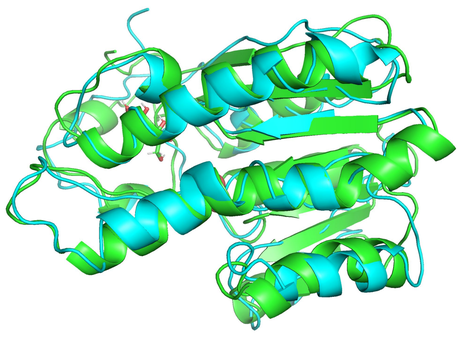
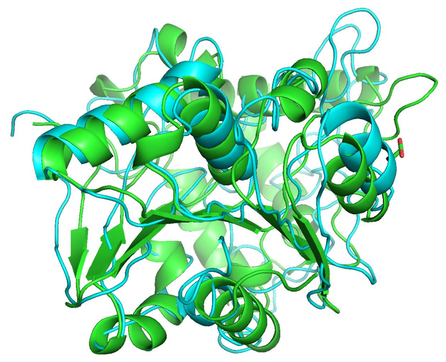
Accurate Protein Structure Prediction by Embeddings and Deep Learning Representations

Iddo Drori, Darshan Thaker, Arjun Srivatsa, Daniel Jeong, Yueqi Wang, Linyong Nan, Fan Wu, Dimitri Leggas, Jinhao Lei, Weiyi Lu
Weilong Fu, Yuan Gao, Sashank Karri, Anand Kannan, Antonio Khalil Moretti, Mohammed AlQuraishi, Chen Keasar, Itsik Pe'er

Columbia University in the City of New York
Cornell University
Harvard Medical School
Ben-Gurion University of the Negev



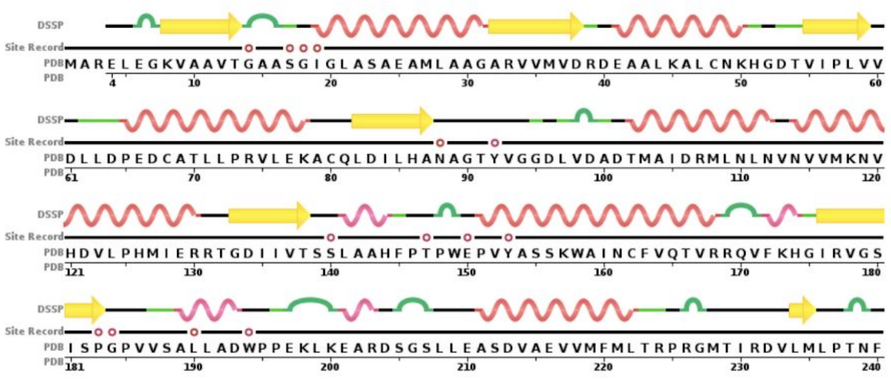


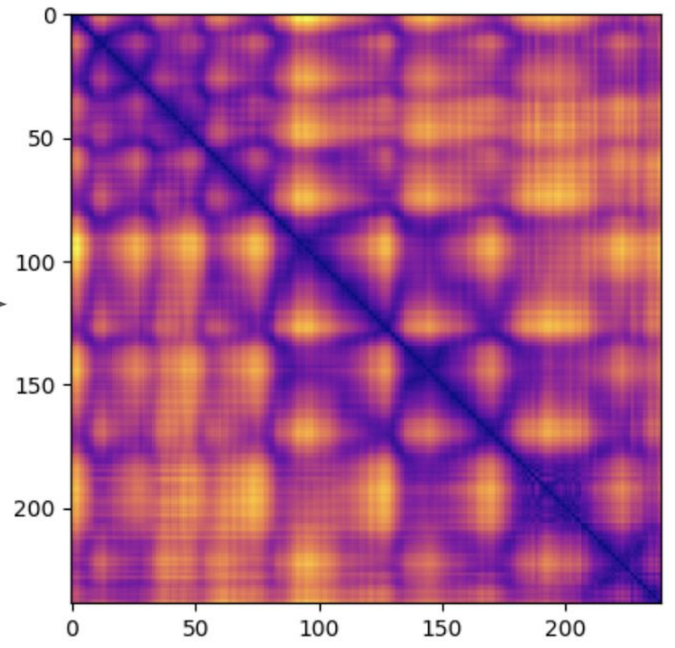
Data, Models, and Code: github.com/idrori/cu-tsp

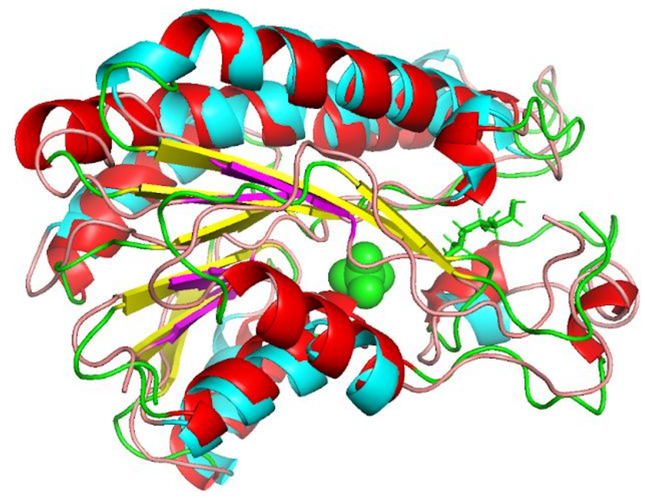
Introduction and Methods

Protein structure prediction (PSP) from amino acid sequences is a fundamental problem in computational biology.
We use 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 proteins by optimization.
Key contributions:

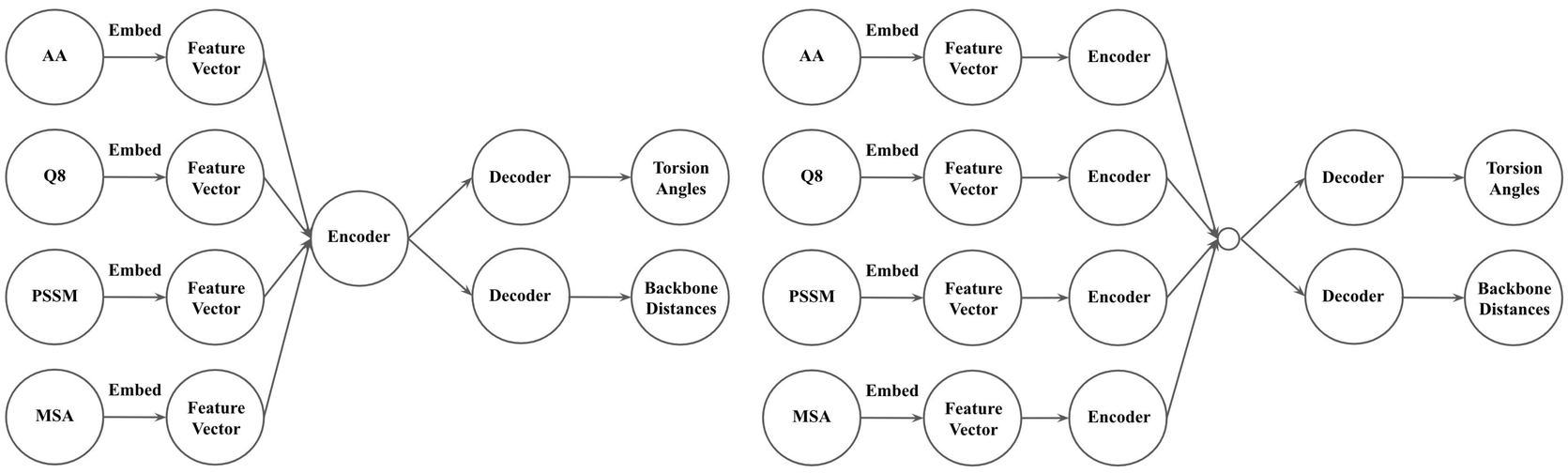
- Gold standard dataset of around 75k proteins which is easy to use in developing deep learning models for PSP.
- Competitive results with the winning teams on Critical Assessment of Techniques for Protein Structure Prediction (CASP13) and a comparison with AlphaFold (A7D), results mostly superseding winning teams (CASP12).
- Publicly available source code for both protein structure prediction using deep learning models and protein reconstruction.







High level flow: Our method operates in three stages by i) predicting backbone atom distance matrices and torsion angles; ii) recovering backbone atom 3D coordinates; and iii) reconstructing the full atom protein by optimization



Feature	Source
AA Sequence	PDB
PSSM	AA/HHBlits
MSA covariance	AA/jackHMMER
Secondary Structure (SS)	DSSP
C_α, C_β Distance Matrices	PDB
Torsion Angles (ϕ, ψ)	DSSP

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

Model architecture: inputs are embedded followed by (left) aggregation, encoding using sequence models, and decoding; (right) encoding using sequence models, aggregation,and decoding.

Results

PDB	CASP	Tar-id	Best RMSD	A7D	Ours
5Z82	13	T0951	1.01 (Seok)	NA	1.79
6D2V	13	T0965	1.72 (A7D)	1.72	**1.60**
6QFJ	13	T0967	1.13 (BAKER)	NA	1.18
6CCI	13	T0969	1.96 (Zhang)	2.27	2.53
6HRH	13	T1003	0.88 (MULTICOM)	2.12	2.95
6QEK	13	T1006	0.58 (YASARA)	0.78	1.02
6N91	13	T1018	1.24 (Wallner)	1.77	3.89
6M9T	13	T1011-D1	1.58 (A7D)	1.58	1.64
5J5V	12	T0861	0.49 (MULTICOM)	NA	1.00
2N64	12	T0865	1.87 (HHPred)	NA	**1.58**
5JMU	12	T0879	1.35 (MULTICOM)	NA	**1.31**
5JO9	12	T0889	1.31 (Seok)	NA	1.79
4YMP	12	T0891	1.10 (GOAL)	NA	1.36
5XI8	12	T0942-D2	1.73 (EdaRose)	NA	**1.60**

Representative comparison between the winning CASP12/13 models for each protein, bestAlphaFold (A7D) model for CASP13, and our model for CASP12/13.

Average length of CASP13 target proteins selected is 325 residues, average for CASP12 is 247 residues. Results of RMSD around 2 Angstrom on test targets are considered accurate in CASP.

Our results supersede winning teams of CASP12 compared with each best team for most individual proteins which highlights the improvement of using deep learning methods. Our approach is on par with winning teams in CASP13, compared with winning teams for individual proteins, which highlights that our methods are state-of-the-art and overall our performance on CASP is highly competitive.

Acknowledgements: We thank the 100 Columbia University graduate students of the Spring 2019 Deep Learning course for their participation in an in-class protein tertiary structure prediction competition.