Retrieval Augmented Protein Language Models for Protein Structure Prediction

Pan Li [*1] **Xingyi Cheng** [*1 2] **Le Song** [1 2] **Eric Xing** [1 2 3]

Abstract

The advent of advanced artificial intelligence technology has significantly accelerated progress in protein structure prediction, with AlphaFold2 setting a new benchmark for prediction accuracy by leveraging the Evoformer module to automatically extract co-evolutionary information from multiple sequence alignments (MSA). To address AlphaFold2's dependence on MSA depth and quality, we propose two novel models: AIDO.RAGPLM and AIDO.RAGFold, pre-trained modules for **R**etrieval-**A**u**G**mented protein language model and structure prediction in an AI-driven Digital Organism (Song et al., 2024). AIDO.RAGPLM integrates pre-trained protein language models with retrieved MSA, surpassing single-sequence protein language models in perplexity, contact prediction, and fitness prediction. When sufficient MSA is available, AIDO.RAGFold achieves TM-scores comparable to AlphaFold2 while operating up to eight times faster, and significantly outperforms AlphaFold2 when MSA is insufficient (ΔTM-score=0.379, 0.116 and 0.059 for 0, 5 and 10 MSA sequences as input). Additionally, we developed an MSA retriever using hierarchical ID generation that is 45 to 90 times faster than traditional methods, expanding the MSA training set for AIDO.RAGPLM by 32%. Our findings suggest that AIDO.RAGPLM provides an efficient and accurate solution for protein structure prediction, particularly in scenarios with limited MSA data. The AIDO.RAGPLM model has been open-sourced and is available on https://huggingface.co/genbio-ai/AIDO.Protein-RAG-3B.

[*]Equal contribution [1]GenBio AI [2]Mohamed bin Zayed University of Artificial Intelligence [3]Carnegie Mellon University. Correspondence to: Le Song <le.song@genbio.ai>, Eric Xing <eric.xing@genbio.ai>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

The advent of advanced artificial intelligence technology has significantly accelerated progress in protein structure prediction. AlphaFold2 (Jumper et al., 2021), a pioneering method in this field, has set a new benchmark for prediction accuracy. Multiple sequence alignment (MSA) plays a crucial role in protein structure prediction. Unlike previous methods that required manual calculation of MSA features (Senior et al., 2020), AlphaFold2 leverages the Evoformer module to automatically extract co-evolutionary information from MSA, thereby enhancing the efficiency of information utilization.

However, the efficacy of structure prediction methods like AlphaFold2 is heavily dependent on the depth and quality of the MSA. Consequently, it is imperative to prepare an extensive sequence database. When the number of homologous sequences is insufficient, the performance of AlphaFold2 deteriorates significantly. To address this limitation, methods based on large-scale pre-trained protein language models have been proposed. For instance, ESMFold (Lin et al., 2023), OmegaFold (Wu et al., 2022), ESM3 (Hayes et al., 2024) and xTrimoPGLM-Fold (Chen et al., 2024b) have demonstrated commendable results using a single sequence as input. Nevertheless, even with 100-billion parameters, models like xTrimoPGLM and ESM3 remain inferior to AlphaFold2 in structure prediction when MSA is used as input, underscoring the importance of MSA. Although several PLM have attempted to integrate multiple sequences for training (see Appendix A), there is currently no validation for using retrieved augmented PLM for end-to-end protein structure prediction.

In this paper, we integrate pre-trained protein language models with retrieved MSA to propose a novel approach termed Protein Language Model with Retrieved Augmented MSA (RAGPLM) (see Figure 1). This approach allows for the incorporation of co-evolutionary information from MSA in structure prediction while compensating for insufficient MSA information through large-scale pre-training. We concatenate the query sequence with aligned homologous sequences into a long sequence (up to 12.8k) and perform pre-training by column span mask strategy based on a transformer encoder framework. Our method surpasses single-sequence protein language models in perplexity, contact

prediction, and fitness prediction. Subsequently, we utilized AIDO.RAGPLM as a feature extractor, integrating it with the folding trunks and Structure Modules to achieve end-to-end structural prediction (AIDO.RAGFold). Our findings indicate that when sufficient MSA is available, our method achieves results comparable to AlphaFold2 and is eight times faster; when MSA is insufficient, our method significantly outperforms AlphaFold2.

To expedite MSA acquisition, we also developed an MSA retriever using hierarchical ID generation. This retriever is 45 to 90 times faster than traditional HHblits (Steinegger et al., 2019) in MSA retrieval, which is used to expand the MSA training set for AIDO.RAGPLM by 32%.

2. Methods

Our method consists of three major components, MSA retriever, AIDO.RAGPLM and AIDO.RAGFold, which we explain more details below.

2.1. MSA retriever

Searching for multiple sequence alignments (MSAs) in large sequence databases is time-consuming. Inspired by (Wang et al., 2023) that generates relevant document identifiers by sequence-to-sequence network in document retrieval, we developed an MSA retriever to generate hierarchical identifiers for homologous sequences for a query protein sequence (see Figure 3). The protocol comprises three steps: (1) Construct hierarchical IDs for each sequence in UniClust30 (UC30) (Mirdita et al., 2016) through hierarchical K-means clustering of embedding; (2) Fine-tune a pretrained casual language model with 3-billion parameters (CLM-3B, (Cheng et al., 2024)) to memorize the ID of each sequence on UC30 dataset; (3) Continue to fine-tune the model to generalize to IDs of homologous sequences on the HHblits MSA dataset. For detailed training information, please refer to Appendix C. During inference, the MSA retriever generates each ID token sequentially, which corresponds to the nodes of the tree, until the UC30 node is reached. We perform multiple generations using different parameters and aggregate all retrieved sequences. Jackhmmer (Johnson et al., 2010) is then used to filter and align the homologous sequences. We use MSA Retriever to expand the MSA training data for AIDO.RAGPLM (see Appendix D).

2.2. AIDO.RAGPLM

We fine-tuned a pretrained masked language model with 3-billion parameters (MLM-3B, (Cheng et al., 2024)) using MSA data by concatenating the query sequence with homologous sequences (see Figure 1). We introduced several modifications to the standard BERT masking strategy (Devlin et al., 2019): (1) We randomly sampled $0.05 \times L$ span positions from a query sequence of length L, with span lengths following a geometric distribution (p=0.2), and capped the maximum length at 10. Our experiments revealed that this settings lead to an average of 15% of the query tokens were masked. (2) To prevent information leakage, when a residue was selected, all residues at the same index across all sequences (the column of the MSA matrix) were also masked. (3) When a column of MSA was selected for masking, the entire column was replaced with the `<MASK>` token in 80% of cases, with random amino acids in 10% of cases, and remained unchanged in the remaining 10% of cases. To help the model distinguish which tokens are from the same chain and which tokens have the same residue index, we use 2D rotary position embedding (Chen et al., 2024a; Su et al., 2023) to encode the tokens (see Figure 4 and Appendix E). For the details of training parameters, please refer to Table 6.

2.3. AIDO.RAGFold

Inspired by ESMFold (Lin et al., 2023), we use AIDO.RAGPLM as a feature extractor, and added the folding trunks (AlphaFold2 Evorformer without the column attention module) and Structure modules as a head to enable end-to-end protein structure prediction. During training, we also fine-tuned the AIDO.RAGPLM base model using LoRA (Rank=16, Alpha=16). We experimented with various numbers of folding trunks and found that 24 blocks were enough, which is half the number used in AlphaFold2 and ESMFold. Additionally, we replaced the ReLU activation function with GEGLU (Shazeer, 2020) in the transition module to enhance model performance. Our training procedure consists of two phases: initial training and fine-tuning. Detailed training parameters are provided in Appendix 7. Please refer to Appendix E for the details of the data description, model training and inference.

3. Results

Please refer to Appendix F for details of test datasets.

3.1. Comparing MSA retriever and HHblits

We employed HHblits and our MSA retriever to obtain MSAs of the test sequences from the UC30 database. For MSA retriever, we experimented with two sets of parameters: (1) beam search to generate 20 UC30 clusters; (2) Top-K (K=10) sampling for 64 UC30 clusters. The results were combined and used as input for AlphaFold2 (checkpoint: `model_3_ptm`). Table 5 demonstrates that although our results are not as favorable as those obtained with HHblits in terms of TM-score, our method is approximately 45 to 85 times faster. To address the issue of missing targets in the retriever (Depth $\leq$ 10), we combined the MSA retriever with HHblits. For samples with a depth of less than 10 in the retriever's results, we used HHblits to retrieve the MSA

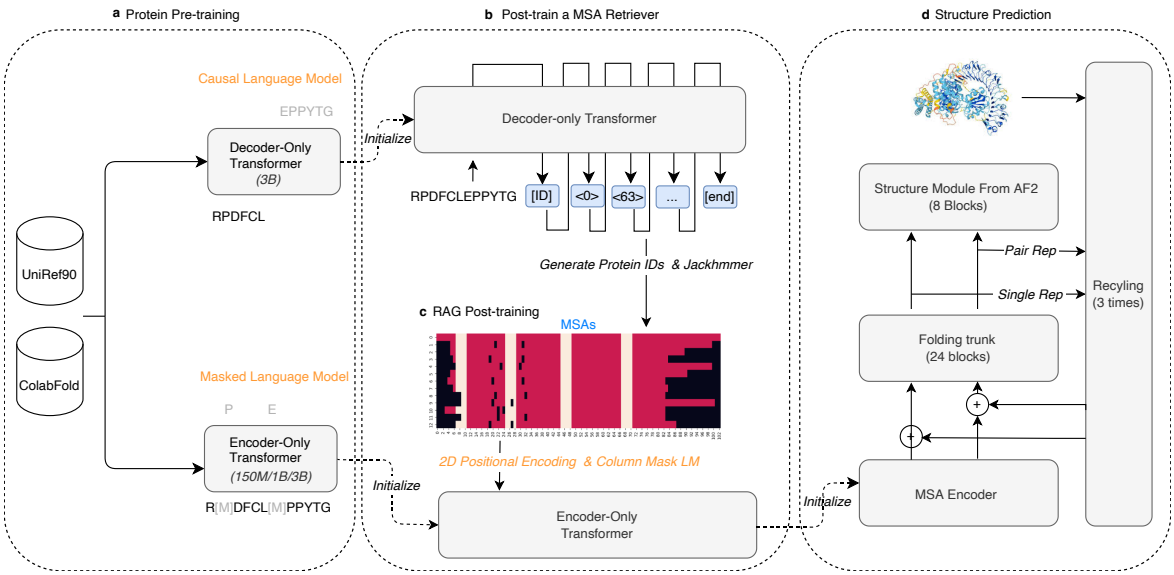

Figure 1. Schematic diagram of MSA retriever, AIDO.RAGPLM, and AIDO.RAGFold. (a) A decoder-only and an encoder-only transformer model are trained using UniRef90 and ColabFold protein sequence databases with CLM and MLM losses, respectively. (b) The MSA Retriever is fine-tuned on an MSA dataset to generate MSA sequence IDs from a query sequence, enabling the creation of a UniRef50 MSA training dataset. (c) AIDO.RAGPLM is trained on the UniRef50 MSA dataset using column span masking and recovery loss. (d) AIDO.RAGPLM acts as a feature extractor for protein structure prediction.

again. We found that the TM-score is comparable across four datasets when using HHblits, while still maintaining a 5 to 70-fold increase in speed.

3.2. AIDO.RAGPLM

Perplexity (PPL): We randomly replace 15% of the tokens in the sequence with <MASK> token. For MSA sequences, residues at the same index (the column of MSA) of masked query are also masked. We then calculate the perplexity of the masked tokens from the query sequence using ESM2-3B, MLM-3B, and AIDO.RAGPLM models. Table 4 shows that PLMRAG has the lowest PPL across all datasets, and as the number of homologous sequences increases, the PPL decreases further.

Unsupervised Contact Prediction: Following the methodology of (Rao et al., 2021), we randomly selected 20 chains as the training set and obtained $H \times L$ attention maps from the model, where H is the number of heads and L is the number of layers. Each attention map was symmetrized and adjusted using the Average Product Correction (APC) independently. Residue pairs with a distance of less than 8Å were defined as contacts. Logistic regression was employed to predict whether a residue pair is a contact (distance less than 8Å) using the $H \times L$ features as input. For AIDO.RAGPLM, only the attention map of the query sequence part was utilized. As shown in Table 1, AIDO.RAGPLM outperforms the two single-sequence models, despite its base model, MLM-3B, performing worse than ESM-3B on the CAMEO and Recent datasets.

Supervised Contact Prediction: We utilized the contact prediction dataset from trRosetta (Yang et al., 2020) to fine-tune the model. For all models, qkvo LoRA (Hu et al., 2021) and MLP LoRA were applied with (Rank=16, Alpha=16). The batch size was set to 8, and training was conducted for 25,000 steps. The checkpoint with the highest validation Top $L/5$ accuracy was used to evaluate the model. As shown in Table 2, AIDO.RAGPLM outperforms ESM2-3B and MLM-3B on both the validation and test sets.

Table 1. Unsupervised contact prediction.

		ESM-3B	MLM-3B	RAGPLM
Top L	CASP14	0.357	0.350	**0.389**
	CASP15	0.420	0.427	**0.451**
	CAMEO	0.493	0.483	**0.513**
	Recent	0.452	0.433	**0.477**
Top L/5	CASP14	0.348	0.365	**0.396**
	CASP15	0.381	0.380	**0.415**
	CAMEO	0.444	0.441	**0.470**
	Recent	0.436	0.403	**0.443**

ProteinGym zero-shot prediction. We obtained the substitutions dataset of Deep Mutational Scanning (DMS) assays from the ProteinGym website (Notin et al., 2023). For each mutation $t_{wt} \rightarrow t_{mut}$, we replace the wildtype token t_{wt} with a special <MASK> token. We then computed the log ratio $log(P_\theta(t_{mut})) - log(P_\theta(t_{wt}))$, where $P_\theta(t_{mut})$ represents the model's probability of the mutated token given the other tokens as input. To evaluate the model's performance, we calculated the Spearman correlation coefficient between the log ratio and the "DMS score" from the downloaded tables. As shown in Table 2, the AIDO.PLMRAG model achieved a higher score compared to the other two

Table 2. Result of supervised contact prediction and fitness prediction. Supervised contact prediction: 1,512 samples for validation set and test set. Fitness prediction: Spearman correlation coefficients of Deep Mutational Scanning (DMS) assays from ProteinGym. The column labeled "All" includes sequences with single and multiple mutations, while the column labeled "Single" includes sequences with only a single mutation. The data size is 207.

	Contact		Fitness	
	Validation	Test	All	Single
ESM2-3B	0.931	0.915	0.439	0.426
MLM-3B	0.916	0.910	0.430	0.408
PLMRAG	**0.938**	**0.927**	**0.462**	**0.437**

Table 3. TM-scores of AlphaFold2, AIDO.RAGFold, and ESMFold on four test datasets. HHblits MSAs were used as input for AlphaFold2 and AIDO.RAGFold. $N_{ensemble} = 4$.

Dataset	AF2	AIDO.RAGFold	ESMFold
CASP14	0.767	**0.776**	0.696
CASP15	**0.728**	0.726	0.639
CAMEO	0.864	**0.871**	0.854
Recent	**0.824**	0.823	0.775

single-sequence PLMs.

3.3. AIDO.RAGFold

We conducted a comparative analysis of TM-scores and runtime between AIDO.RAGFold and AlphaFold2 (checkpoint: model_3_ptm) using HHblits retrieved MSAs as input. The number of recycle ($N_{recycle}$) was fixed at three, and the maximum context length for RAG was constrained up to 25,600. Both AlphaFold2 and AIDO.RAGFold were executed with varying $N_{ensemble}$ (1, 2 and 4), resulting in AIDO.RAGPLM processing 4, 8, and 16 different MSAs, respectively (see Algorithm 1). Table 3 and Table 12 presents the TM-score of the two models. Table 10 presents the inference time, RMSD and LDDT. Our findings indicate that:

MSA ensembling enhances AIDO.RAGFold's performance: This improvement is primarily due to RAG's limited MSA context usage. Increasing $N_{ensemble}$ allows AIDO.RAGFold to use more homologous sequence information.

AIDO.RAGFold's performance is comparable to AlphaFold2: AIDO.RAGFold demonstrates a significantly faster inference speed, ranging from 8 times faster.

AIDO.RAGFold outperforms ESMFold: The inclusion of MSA significantly boosts AIDO.RAGFold's performance compared to ESMFold.

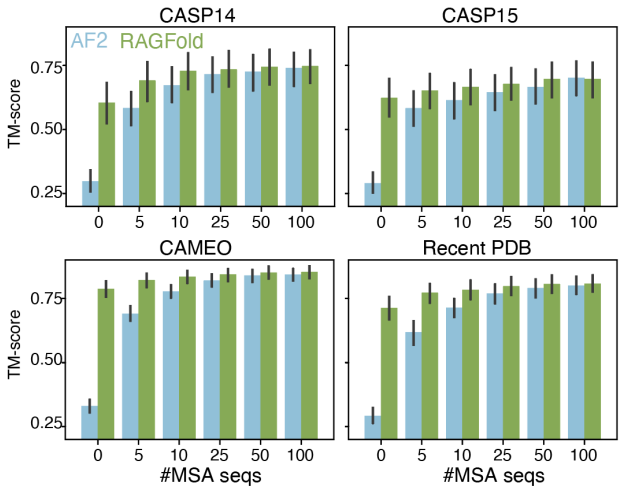

Figure 2. TM-scores of AlphaFold2 and AIDO.RAGFold on four test datasets with limited MSA sequences as input. AlphaFold2 and AIDO.RAGFold are represented by blue and green bars respectively. The x axis represents the upper bound of the MSA number.

To investigate the impact of the number of MSAs on AIDO.RAGFold's structural prediction accuracy, we randomly sampled 0, 5, 10, 25, 50, and 100 sequences from the HHblits MSA as input for both AlphaFold2 and AIDO.RAGFold. Table 11 and Figure 2 illustrate that AIDO.RAGFold's TM-scores surpass those of AlphaFold2 when the number of MSAs is limited. For instance, using the Recent PDB dataset, AIDO.RAGFold outperforms AlphaFold2 by margins of 0.420, 0.155, 0.070, 0.016, and 0.007 for 0, 5, 10, 25, 50, and 100 MSAs, respectively. However, it is noteworthy that without any MSA input, AIDO.RAGFold's performance lags behind ESMFold. Nevertheless, providing more than 5 MSAs enables AIDO.RAGFold to match ESMFold's performance, with the exception of the CAMEO dataset.

4. Conclusion

Our study introduces a novel MSA retrieval method based on ID generation, significantly accelerating MSA acquisition compared to traditional approaches. Utilizing this method, we expanded the existing MSA dataset and trained an MSA retrieval-enhanced protein language model. Our findings indicate that this model outperforms single-sequence models in tasks such as contact prediction and fitness prediction. Furthermore, we employed the embeddings from this language model for downstream end-to-end structure prediction, achieving results comparable to AF2, but with an approximately eightfold increase in speed. Notably, in scenarios with insufficient MSAs, our model substantially surpasses AF2, underscoring the critical importance of pre-trained models.

References

Ahdritz, G., Bouatta, N., Kadyan, S., Jarosch, L., Berenberg, D., Fisk, I., Watkins, A. M., Ra, S., Bonneau, R., and AlQuraishi, M. Openproteinset: Training data for structural biology at scale, 2023. URL https://arxiv.org/abs/2308.05326.

au2, T. F. T. J. and Bepler, T. Poet: A generative model of protein families as sequences-of-sequences, 2023. URL https://arxiv.org/abs/2306.06156.

Chen, B., Bei, Z., Cheng, X., Li, P., Tang, J., and Song, L. Msagpt: Neural prompting protein structure prediction via msa generative pre-training. *arXiv preprint arXiv:2406.05347*, 2024a.

Chen, B., Cheng, X., Li, P., ao Geng, Y., Gong, J., Li, S., Bei, Z., Tan, X., Wang, B., Zeng, X., Liu, C., Zeng, A., Dong, Y., Tang, J., and Song, L. xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein, 2024b. URL https://arxiv.org/abs/2401.06199.

Cheng, X., Chen, B., Li, P., Gong, J., Tang, J., and Song, L. Training compute-optimal protein language models. *bioRxiv*, pp. 2024–06, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024. doi: 10.1101/2024.07.01.600583. URL https://www.biorxiv.org/content/early/2024/07/02/2024.07.01.600583.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Johnson, L. S., Eddy, S. R., and Portugaly, E. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC Bioinformatics*, 11(1):431, Aug 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-431. URL https://doi.org/10.1186/1471-2105-11-431.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Ma, C., Zhao, H., Zheng, L., Xin, J., Li, Q., Wu, L., Deng, Z., Lu, Y., Liu, Q., and Kong, L. Retrieved sequence augmentation for protein representation learning, 2023. URL https://arxiv.org/abs/2302.12563.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., and Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 45(D1):D170–D176, 11 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1081. URL https://doi.org/10.1093/nar/gkw1081.

Notin, P., Kollasch, A. W., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Orenbuch, R., Gal, Y., and Marks, D. S. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*, 2023. doi: 10.1101/2023.12.07.570727. URL https://www.biorxiv.org/content/early/2023/12/08/2023.12.07.570727.

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. *bioRxiv*, 2021. doi: 10.1101/2021.02.12.430858. URL https://www.biorxiv.org/content/early/2021/02/13/2021.02.12.430858.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, Jan 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1923-7. URL https://doi.org/10.1038/s41586-019-1923-7.

Sgarbossa, D., Malbranke, C., and Bitbol, A.-F. Protmamba: a homology-aware but alignment-free protein state space model. *bioRxiv*, 2024. doi: 10.1101/2024.05.24.595730. URL https://www.biorxiv.org/content/early/2024/05/28/2024.05.24.595730.

Shazeer, N. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Song, L., Segal, E., and Xing, E. Toward AI-Driven Digital Organism: A System of Multiscale Foundation Models for Predicting, Simulating, and Programming Biology at All Levels . *Technical Report*, 2024.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1):473, Sep 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3019-7. URL https://doi.org/10.1186/s12859-019-3019-7.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Sun, H., Chen, Q., Xia, Y., Chi, C., Zhao, G., Liu, Z., Xie, X., Sun, H. A., Deng, W., Zhang, Q., and Yang, M. A neural corpus indexer for document retrieval, 2023. URL https://arxiv.org/abs/2206.02743.

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., and Peng, J. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020. doi: 10.1073/pnas.1914677117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1914677117.

A. Retrieval Augmented Protein Language Models.

Recent advancements in protein language models have attempted to integrate multiple homologous sequences for training. For example, the MSA Transformer (Rao et al., 2021), a model with 150 million parameters, utilizes aligned homologous sequences as input and employs self-supervised learning through random masking. This model has demonstrated superior performance compared to single-sequence models in downstream tasks such as fitness and contact prediction. Similarly, PoET (au2 & Bepler, 2023), an autoregressive generative model, concatenates unaligned sequences and trains them using next-token prediction. This enables the generation of entirely new sequences within the same family and the prediction of variant fitness. RSA (Ma et al., 2023) retrieves homologous sequences of the query using its dense sequence retriever and aggregates the information from the query and homologous sequences in pairs for downstream task prediction. This method not only achieves a retrieval speed significantly faster than traditional MSA methods but also delivers superior results in tasks such as fold classification, contact prediction, and localization. ProtMamba (Sgarbossa et al., 2024), leveraging the Mamba framework, extends the maximum sequence length up to 131k. By integrating autoregressive modeling and masked language modeling (MLM) with a fill-in-the-middle objective, ProtMamba can generate protein sequences and be utilized for downstream tasks such as fitness prediction.

B. MLM-3B model and CLM-3B model

The MLM-3B model (Cheng et al., 2024)) is a transformer encoder framework with 2.8 billion parameters. We utilize the same hyperparameters as ESM-3B, specifically: 36 layers, 40 heads, a hidden size of 2560, and an FFN hidden size of 6832. The training data is a mixture of the UniRef database and ColabFoldDB. We follow the BERT masking strategy: 15% of the tokens are selected for masking, with 80% replaced by special MASK tokens, 10% replaced by random amino acids, and the remaining 10% left unchanged. The learning rate schedule includes a 3% warm-up phase from 0 to 2.5e-4 followed by cosine decay from 2.5e-4 to 2.5e-5. Please refer to Table 6 for detailed information about MLM-3B. We train on 1,000 billion tokens and evaluate the model on two out-of-distribution datasets, with maximum identity to the training set being less than 0.9 and 0.5, respectively. The results are presented in Table 5.

The CLM-3B model (Cheng et al., 2024)) is a transformer decoder framework. It shares the same hyperparameters as the MLM-3B model and is trained on the same dataset. Our approach follows the training methodology of GPT, predicting the next token based on the given prefix. For detailed information about CLM-3B, please refer to Table 6.

C. MSA retriever

As described in the main text, training the MSA retriever involves three steps. Below, we detail the methods for each step.

C.1. Construct Hierarchical IDs for Each Sequence in UniClust30 via Hierarchical K-means Clustering of Embedding

The UC30 database (v2021_03) comprises 29 million clusters containing a total of 263 million sequences. The hierarchical ID is a multi-layer tree structure (see Figure 3), with each node having no more than 64 child nodes. Each leaf node corresponds to a sequence in UC30. To ensure the hierarchical ID reflects sequence similarity semantics (e.g., the similarity between 23-43-52-0 and 23-43-52-1 is higher than that between 23-43-52-0 and 23-43-34-5), we assign the ID tokens by clustering the embedding of sequences. So we use the MLM-3B model to generate embedding (dimension of 2560) for the 263 million sequences.

The ID of a sequence consists of two parts: (1) ID_center: derived from clustering the 29 million cluster centers; (2) ID_member: derived from clustering the members within a UC30 cluster.

ID_center: For each UC30 cluster, the longest sequence is selected as the representative sequence, and its embedding is used as the representative embedding of the cluster. We perform hierarchical K-means clustering (K=64) on the 29 million representative embedding, resulting in a tree with a degree of 64. We label all child nodes of each node from 0 to 63. Thus, for any node, we traverse from the root node to it in sequence to obtain its hierarchical ID, which is ID_center.

ID_member: For UC30 clusters with more than 64 members, we perform the same hierarchical K-means clustering on all members' embeddings to build ID_member.

The final ID for each sequence is obtained by concatenating ID_center and ID_member.

C.2. Fine-tune the CLM-3B Model to Memorize the ID of Each Sequence

We first build a Seq-ID dataset from UC30 dataset. Each sample comprises the query sequence, a special <ID> token, the hierachical ID tokens, and an <EOS> token. The CLM-3B model is trained with 500 billion tokens. After training, the model can generate the ID tokens with the query sequence and the <ID> token as a prefix until the <EOS> token or the UC30 cluster level token (purple circle in Figure 3). The learning rate is warmed up from 0 to 2.0e-5 for the first 2.5% of training tokens and then decays to 0 using a cosine schedule.

C.3. Continue to fine-tune the model to Generalize to IDs of Homologous Sequences

We use HHblits to search for MSAs from UC30 using UniRef50 (UR50) as query sequences, obtaining 23.7 million MSAs. We refer to this dataset as HHblits_MSA. When fine-tuning CLM-3B on this dataset, each sample comprises a query sequence, a special <ID> token, the ID tokens (randomly sampled from its homologous sequences), and a <EOS> token. We train 10 billion tokens on this dataset. Please refer to Table 6 for detailed information.

D. AIDO.RAGPLM training dataset

We utilized sequences from UniRef50 as queries to search for homologous sequences in UniClust30, subsequently constructing multiple sequence alignments (MSAs). UniRef50 comprises a total of 53.6 million sequences. Using HHblits, we searched all sequences, identifying over 25 homologous sequences for 23.7 million of them. This dataset was directly used as the training set, referred to as HHblits_MSA. The remaining 29.9 million sequences were input into MSA Retriever, resulting in 7.7 million sequences with more than 25 homologous sequences. This dataset was designated as Retriever_MSA. During training, AIDO.RAGPLM randomly sampled from the two datasets with probabilities of 0.75 and 0.25, respectively. Detailed information is provided in Figure 8.

E. Detailed description of AIDO.RAGFold architecture and inference.

We used the PDB database (release prior to January 1, 2024), the AlphaFold Database (with mean pLDDT ≥ 90) and OpenProteinSet residues with pLDDT ≥ 90 (Ahdritz et al., 2023) as the training set. Detailed information about the data is provided in Table 9. We ensured that all samples with sequence identity greater than 0.5 with the test set were excluded. The open-source OpenFold framework was employed to train our RAG-Fold model.

To feed the query tokens ($\in \mathbb{R}^L$, where L is the length of the query sequence) and MSA tokens ($\in \mathbb{R}^{N\times L}$, where L is the length of the query sequence) into the AIDO.RAGPLM model, the MSA tokens are flattened into the shape of $\mathbb{R}^{NL}$. We initialize a 2D positional encoding ($\in \mathbb{R}^{2\times L}$), where the first dimension represents the residue index for each sequence and the second dimension represents the sequence index (Chen et al., 2024a). To reduce the length of the sample, we remove G gap tokens that contain no information in the sequence. This adjustment changes the dimension of the sample to $\mathbb{R}^{NL-G}$ and the dimension of the positional encoding to $\mathbb{R}^{2\times(NL-G)}$. Figure 4 illustrates this process.

The output of the AIDO.RAGPLM model includes the embeddings of homologous sequences. We retain only the hidden states corresponding to the query tokens and input them into the downstream modules. Linear modules are employed to transform these hidden states into the MSA representation and Pair representation of folding trunks. For a detailed description, please refer to Algorithm 1.

During inference, due to the limitation of the input sample length (up to 25,600), the information from homologous sequences that the AIDO.RAGPLM model can utilize is restricted. To address this, we adopted the MSA ensembling method from AlphaFold2. Specifically, we sample a subset of up to 25,600 sequences from the all MSA sequences each time and run the AIDO.RAGPLM $N_{ensemble}$ times to average the resulting representations. This approach enables us to maximize the utilization of information from homologous sequences.

F. Test datasets

- **CASP14** (N=50): Protein targets obtained from the CASP14 website, accompanied by ground-truth structures.
- **CASP15** (N=53): Protein targets sourced from the CASP15 website, with corresponding ground-truth structures.
- **CAMEO** (N=194): Protein domains retrieved from the CAMEO website, covering the period from July 1, 2021, to

Table 4. Perplexity of various models and inputs across six sequence datasets. (N) denotes the dataset size, while (D) represents the number of homologous sequences used as input for AIDO.RAGPLM.

	CASP14	CASP15	CAMEO	Recent PDB	MaxID0.5	MaxID0.9
N	50	53	194	107	5,012	6,907
ESM2-3B	10.658	5.963	5.959	6.223	10.753	6.703
MLM-3B	8.905	6.355	5.631	5.671	10.959	6.816
RAGPLM (D=1)	10.114	6.599	6.321	6.357	10.718	6.816
RAGPLM (D=8)	9.303	6.360	6.212	5.999	10.222	6.494
RAGPLM (D=16)	8.980	6.167	6.167	5.666	9.989	6.317
RAGPLM (D=64)	8.724	5.803	5.995	5.329	9.381	5.840
RAGPLM (D=128)	8.391	**5.612**	5.707	5.296	**9.341**	**5.833**
RAGPLM (D=256)	**8.072**	5.741	**5.635**	**5.266**	9.359	5.871

Table 5. Performance Comparison of Various MSA Search Tools. In the case of Ours + hhblits, Ours MSAs with a depth of fewer than 10 were replaced with HHblits MSAs.

		Average time (s)	#(Depth $\geq$ 100) $\uparrow$	#(Depth $\leq$ 10) $\downarrow$	AlphaFold2 TM-score $\uparrow$
CASP14	hhblits	899	27	9	**0.757**
	Ours	19	31	11	0.696
	Ours + hhblits	172	31	7	0.748
CASP15	hhblits	2928	47	2	**0.731**
	Ours	32	42	8	0.668
	Ours + hhblits	94	46	2	0.723
CAMEO	hhblits	2761	172	5	**0.865**
	Ours	43	163	24	0.831
	Ours + hhblits	127	171	5	0.862
Recent PDB	hhblits	3138	91	1	0.824
	Ours	43	94	10	0.783
	Ours + hhblits	104	96	1	**0.827**

June 1, 2022.

- **Recent PDB** (N=107): Protein chains extracted from the PDB database, with release dates ranging from January 1, 2024, to July 1, 2024. The following criteria were applied to filter the chains: (1) a length range between 50 and 1500 residues; (2) exclusion of sequences containing non-standard amino acid types; (3) removal of sequences with repeat fragments, defined as having a bi-gram entropy greater than 4; (4) exclusion of sequences with more than 50% identity to the training set; (5) clustering of sequences at a 50% identity cutoff, selecting one representative sequence per cluster.

Table 6. Detailed training information of MLM-3B, CLM-3B, MSA Retriever and AIDO.RAGPLM.

	MLM-3B	CLM-3B	Retriever Step 1	Retriever Step 2	RAGPLM
Training data	UniRef + ColabFoldDB	UniRef + ColabFoldDB	UniClust30	HHblits_MSA	HHblits_MSA Retriever_MSA
Initial params	Random	Random	CLM-3B	Retriever Step 1	MLM-3B
Learning rate	2.5e-4	1.2e-4	2e-4	1.2e-4	1e-4
Training tokens	1000B	2300B	300B	10B	100B
Batch size	2560	2048	2048	1024	256
Micro batch size	4	4	4	4	1
Sample length	1024	2048	2048	1024	12800
Attention	Bi-directional	Causal	Causal	Causal	Bi-directional

Table 7. Detailed training information of AIDO.RAGFold

	Initial training	Fine-tuning
Sequence crop size	256	368
Maximum context length of RAG	16,384	12,800
Exponential moving average	Enabled	Enabled
Learning rate of LoRA	A: 1e-4, B: 1.6e-3	A: 1e-4, B: 1.6e-3
Learning rate of folding trunks Structural modules	First 90%: 1e-3 Last 10%: 5e-4	5e-4
Batch size	First 90%: 128 Last 10%: 256	256
Warm up	First 2000 steps	N/A
Structural violation loss weight	0	0.1
”Experimentally resolved” loss weight	0	0.01
Training samples (million)	10	6

Table 8. Training data of AIDO.RAGPLM. 23.7 million MSAs are collected by HHblits and 7.7 million MSAs are collected by MSA Retriever.

	#Seqs	#Query tokens	Sample Weight
HHblits_MSA	23.7M	6.5B	0.75
Retriever_MSA	7.7M	2.4B	0.25

Table 9. **Training data of AIDO.RAGFold.**

Dataset	#Chains	#Clusters	Sample ratio
PDB	440,952	34,961	25%
AlphaFold DB Distil	4,457,794	1,829,120	N/A
OpenProteinSet Distil	259,343	242,079	N/A
Distil mixed	4,711,621	2,002,005	75%

Table 10. Inference time, RMSD and LDDT of AlphaFold2 (AF2), AIDO.RAGFold (AIDO.RF), and ESMFold on four test datasets.

	Dataset	AF2			AIDO.RF			ESMFold
		ens=1	ens=2	ens=4	ens=1	ens=2	ens=4	
Inferene Time (wo MSA search)	CASP14	93.9	121.9	163.8	8.7	17.5	34.4	8.3
	CASP15	95.4	127.4	171.6	11.4	22.8	45.2	8.5
	CAMEO	90.0	116.1	149.3	11.3	22.5	44.5	6.0
	Recent	99.6	130.2	175.4	11.7	23.2	45.9	9.1
RMSD	CASP14	6.152	5.767	5.726	6.788	6.521	6.281	8.558
	CASP15	15.479	15.387	15.351	12.375	13.451	12.930	16.055
	CAMEO	3.555	3.607	3.597	3.670	3.635	3.633	4.131
	Recent	5.428	5.431	5.213	6.263	6.161	6.071	7.080
LDDT	CASP14	0.784	0.794	0.797	0.795	0.804	0.813	0.732
	CASP15	0.841	0.843	0.842	0.836	0.840	0.840	0.777
	CAMEO	0.890	0.890	0.890	0.893	0.894	0.896	0.876
	Recent	0.893	0.893	0.893	0.891	0.894	0.893	0.853

Table 11. TM-scores of AlphaFold2, AIDO.RAGFold, and ESMFold on four test datasets with limited MSA sequences as input.

	#MSA=0		#MSA=5		#MSA=10		#MSA=25		#MSA=50		#MSA=100		
	AF2	RAG	AF2	RAG	AF2	RAG	AF2	RAG	AF2	RAG	AF2	RAG	ESMFold
CASP14	0.298	0.604	0.584	0.692	0.672	0.728	0.716	0.735	0.726	0.744	0.740	0.748	0.696
CASP15	0.290	0.624	0.583	0.652	0.614	0.666	0.645	0.678	0.666	0.697	0.701	0.697	0.639
CAMEO	0.330	0.787	0.690	0.822	0.777	0.834	0.820	0.844	0.840	0.851	0.843	0.853	0.854
Recent	0.292	0.712	0.618	0.773	0.714	0.784	0.769	0.798	0.790	0.806	0.800	0.807	0.775

Table 12. TM-scores of AlphaFold2, AIDO.RAGFold, and ESMFold on four test datasets. HHblits MSAs were used as input for AlphaFold2 and AIDO.RAGFold. "ens" denotes the number of MSA ensembles.

Dataset	AF2			AIDO.RAGFold			ESMFold
	ens=1	ens=2	ens=4	ens=1	ens=2	ens=4	
CASP14	0.754	0.766	0.767	0.752	0.764	**0.776**	0.696
CASP15	0.725	0.727	**0.728**	0.722	0.727	0.726	0.639
CAMEO	0.864	0.863	0.864	0.868	0.869	**0.871**	0.854
Recent	**0.824**	0.823	**0.824**	0.820	0.823	0.823	0.775

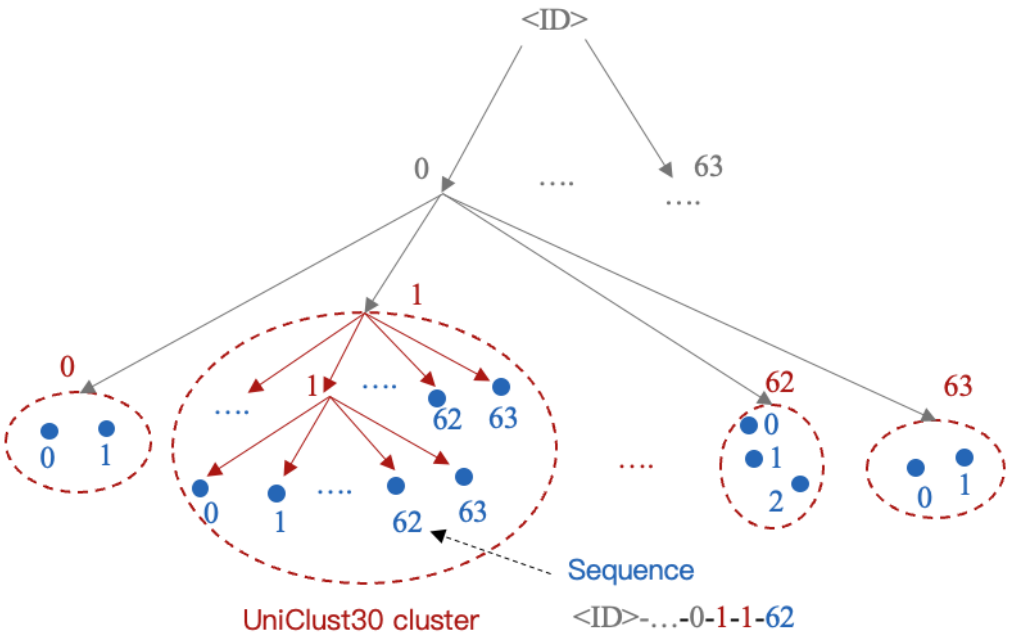

Figure 3. Schematic Diagram of Hierarchical ID of UniClust30 Sequences. The UC30 sequences are organized into a tree structure with a branching factor of 64. Each leaf node represents an individual sequence, while each UC30 cluster corresponds to an internal node of the tree. The hierarchical ID of a sequence is determined by traversing from the root node to the corresponding leaf node.

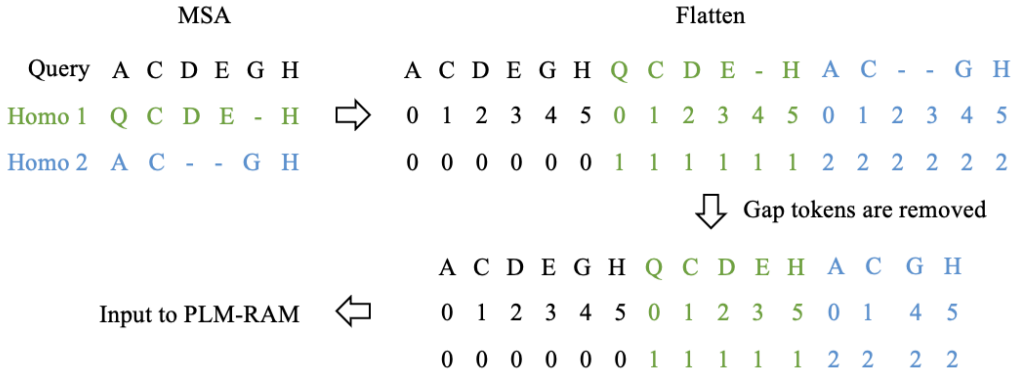

Figure 4. Schematic Diagram of AIDO.RAGPLM input.

Algorithm 1 AIDO.RAGFold

Input: query_tokens $\in \{0, \ldots, 20\}^L$ $\{L$: Number of residues$\}$
Input: msa_tokens $\in \{0, \ldots, 20\}^{N \times L}$ $\{N$: Number of sequences$\}$
Input: $N_{recycle}$ {Number of recycles}
Input: $N_{ensemble}$ {Number of ensemble}
msa_emb_prev, pair_emb_prev, cbeta_prev = 0, 0, 0
for $i\_rec \in 0, \ldots, N_{recycle}$ **do**
  msa_emb, pair_emb = 0, 0
  for $i\_ens \in 1, \ldots, N_{ensemble}$ **do**
    msa_tokens_ens = GreedyMaxSample(msa_tokens) {GreedyMaxSample sample a subset of MSA with maximum diversity}
    msa_emb, pair_emb += RAGPLM-Embedder(query_tokens, msa_tokens_ens)
  end for
  msa_emb, pair_emb /= $N_{ensemble}$
  msa_emb, pair_emb += RecyclingEmbedder(msa_emb_prev, pair_emb_prev, cbeta_prev)
  msa_emb, pair_emb = FoldTrunk(msa_emb, pair_emb)
  atom_pos, plddt = StructureModule(msa_emb, pair_emb)
  msa_emb_prev, pair_emb_prev, cbeta_prev = msa_emb, pair_emb, get_cbeta(atom_pos)
end for
Output: atom_pos, plddt

Algorithm 2 RAGPLM-Embedder

Input: query_tokens $\in \{0, \ldots, 20\}^L$ $\{L$: Number of residues$\}$
Input: msa_tokens $\in \{0, \ldots, 20\}^{N \times L}$ $\{N$: Number of sequences$\}$
hid_stat = RAGPLM (query_tokens, msa_tokens) {hid_stat $\in \mathbb{R}^{(NL-G) \times D}$}
hid_stat = hid_stat[:L] {hid_stat $\in \mathbb{R}^{L \times D}$}
msa_emb = MSA_Transform(hid_stat) {msa_emb $\in \mathbb{R}^{L \times 256}$}
pair_emb = PAIR_Transform(hid_stat) {pair_emb $\in \mathbb{R}^{L \times L \times 128}$}
pair_emb $+=$ relpos(res_ind) {res_ind is short for residue index}
msa_emb $+=$ aa_embedder(query_tokens)
Output: msa_emb, pair_emb

Algorithm 3 MSA_Transform

Input: hid_stat $\in \mathbb{R}^{L \times D}$
msa_emb = Linear(hid_stat) {msa_emb $\in \mathbb{R}^{L \times 256}$}
Output: msa_emb

Algorithm 4 PAIR_Transform

Input: hid_stat $\in \mathbb{R}^{L \times D}$
hid_stat = LayerNorm(hid_stat)
pair_emb = OuterAdd(Linear(hid_stat), Linear(hid_stat)) {pair_emb $\in \mathbb{R}^{L \times L \times 128}$}
Output: msa_emb