Accurate Protein Structure Prediction by Embeddings and Deep Learning Representations

Iddo Drori
Columbia University and Cornell University
idrori@cs.columbia.edu,idrori@cornell.edu

Darshan Thaker
Columbia University
dbt2120@columbia.edu

Arjun Srivatsa
Columbia University
a.srivatsa@columbia.edu

Daniel Jeong
Columbia University
dpj2108@columbia.edu

Yueqi Wang
Columbia University
yueqi.wang@columbia.edu

Linyong Nan
Columbia University
ln2401@columbia.edu

Fan Wu
Columbia University
fw2321@columbia.edu

Dimitri Leggas
Columbia University
ddl2133@columbia.edu

Jinhao Lei
Columbia University
jl5207@columbia.edu

Weiyi Lu
Columbia University
wl2681@columbia.edu

Weilong Fu
Columbia University
wf2232@columbia.edu

Yuan Gao
Columbia University
yg2541@columbia.edu

Sashank Karri
Columbia University
srk2177@columbia.edu

Anand Kannan
Columbia University
ak4250@columbia.edu

Antonio Khalil Moretti
Columbia University
amoretti@cs.columbia.edu

Chen Keasar
Ben-Gurion University
keasar@cs.bgu.ac.il

Itsik Pe'er
Columbia University
itsik@cs.columbia.edu

Abstract

Proteins are the major building blocks of life, and actuators of almost all chemical and biophysical events in living organisms. Their native structures in turn enable their biological functions which have a fundamental role in drug design. This motivates predicting the structure of a protein from its sequence of amino acids, a fundamental problem in computational biology. In this work, we demonstrate state-of-the-art protein structure prediction (PSP) results using embeddings and deep learning models for prediction of backbone atom distance matrices and torsion angles. We recover 3D coordinates of backbone atoms and reconstruct full atom protein by optimization. We create a new gold standard dataset of proteins which is comprehensive and easy to use. Our dataset consists of amino acid sequences, Q8 secondary structures, position specific scoring matrices, multiple sequence alignment co-evolutionary features, backbone atom distance matrices, torsion angles, and 3D coordinates. We evaluate the quality of our structure prediction by RMSD on the latest Critical Assessment of Techniques for Protein Structure Prediction (CASP13) test data and demonstrate competitive results with winning teams and AlphaFold. We make our data, models, and code publicly available upon publication.

Preprint. Under review.

1 Introduction

Proteins are necessary for a variety of functions within cells including transport, antibody, enzyme and catalysis. They are polymer chains of amino acid residues, whose sequences (aka primary structure) dictate stable spatial conformations, with particular torsion angles between successive monomers. To perform their functions, the amino acid residues must fold into proper configurations. The sequence space of proteins is vast, 20 possible residues per position, and evolution has been sampling it over billions of years. Thus, current proteins are highly diverse in sequences, structures and functions. The high throughput acquisition of DNA sequences, and therefore ubiquity of known protein sequences stands in contrast to the limited availability of 3D structures, which are more functionally relevant. From a physics standpoint, the process of protein folding is a search for the minimum energy conformation which happens in nature paradoxically fast [11]. Unfortunately, explicitly computing the energy of a protein conformation, along with its surrounding water molecules, is extremely complex. Forward simulation approaches have been prohibitively slow on even modest-length proteins, with successes reported only for smaller peptides across timescales that are orders of magnitude too short, and even these required specialized hardware [13] or massive crowd sourcing [10]. Inferring local *secondary structure* [7] consists of linear annotation of structural elements along the sequence [9]. Inferring *tertiary structure* consists of resolving the 3D atom coordinates of proteins. When highly similar sequences are available with known structures, this homology can be used for modeling. Recently, PSP had been tackled by first predicting contact points between amino acids, and then leveraging the contact map to infer structure. A primary indicator of contact between a pair of amino acids is their tendency to have correlated and compensatory mutations during evolution. The availability of large scale data on DNA, and therefore protein sequences allows detection of such co-evolutionary constraints from sets of sequences that are homologous to a protein of interest. Registering such contacting pairs in a matrix facilitates a framework for their probabilistic prediction. This contact map matrix can be generalized to register distances between amino acids [15].

Machine learning approaches garnered recent success in PSP [3, 2] and its sub-problems [14]. These leverage available repositories of tertiary structure [8] and its curated compilations [12] as training data for models that predict structure from sequence. Specifically, the recent biannual critical assessment of PSP methods [4] featured multiple such methods. Most prominently, a closed-source, ResNet-based architecture [5] has achieved impressive results in the CASP evaluation settings, based on representing protein structures both by their distance matrices as well as their torsion angles. In this work, we demonstrate state-of-art protein structure prediction results using deep learning models to predict backbone atom distance matrices and torsion angles, recover backbone atom 3D coordinates and reconstruct the full atom protein by optimization. Our three key contributions are:

- A gold standard dataset of around 75k proteins described in Table 1 which we call the CUProtein dataset which is easy to use in developing deep learning models for PSP.
- Competitive results on CASP13 and a comparison with AlphaFold (A7D) [5].
- Publicly available source code for both protein structure prediction using deep learning models and protein reconstruction.

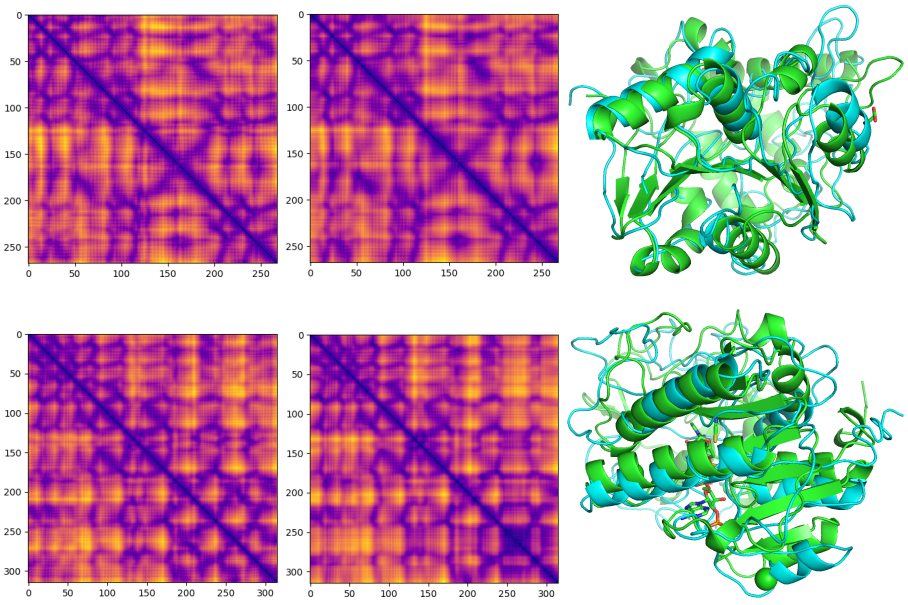
This work explores encoded representation for sequences of amino acids alongside their auxiliary information. We offer full access to data, models, and code, which remove significant barriers to entry for investigators and make publicly available methods for this important application domain.

2 Methods

We address two problems as shown in Figure 1: (i) predicting backbone distance matrices and torsion angles of backbone atoms from amino acid sequences, Q8 secondary structure, PSSMs, and co-evolutionary multiple sequence alignment; and (ii) reconstruction of all atom coordinates from the predicted distance matrices and torsion angles.

2.1 Predicting Backbone Distance Matrices and Torsion Angles

Our inputs are AA sequences, Q8 sequences, PSSMs, and MSA covariance matrices. Our outputs are backbone distance matrices and torsion angles.



(a) (b) (c)

Figure 1: Visualization of CASP13 results: top row denotes protein 5Z82 and bottom row protein 6HRH. Columns denote (a) ground truth and (b) predicted C_α distance matrices; (c) native (green) and predicted (cyan) structures using our model.

Feature	**Notation**	**Source**	**Dims**	**IO**	**Type**
AA Sequence	x_1	PDB	$n \times 1$	Input	21 chars
PSSM	x_2	AA/HHBlits	$n \times 21$	Input	Real [0,1]
MSA covariance	x_3	AA/jackHMMER	$n \times n$	Input	Real [0,1]
Secondary Structure (SS)	x_4	DSSP	$n \times 8$	Input	8 chars
C_α Distance Matrix	$g_1(\cdot)$	XYZ	$n \times n$	Output	Ångstrom
Torsion Angles (ϕ, ψ)	$g_2(\cdot)$	DSSP	$n \times 2$	Output	Radians

Table 1: Our CUProtein dataset consists of amino acid sequences, secondary structure, PSSM's, MSA covariance matrices, backbone distance matrices, and torsion angles.

Embeddings. We begin with a one-hot representation of each amino acid and secondary structure sequence, and real valued PSSM's and MSA covariance matrices. These are passed through *embedding layers* and then onto an encoder-decoder architecture. To leverage sequence homology, we compute the covariance matrix of the MSA features by embedding the homology information along a k dimensional vector to form a 3-tensor and contract the tensor A_{ijk} along the k dimensional embedding which is then passed as input to the encoder:

$$\Sigma = A^k_{ji} A_{ijk}, \tag{1}$$

Encoder-Decoder architectures We use encoder-decoder models with a bottleneck to train prediction models. The encoder f receives as input the aggregation A (by concatenation) of the embeddings e_i of each input x_i, and two separate decoders g_1 and g_2 that output distance matrices and torsion angles for $i \in \{1, \ldots, 4\}$. In addition, we also use a model which consists of separate encoders f_i for each embedded input $e_i(x_i)$, that are aggregated by concatenation after encoding, and separate decoders g_1 and g_2 for distance matrices and torsion angles.

$$g_j\left(f\left(\underset{x_i \in \mathcal{X}}{\text{Agg}}\left(e_i(x_i)\right)\right)\right), \quad g_j\left(\underset{x_i \in \mathcal{X}}{\text{Agg}}\left(f_i(e_i(x_1))\right)\right). \tag{2}$$

Using a separate encoder model involves a larger number of trainable parameters. Our models differ in the use of embeddings for the input, their models, and loss functions.

Deep learning models Building on techniques commonly used in natural language processing (NLP) our models use embeddings and a sequence of bi-directional GRU's and LSTM's with skip connections, and include batch normalization, dropout, and dense layers. We experimented with various loss functions including MAE, MSE, Frobenius norm, and distance logarithm to handle the dynamic range of distances. To learn the loss function we have also implemented distance matrix prediction using conditional GAN's and VAE's for protein sub-sequences.

2.2 Reconstructing Proteins

Once backbone distance matrices and torsion angle are predicted we address two reconstruction sub-problems: (i) reconstructing the protein backbone coordinates from their distance matrices, and (ii) reconstructing the full atom protein coordinates from the C_α or C_β coordinates and torsion angles.

Backbone coordinate reconstruction. We employ three different techniques for reconstructing the 3D coordinates X between backbone atoms of a protein from the predicted matrix of their pairwise distances [6]. Given a predicted distance matrix $\hat{D}$ our goal is to recover 3D coordinates $\hat{X}$ of n points. We notice that $D(X)$ depends only on the Gram matrix X^TX:

$$D(X) = \mathbf{1}\text{diag}(X^TX)^T - 2X^TX + \text{diag}(X^TX)\mathbf{1}^T, \tag{3}$$

Multi-dimensional scaling (MDS):

$$\underset{\hat{X}}{\text{minimize}} \left\| D(\hat{X}) - \hat{D} \right\|_F^2. \tag{4}$$

Semi-definite programming (SDP) and relaxation (SDR):

$$\underset{G}{\text{minimize}} \left\| K(G) - \hat{D} \right\|_F^2 \qquad \text{s.t.} \quad G \in C, \tag{5}$$

where $K(G) = \mathbf{1}\text{diag}(G)^T - 2G + \text{diag}(G)\mathbf{1}^T$ operates on the Gram matrix G.

Alternating direction method of multipliers (ADMM) [3]:

$$\underset{G,Z,\eta}{\text{minimize}} \ \lambda \left\|\eta\right\|_1 + \frac{1}{2}\left(\sum_{i,j=1}^{n} (G_{ii} + G_{jj} - 2G_{ij} + \eta_{ij} - \hat{D}_{ij}^2) \right)^2 + \mathbb{1}\{Z \in S_+^n\} \ \text{s.t.} \ G - Z = 0. \tag{6}$$

We have found multi-dimensional scaling to be the fastest and most robust method of the three, without depending on algorithm hyper-parameters, which is most suitable for our purposes. To handle the chirality of these coordinates we calculate all alpha-torsion angles (the torsion angles defined by four consecutive C_α atoms) and compare their distribution to the characteristic, highly non-symmetric distribution of protein alpha-torsions. If the predicted alpha-torsions are unlikely to result from the distribution of the protein dataset, we invert the coordinates chirality.

Full atom coordinate reconstruction. Given backbone coordinates we assign plausible coordinates to the rest of the protein's atoms. We begin with an initial guess or prediction for ϕ and ψ torsion angles. We maintain a look-up table of mean ϕ, ψ values for each combination of two consecutive α torsions and three α-angles (the angles defined by three consecutive atoms). Using these values and the C_α positions we generate an initial model. This model is then relaxed by a series of energy minimization simulations under an energy function that includes: standard bonded terms (bond, angle, plane and out-of-plane), knowledge based Ramachandran and pairwise terms, torsion constraints on the ϕ and ψ angles, and tether constraints on the C_α position. The latter term reduces the perturbations of the initial high forces. Finally, we add side-chains using a rotamer library, and remove clashes by a series of energy minimization simulations. We develop a very similar method for reconstructing the full atom protein from C_β atom distance matrices.

3 Results

We have compared our predictions on CASP12 and CASP13 [1] targets. Deep learning methods were widely used starting from CASP13, significantly improving performance over CASP12, due

PDB	Tar-id	Best RMSD	A7D	Ours
5Z82	T0951	1.01 (Seok)	NA	1.93
6D2V	T0965	1.72 (A7D)	1.72	2.08
6QFJ	T0967	1.13 (BAKER)	NA	1.33
6CCI	T0969	1.96 (Zhang)	2.27	2.53
6HRH	T1003	0.88 (MULTICOM)	2.12	2.95
6QEK	T1006	0.58 (YASARA)	0.78	1.17
6N91	T1018	1.24 (Wallner)	1.77	3.90
6M9T	T1011-D1	1.58 (A7D)	1.58	1.64

Table 2: Comparison between winning CASP13 model, best AlphaFold (A7D) model, and our model.

to availability of deep learning frameworks. Table 2 provides a sample comparison between the winning CASP13 competition models, AlphaFold models and our models. We measure the sequence independent RMSD which is consistent with the CASP13 evaluation reports and matches the deposited structures and our predictions. CASP competitors such as AlphaFold provide predictions for selected proteins. Overall, our performance on CASP13 is highly competitive. Training our models on a cloud instance takes two days using GPUs. Prediction of backbone distance matrices and torsion angles takes a few seconds per protein, and reconstruction of full atom proteins from distance matrices and torsion angles takes a few minutes per protein, depending on protein length. Limitations of this work are: (i) we only handle single domain proteins and not complexes with multiple chains, (ii) PSSM and MSA data for several of the CASP targets are limited to a subsequence of the full length protein, and (iii) we do not use available methods for detecting and reconstructing beta-sheets.

References

[1] Luciano A Abriata, Giorgio E Tamò, and Matteo Dal Peraro. A further leap of improvement in tertiary structure prediction in casp13 prompts new routes for future assessments. *Proteins: Structure, Function, and Bioinformatics*, 2019.

[2] Mohammed AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC bioinformatics*, 20(1):311, 2019.

[3] Namrata Anand and Possu Huang. Generative modeling for protein structures. In *Advances in Neural Information Processing Systems*, pages 7494–7505, 2018.

[4] Protein Structure Prediction Center. 13th community wide experiment on the critical assessment of techniques for protein structure prediction, 2018.

[5] Google DeepMind. AlphaFold: De novo structure prediction with deep-learning based scoring, 2018.

[6] Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

[7] Iddo Drori, Isht Dwivedi, Pranav Shrestha, Jeffrey Wan, Yueqi Wang, Yunchu He, Anthony Mazza, Hugh Krogh-Freeman, Dimitri Leggas, Kendal Sandridge, et al. High quality prediction of protein q8 secondary structure by diverse neural network architectures. *NeurIPS Workshop on Machine Learning for Molecules and Materials*, 2018.

[8] Z. Feng G. Gilliland T.N. Bhat H. Weissig I.N. Shindyalov P.E. Bourne H.M. Berman, J. Westbrook. The protein data bank. *Nucleic Acids Research*, 28, 2000.

[9] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[10] Stefan M Larson, Christopher D Snow, Michael Shirts, and Vijay S Pande. Folding@ home and genome@ home: Using distributed computing to tackle previously intractable problems in computational biology. *arXiv preprint arXiv:0901.0866*, 2009.

[11] Cyrus Levinthal. How to fold graciously. In *Proceedings of a meeting held at Allerton House. P. Debrunner, JCM Tsibris, and E. Munck, editors. University of Illinois Press, Urbana, IL*, 1969.

[12] Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[13] David E. Shaw, Ron O. Dror, John K. Salmon, J. P. Grossman, Kenneth M. Mackenzie, Joseph A. Bank, Cliff Young, Martin M. Deneroff, Brannon Batson, Kevin J. Bowers, Edmond Chow, Michael P. Eastwood, Douglas J. Ierardi, John L. Klepeis, Jeffrey S. Kuskin, Richard H. Larson, Kresten Lindorff-Larsen, Paul Maragakis, Mark A. Moraes, Stefano Piana, Yibing Shan, and Brian Towles. Millisecond-scale Molecular Dynamics Simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 39:1–39:11, 2009.

[14] Jingxue Wang, Huali Cao, John ZH Zhang, and Yifei Qi. Computational protein design with deep learning neural networks. *Scientific reports*, 8(1):6349, 2018.

[15] Jinbo Xu. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34):16856–16865, 2019.