077	.8456	0.525
w/ SFT	2.370	.9270	5.847	.8442	-
Geo + Bit	2.551	.9254	5.955	.8520	0.900
Geo + Bit + FM	2.443	.9261	6.172	.8404	0.575
Geo + Bit + REPA	2.507	.9264	6.192	.8412	0.875
w/ SFT	2.404	.9322	5.754	.8424	-
*All**	2.379	.9297	6.200	.8398	-

Data Exploration(Φ)

Table 8. **Statistics of PDB-Multimer.** We curate a dataset of multi-chain proteins from PDB to analyze their effects on structural modeling.

Dataset	# proteins Train / Val	# chains	Protein Length	Chain Length
PDB-Multimer	11614/291	2.88 ± 1.66	661.57 ± 416.37	229.39 ± 167.00

Table 9. **Effects of monomer data on tokenizer reconstruction for multimer.** Scaling *monomer* data significantly improves the structure tokenizer reconstruction on PDB-Multimer, suggesting the relevance between multimer and monomer modeling. We also provide results of monomer on CAMEO dataset.

Training Data	Size	PDB-Multimer		CAMEO 2022	
		RMSD↓	TMscore↑	RMSD↓	TMscore↑
PDB & SwissProt	200K	9.973	0.694	2.589	0.930
AFDB_Rep	+1.2M	**6.873**	**0.784**	**2.245**	**0.938**

Table 10. **Applying chain linker and position offset in multimer modeling.** We present the folding results on PDB-Multimer and report the reconstruction performance of structure tokenizer. ESMFold used G-linker of length 25 by default in multimer folding.

Method	PDB-Multimer	
	RMSD↓	TMscore↑
Tokenizer Reconstruction		
DPLM-2 (monomer) tokenizer	6.873	0.784
w/ Pos. Offset	5.886	0.812
Folding		
ESMFold	17.297	0.850
DPLM-2	19.110	0.768
w/ Chain Linker	**17.966**	0.771
w/ Pos. Offset	18.338	0.767

Table 11. **Fine-tuning with multimer and monomer data.** We evaluate the effects of fine-tuning with PDB-Multimer and Swissprot on structure prediction. Incorporating multimer data improves both monomer and multimer folding. SFT: supervised fine-tuning with folding objective.

Training Data		SFT	PDB-Multimer		CAMEO 2022	
PDB -Multimer	Swissprot		RMSD↓	TMscore↑	RMSD↓	TMscore↑
	✓		17.966	0.771	7.703	0.793
	✓	✓	19.615	**0.799**	6.612	0.823
✓		✓	**16.146**	0.775	10.989	0.686
✓	✓	✓	16.674	0.798	**6.410**	**0.831**

Outline

Conclusion

They identify the limitations in structural modeling for multimodal protein language models and propose an effective design space to bridge the gap.

- They demonstrate that tokenization quantization loss can be effectively mitigated with bit-label supervision and flow-matching.
- They introduce geometric inductive biases through architectural design and leverage representation learning to refine generation diversity.
- Building on the strengths of each component, we further investigate their orthogonality, which informs the final recommended setting.
- Lastly, to tackle the scarcity of structure data, we explore the data coverage to include multimers, ensuring broader 3D structural understanding.

Thank you!