NEURAL INFORMATION PROCESSING SYSTEMS

清華大學 Tsinghua University

BioMap

MSAGPT: Neural Prompting Protein Structure Prediction via MSA Generative Pre-Training

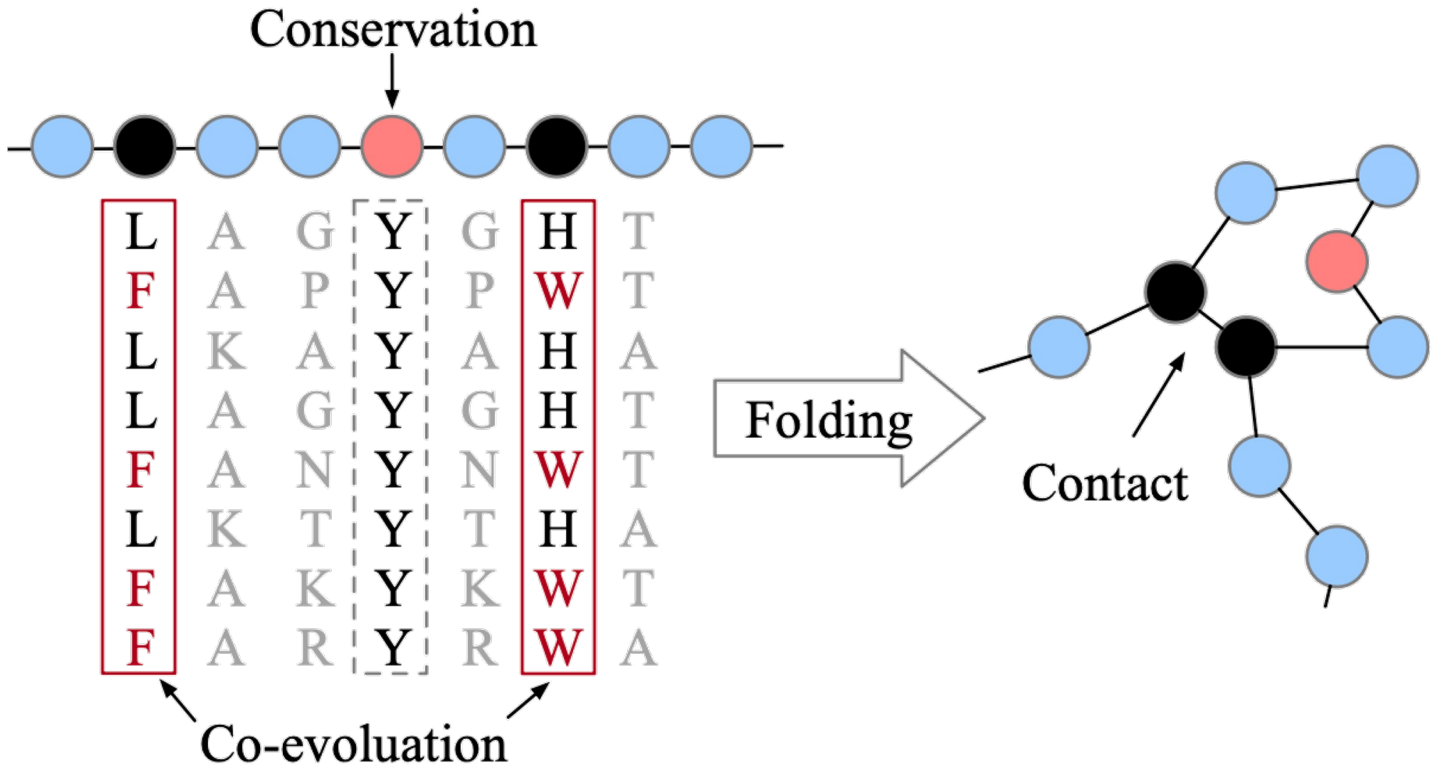
Bo Chen*, Zhilei Bei*, Xingyi Cheng, Pan Li, Jie Tang, Le Song

Tsinghua University, BioMap Research, MBZUAI

https://github.com/THUDM/MSAGPT

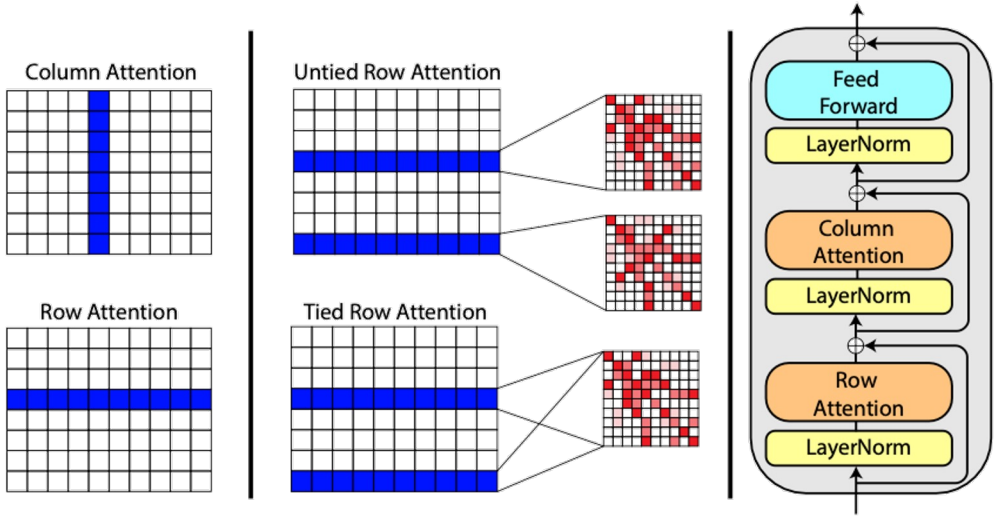
Multiple sequence alignment (MSA) facilitates protein structure prediction (PSP)

- Current PSP models **rely on MSA** for high accuracy
  - AlphaFold
  - RoseTTAFold
- "**Orphan**": 1/5 of all metagenomic proteins & 11% of eukaryotic proteins **lack sequence homologs**, compromising PSP accuracy

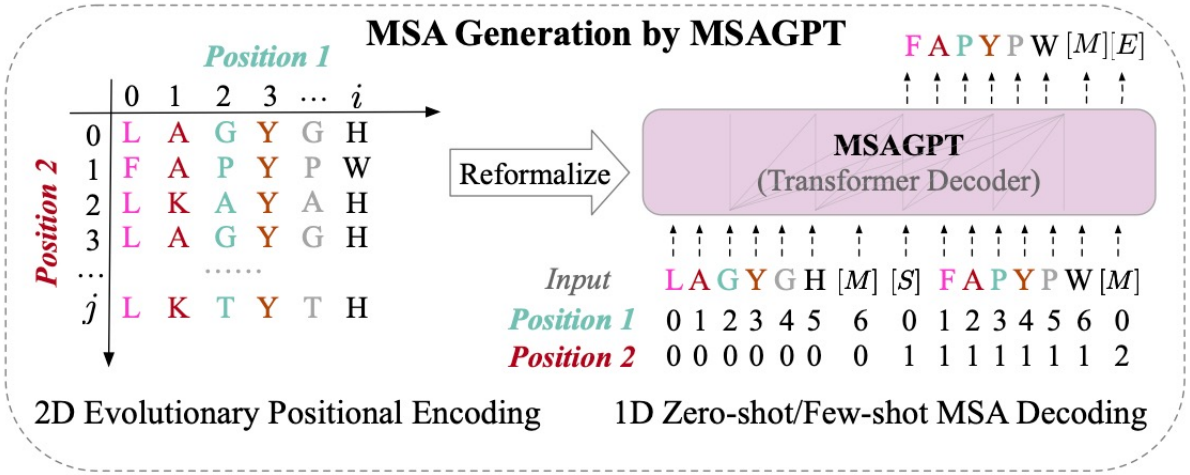


A simple yet effective decoding framework

- Previous MSA-based PLMs usually adopt **Axial Attention**
  - Constrained information fusion: Only allow row- or column-wise
  - Low Efficiency： Sequential attention in a transformer block
- We propose the **2D Evolutionary Position Encoding**
  - To relax the co-evolutionary information modeling from constrained attention flows to the 2D positional encoding
  - Re-formalizes MSA generation as a 1D sequence generation task, enables MSAGPT to conduct zero- or few-shot MSA generation under a flexible in-context learning framework



MSA Transformer (Axial Attention)



MSAGPT (Full Attention with 2D Evolutionary Positional Encoding)

NEURAL INFORMATION PROCESSING SYSTEMS

Tsinghua University

BioMap

Generate virtual MSA to solve the problem

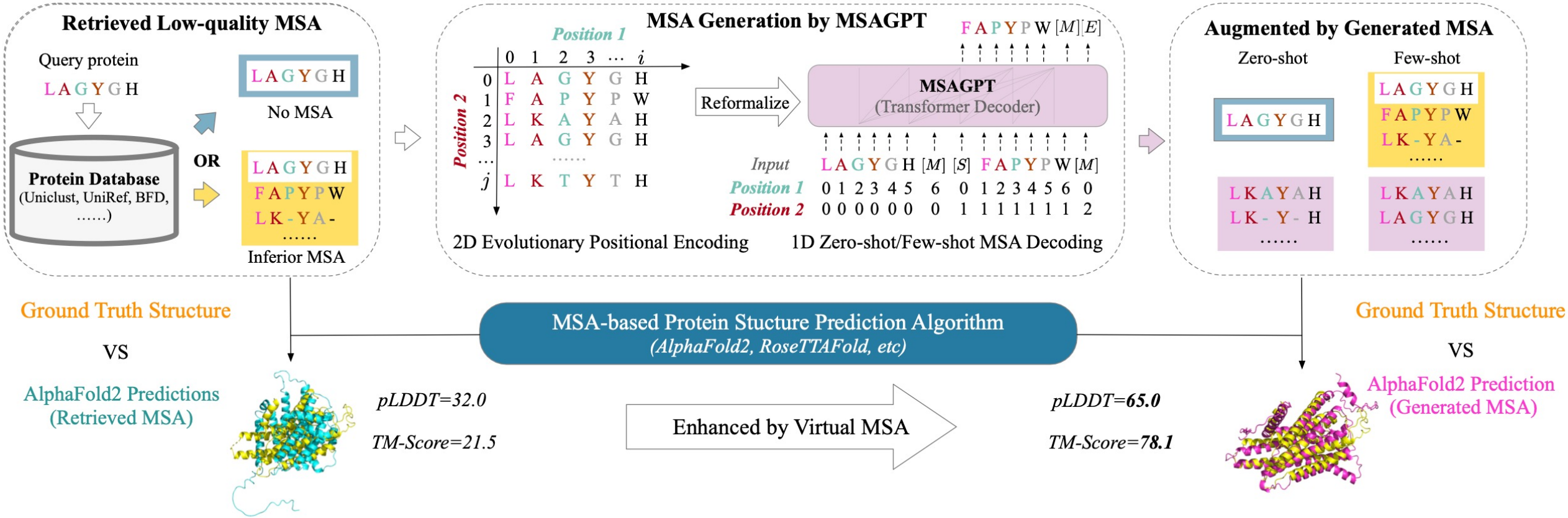
Low quality retrieved MSA
Low structural prediction accuracy

→

Enhanced by MSAGPT-generated virtual MSA

→

High quality augmented MSA
High structural prediction accuracy



Learning from AlphaFold2 feedback

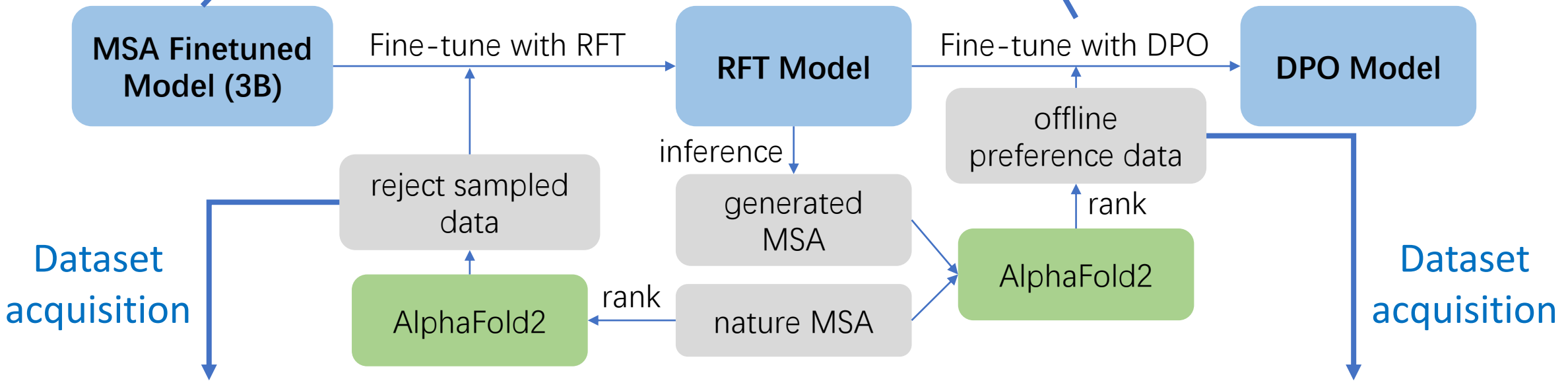
Post-training to alleviate the hallucination scene of MSA generation

- RFT stage: First fine-tune the model using high-quality natural MSA
- DPO stage: Then Use AlphaFold2 as a reward model and further fine-tune based on its feedback

$$L_{\text{ce}} = \mathbb{E}_{\mathbf{M}^f}\left[\sum_{i=0}^{N\times L} -\log p(\mathbf{M}_i^f|\mathbf{M}_{<i}^f,\theta)\right] \quad L_{\text{DPO}} = \mathbb{E}_{(Q,m_{\text{w}},m_{\text{l}})\in\mathcal{D}_{\text{DPO}}}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(m_{\text{w}}|Q)}{\pi_{\text{ref}}(m_{\text{w}}|Q)} - \beta\log\frac{\pi_\theta(m_{\text{l}}|Q)}{\pi_{\text{ref}}(m_{\text{l}}|Q)}\right)\right]$$

Loss calculation

Loss calculation



$$\mathcal{D}_{\text{RFT}} = \{(Q,m_i)|(\mathbb{I}_{acc}(Q,m_i)) > \theta_1 \cap (\mathbb{I}_{acc}(Q,m_i) - \mathbb{I}_{acc}(Q,-)) > \theta_2\} \quad \mathcal{D}_{\text{DPO}} = \{(Q,m_{\text{w}},m_{\text{l}})|\,(\mathbb{I}_{acc}(Q,m_{\text{w}}) - \mathbb{I}_{acc}(Q,m_{\text{l}})) > \theta_3\}$$

NEURAL INFORMATION PROCESSING SYSTEMS

Tsinghua University

BioMap

Experimental Result

- Protein Structure Prediction in low-MSA cases
  - **Natural** MSA-scarce benchmark: low retrieved MSA (<20) from uniclust30
    - **Zero-shot**: only use the generated MSA
    - **Few-shot**: Retrieved low-quality MSA + generated MSA

	Model	CAMEO (avg. Depth = 8.5)				CASP (avg. Depth = 4.6)				PDB (avg. Depth = 2.6)			
		Zero-Shot		Few-Shot		Zero-Shot		Few-Shot		Zero-Shot		Few-Shot	
		pLDDT	TM	pLDDT	TM	pLDDT	TM	pLDDT	TM	pLDDT	TM	pLDDT	TM
	AF2 MSA	63.8	55.4	77.4	71.4	44.0	32.6	54.2	44.1	55.2	45.6	61.0	52.3
	MSA-Aug.	67.7	59.2	77.4	72.1	56.8	36.6	63.4	46.3	61.9	49.8	66.0	55.3
	EvoGen	66.1	60.3	78.6	75.3	48.2	38.4	55.1	48.5	57.6	49.5	62.8	55.4
w/ virtual MSA	MSAGPT	**70.8**	61.4	**80.8**	75.2	**59.0**	39.8	**65.4**	51.0	**68.6**	53.4	**71.3**	59.6
	+ **RFT**	68.0	60.5	79.8	76.4	56.8	40.2	64.0	53.6	66.8	53.4	70.3	**60.1**
	+ **DPO**	68.9	**62.7**	80.2	**76.7**	54.2	**43.7**	62.7	**57.0**	64.5	**53.6**	68.0	59.7
		(+3.1)	(+2.4)	(+2.2)	(+1.4)	(+2.2)	(+5.3)	(+2.0)	(+8.5)	(+6.7)	(+3.8)	(+5.3)	(+4.7)

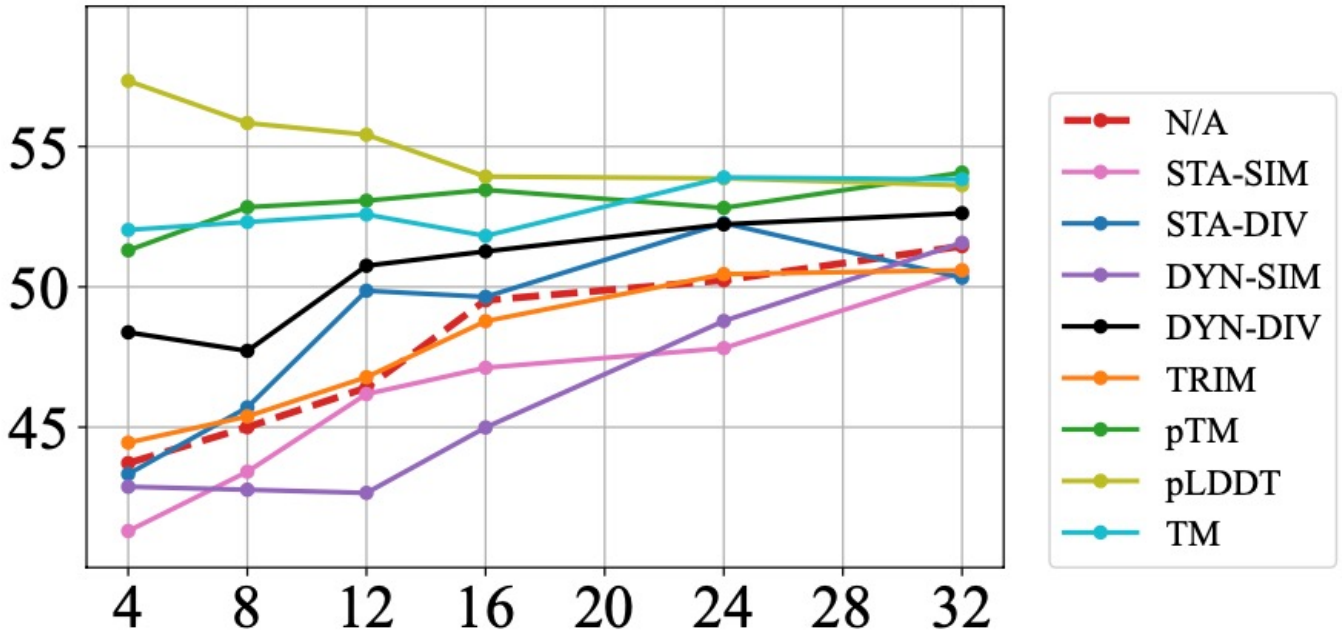
The post-training process significantly reduced hallucinations (Low Predictive metric) generated by the model and improved its performance (High Golden Metric).

Rethinking the MSA Selection Strategy

MSA Selection Criteria

- **1D Sequence** Diversity Measure
- **3D Structure** Validity Measure

Model	CAMEO	CASP	PDB
	TM	TM	TM
MSAGPT-DPO	76.7	57.0	59.7
+ **pLDDT Selection**	**77.5**	**57.6**	**60.5**



Sequence Diversity + Structure Validity → Informative MSA

Transfer Learning on Other Tasks

- Fine-tune MSA Transformer & task-specific head w/ or w/o virtual MSA generated by DPO model:

	Protein Structure		Protein Function	
Model	**CtP**	**SsP**	**LocP**	**MIB**
	ACC	ACC	ACC	ACC
w/o Virtual MSA	11.6	66.5	**58.3**	57.5
w/ Virtual MSA	**13.1**	**69.0**	56.4	**60.3**

Incorporating MSA from MSAGPT > Using single sequence only

NEURAL INFORMATION PROCESSING SYSTEMS

Tsinghua University

BioMap

TL;DR:

Employing a 2D evolutionary positional encoding scheme
and learning from AlphaFold2 Feedback,
MSAGPT generates constructive virtual MSA
to enable accurate protein structure predictions
in situations where natural co-evolutionary information is scarce

Code Repo: https://github.com/THUDM/MSAGPT