DISTILLATION OF MSA EMBEDDINGS TO FOLDED PROTEIN STRUCTURES WITH GRAPH TRANSFORMERS

A PREPRINT

Allan Costa,^{1,2,†} Manvitha Ponnappati,^{1,2} Joseph M. Jacobson,^{1,2} Pranam Chatterjee^{1,2}

¹Center for Bits and Atoms, ²Media Lab,
Massachusetts Institute of Technology

[†]Corresponding author: allanc@mit.edu

June 2, 2021

ABSTRACT

Determining the structure of proteins has been a long-standing goal in biology. Language models have been recently deployed to capture the evolutionary semantics of protein sequences. Enriched with multiple sequence alignments (MSA), these models can encode protein tertiary structure. In this work, we introduce an attention-based graph architecture that exploits MSA Transformer embeddings to directly produce three-dimensional folded structures from protein sequences. We envision that this pipeline will provide a basis for efficient, end-to-end protein structure prediction.

Introduction

Elucidating protein structure is critical for understanding protein function. However, structure determination via experimental methods such as x-ray crystallography [Smyth, 2000] or cryogenic electron microscopy (cryo-EM) [Murata and Wolf, 2018], is a time-consuming, difficult, and expensive task. Classical modeling methods have attempted to solve this task *in silico*, yet are computationally prohibitive [Rohl et al., 2004, Hollingsworth and Dror, 2018, Wang et al., 2016]. Recently, machine learning approaches have been deployed to harvest available structural data and efficiently map sequence to structure [Yang et al., 2020, Senior et al., 2020].

Transformer models are sequence-to-sequence architectures that have been shown to capture the contextual semantics of words [Vaswani et al., 2017] and have been widely deployed as language models [Devlin et al., 2019, Brown et al., 2020]. The sequential structure of proteins, imposed by the central dogma, along with their hierarchical semantics, developed by Darwinian evolution, makes them a natural target for language modeling. Recently, transformers have been deployed to learn protein sequence distributions, and generate latent embeddings that grasp relevant structure [Rives et al., 2021, Elnaggar et al., 2020, Vig et al., 2020], most notably tertiary structural information [Rao et al., 2020]. Augmenting input sequences with their evolutionarily-related counterparts, in the form of a multiple sequence alignment (MSA), further strengthens the predictive power of these transformer architectures, as demonstrated by state-of-art contact prediction results [Rao et al., 2021].

In this study, we leverage MSA Transformer embeddings within a geometric deep learning architecture to directly map protein sequences to folded, three-dimensional structures. In contrast to existing architectures, we directly estimate point coordinates in a learned, canonical pose, which removes the dependency on classical methods for resolving distance maps, and enables gradient passing for downstream tasks, such as side chain prediction and protein refinement. Overall, our results provide a bridge to a complete, end-to-end folding pipeline.

Methodology and Results

We treat the protein folding problem as a graph optimization problem. First, we harvest information-dense embeddings produced by the MSA Transformer [Rao et al., 2021], and use these embeddings to produce initial node and edge hidden representations in a complete graph. To process and structure geometric information, we employ the attention-based architecture of the Graph Transformer, as proposed by [Shi et al., 2021]. Final node representations are then projected into Cartesian coordinates through a learnable transformation, and the resulting induced distance maps are compared to ground truth to define the loss for training.

MSA Transformer Data Augmentation

The MSA Transformer is an unsupervised protein language model that produces information-rich residue embeddings [Rao et al., 2021]. In contrast to other protein language models, it operates on two dimensional inputs consisting of a length- N query sequence along with its MSA sequences. It utilizes an Axial Transformer [Ho et al., 2019] as an efficient attention-based architecture for performing computation on its layers’ $O(N \cdot S)$ representations, where S is the total number of input MSA sequences.

Our model seeks to operate on graph features distilled from MSA Transformer encodings. Last-layer residue embeddings capture individual and contextual residue properties. Similarly, the vector formed by pairwise attention scores at each layer and head captures attentive interactions between residue pairs. The richness of information present at these vectors was demonstrated in state-of-the-art contact prediction [Rao et al., 2021]. We perform the natural step of extending those individual and pairwise embeddings to node and edge representations, and show how learning over the resulting graph can resolve a protein’s three-dimensional structure.

For this study, we employed the 100 million parameter-sized ESM-MSA-1 model [Rao et al., 2021], which was trained on 26 million MSAs queried from UniRef50 and sourced from UniClust30. ESM-MSA-1 produces N residue embeddings $h_i^* \in \mathbb{R}^{768}$ and $N \times N$ attention score traces $h_{ij}^* \in \mathbb{R}^{144}$ for each input sequence. Since the MSA Transformer is computationally expensive to evaluate for large S even in the context of inference, we precomputed the encodings and made them readily available for training. We used $S = 64$ and stored residue embeddings $\{h_i^*\}$ and attention score traces, $\{h_{ij}^*\}_{j>i}$ for each query sequence.

For training and validation, we used the ESM Structural Split [Rives et al., 2021], which builds upon trRosetta’s training dataset [Yang et al., 2020]. To overcome the bottleneck associated with reading large encodings directly from the file system, we fixed our splits to the first superfamily split as specified in Rives, et al., and serialized its MSA Transformer encodings into tar shards. A virtual layer of data shuffling was added through the WebDataset framework [Aizman et al., 2020]. The resulting dataset of graph features has 0.25 TB.

Graph Building

We treat a protein as an attributed complete graph. Let H_V and H_E be the dimensionalities of node and edge representations, respectively. These attributes are extracted from MSA-Transformer embeddings through standard deep neural networks:

$$h_i = \sigma(W_E^{(D_V)} \dots (\sigma(W_V^{(0)} h_i^*)) \dots)$$

$$h_{ij} = \sigma(W_D^{(D_E)} \dots (\sigma(W_E^{(0)} h_{ij}^*)) \dots)$$

Where $h_i \in \mathbb{R}^{H_V}$, $h_{ij} \in \mathbb{R}^{H_E}$, $\sigma(\cdot)$ is a ReLU nonlinearity, and D_V and D_E are the depths of node and edge information extractors, respectively. W denotes dense learnable parameters, and here and in the following equations we omit bias terms.

Graph Transformer

The Graph Transformer was introduced in [Shi et al., 2021] to incorporate edge features directly into graph attention. This is possible by directly summing transformations of edge attributes to the original keys and values of the attention mechanism. We approach protein folding with a variation of this architecture. Consider layer l node hidden states, $\{h_i^l\}$, and similarly learned edge latent states $\{e_{ij}\}$. If we employ C attention heads, a layer update can be written as

$$h_i^{l+1} = W_R^{(l)} h_i^l + \sigma \left(W_A^{(l)} \bigoplus_{c=1}^C \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}^{(l,c)} (v_j^{(l,c)} + e_{ij}^{(c)}) \right)$$

Figure 1: An overview of a sequence-to-structure pipeline utilizing the MSA-Transformer and a Graph Transformer. (a) We first augment a length- N protein sequences to S of its MSA. The MSA-Transformer operates over this token matrix to produce enriched individual and pairwise embeddings. We store those embeddings that are from the original query sequence. (b) Deep neural networks extract relevant features and structure latent states for a downstream graph transformer. Individual and pairwise embeddings are assigned to nodes and edges, respectively. (c) A graph transformer operates on node representations through an attention-based mechanism that considers pairwise edge attributes. The final node encodings are projected directly to $\mathbb{R}^3$, and the induced distogram is computed for the loss.

Where $\oplus$ denotes concatenation, $W_A^{(l)}$ and $W_R^{(l)}$ are learnable projections. As in the original architecture, we apply batch normalization to each layer. The attention scores $\alpha_{ij}^{(l,c)}$, node values $v_j^{(l,c)}$ and edge values $e_{ij}^{(c)}$ are obtained from learnable transformations of the original node hidden states and edge attributes:

$$\begin{aligned} q_i^{(l,c)} &= W_q^{(l,c)} h_i^{(l)} & k_i^{(l,c)} &= W_k^{(l,c)} h_i^{(l)} \\ v_i^{(l,c)} &= W_v^{(l,c)} h_i^{(l)} & e_{ij}^{(c)} &= W_e^{(c)} h_{ij} \end{aligned}$$

The attention scores are normalized according to graph attention:

$$\bar{\alpha}_{i,j}^{(l,c)} = (q_i^{(l,c)})^\top (k_i^{(l,c)} + e_{ij}^{(c)}) \quad \alpha_{i,j}^{(l,c)} = \frac{\exp[\bar{\alpha}_{i,j}^{(l,c)}]}{\sum_{u \in \mathcal{N}(i)} \exp[\bar{\alpha}_{i,u}^{(l,c)}]}$$

66 To hold computational costs roughly constant, we let $\{q_i^c, v_i^c, k_i^c, e_{ij}^c\} \in \mathbb{R}^{H_v/C}$ as in standard Transformer architec-
67 tures.

68 Cartesian Projection and Loss

We train a predictor to recover coordinates of each residue in a learned canonical pose:

$$X_i = W_X h_i^{(L)}$$

Where $X_i \in \mathbb{R}^3$. To train our network, we use a distogram-based loss function on the resulting distance map. Let $\hat{D}_{ij} = \|X_i - X_j\|_2$ be the induced Euclidean distance between the Cartesian projections of nodes i and j , and D_{ij} be the ground truth distance. Our loss is based on the L_1 -norm of the difference between those values:

$$\mathcal{L} = \frac{1}{N^2} \sum_i \sum_j^N \|\hat{D}_{ij} - D_{ij}\|_1$$

Figure 2: Comparison of predicted distograms and three-dimensional arrangement with their ground truth counterparts for samples from (A) ESM Structural Split dataset, and (B) CASP13 free modeling targets. For each prediction-ground truth pair, a PDB name is indicated on the left. Aligned $C\alpha$ traces and metrics are indicated on the right. Blue traces denote model predictions, red traces denote ground truth. Traces were produced by fitting splines to the sequence of predicted $C\alpha$ coordinates.

Model Training

To optimize our trained model, we performed a shallow random hyperparameter search for $H_V \in \{32, 64, 128, 256\}$, $H_E \in \{32, 64, 128\}$, $L \in \{3, 6, 10, 15\}$, $C \in \{1, 2, 4\}$. We utilized the Adam Optimizer with $\text{lr} \in \{1 \times 10^{-3}, 3 \times 10^{-4}, 1 \times 10^{-4}, 3 \times 10^{-5}, 1 \times 10^{-5}\}$. We further tested variations of the loss function, testing the MSE loss and weighted versions of L_1 and MSE for batch sizes $B \in \{10, 15, 30\}$.

To handle GPU memory constraints, we employed gradient checkpointing at each Graph Transformer layer. Models were trained in parallel on NVIDIA V100s provided by the MIT SuperCloud HPC [Reuther et al., 2018].

In total, we performed 40 search training runs, with a maximum of 70 epochs and an early stop with a patience of 3 for validation loss. Our best model trained for 17 hours without registering early stopping. With $H_V = H_E = 64$, $L = 10$ and $C = 1$, this model only possesses a total of 382K parameters. We used $\text{lr} = 3 \times 10^{-4}$ and $B = 30$, as well as an L_1 loss, and achieved $\mathcal{L}_{\text{val}} = 2.25$ and $\text{GDT_TS}_{\text{val}} = 40.58$.

CASP13 Evaluation

To investigate the generalization of our model, we evaluated it on the free modeling targets from the 13th edition of the Critical Assessment of Protein Structure Prediction (CASP13). We chose to benchmark our model against the performance of the current state-of-the-art public architecture: trRosetta [Yang et al., 2020]. trRosetta considers a sequence's MSA to predict distance probability volumes as well as relevant interresidue orientations. In contrast to our model, trRosetta relies on restraints derived from the predicted distance and orientations for downstream Rosetta minimization protocols [Rohl et al., 2004]. For each distance, we consider trRosetta's best prediction as its expected value or its maximum likelihood estimate. We utilized dRMSD (distogram RMSD) between predicted distances and ground truth as our evaluation metric. To make a direct comparison, we only consider distances that lie within trRosetta's binning range (2-20 Å).

Our results demonstrate that the Graph Transformer model, despite its size, is competitive to trRosetta's estimates (Figure 3, Table 1). It is worth noting that our architecture resolves backbone structure as its main output and uniquely and deterministically produces distances, whereas trRosetta operates within a probabilistic domain that does not need three-dimensional resolution. These early results thus suggest potential for improved predictive capability with larger model capacity and downstream protein refinement.

Figure 3: Qualitative assessment of model predictions for CASP13 free modeling targets. Note that our model is able to capture long range interactions, whereas trRosetta by construction is limited to short range dependencies. We highlight T0950 and T0963D2 as examples of challenging reconstructions for our network.

	T0987	T0969	T0955d1	T0998
Graph Transformer	3.722	3.080	5.346	3.476
trRosetta (argmax)	2.135	1.583	2.400	1.482
trRosetta (expectation)	1.638	1.288	2.160	1.247
	T0990	T0958d1	T0968s2d1	T0963d2
Graph Transformer	3.017	2.886	3.380	7.853
trRosetta (argmax)	1.356	1.947	1.927	4.039
trRosetta (expectation)	1.078	1.796	1.695	2.982
	T0953s2d3	T1010	T0968s1d1	T0957s2d1
Graph Transformer	5.404	4.002	3.905	2.559
trRosetta (argmax)	4.647	2.048	2.226	1.700
trRosetta (expectation)	3.681	1.531	1.797	1.492
	T0950	T0953s1	T0953s2	T1022s1
Graph Transformer	3.392	2.698	4.158	2.604
trRosetta (argmax)	1.542	1.897	3.868	1.665
trRosetta (expectation)	1.148	1.618	3.144	1.433

Table 1: CASP13 Free Modeling benchmarks of dRMSD for our architecture’s induced distances and trRosetta’s expectation and argmax distances, against ground truth. We only consider distances that lie within trRosetta’s binning range.

Discussion

In this work, we revisit the protein folding problem and highlight the role of unsupervised language models in providing a meaningful basis for the sequence-to-structure prediction task. We present a strategy to encapsulate MSA Transformer embeddings and attention traces in a geometric framework, and formalize a graph learning pipeline to reason positional information.

Overall, our results demonstrate the remarkably expressive power of language models and, in particular, of MSA-augmented architectures. To demonstrate a versatile bridge between sequence and three-dimensional structure, we trained a downstream model to produce $C\alpha$ -traces which, before any refinement is performed, induce distograms with high similarity to ground truth.

We emphasize that our model, in its current state, tackles only a step of the protein structure prediction problem. With only 382K parameters, it serves as a fast and scalable solution to resolving the position of protein backbones. Furthermore, it extends learning beyond distogram prediction and provides a natural foundation for downstream tasks, such as side chain prediction and protein refinement. Most importantly, we hypothesize that by increasing model capacity, dataset size, and training time, our model's predictive capability can improve significantly.

In total, this study builds upon the recent groundbreaking work in protein representation learning and protein language modeling by the community. The integration of diverse network architectures and pretrained models, as demonstrated here, will enable the eventual efficient solution of the protein structure prediction problem.

Acknowledgements

We thank the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing the HPC and database resources that have contributed to the research reported within this manuscript. We thank Thrasyvoulos Karydis for the fruitful discussions and insights for this project.

References

- [Aizman et al., 2020] Aizman, A., Maltby, G., and Breuel, T. (2020). High performance i/o for large scale deep learning.
- [Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.
- [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- [Elnaggar et al., 2020] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2020). ProtTrans: Towards cracking the language of life’s code through self-supervised learning.
- [Ho et al., 2019] Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. (2019). Axial attention in multidimensional transformers.
- [Hollingsworth and Dror, 2018] Hollingsworth, S. A. and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143.
- [Murata and Wolf, 2018] Murata, K. and Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(2):324–334.
- [Rao et al., 2021] Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., Sercu, T., and Rives, A. (2021). MSA transformer.
- [Rao et al., 2020] Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. (2020). Transformer protein language models are unsupervised structure learners.
- [Reuther et al., 2018] Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., et al. (2018). Interactive supercomputing on 40,000 cores for machine learning and data analysis. pages 1–6.
- [Rives et al., 2021] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.
- [Rohl et al., 2004] Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein structure prediction using rosetta. In *Methods in Enzymology*, pages 66–93. Elsevier.
- [Senior et al., 2020] 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. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.
- [Shi et al., 2021] Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., and Sun, Y. (2021). Masked label prediction: Unified message passing model for semi-supervised classification.
- [Smyth, 2000] Smyth, M. S. (2000). x ray crystallography. *Molecular Pathology*, 53(1):8–14.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
- [Vig et al., 2020] Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. (2020). BERTology meets biology: Interpreting attention in protein language models.
- [Wang et al., 2016] Wang, S., Li, W., Zhang, R., Liu, S., and Xu, J. (2016). CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Research*, 44(W1):W361–W366.
- [Yang et al., 2020] Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503.