



The protein structure prediction problem is a long-standing challenge within the scientiﬁc community with the potential to revolutionize the healthcare and drug development industries. Structural knowledge on target proteins allows scientists the capability to evaluate how a particular protein may interact with drug candidates and other molecules. Various methods have been employed in pursuance of generating probable native structures given available amino acid sequences with varying results. Here, we present an approach that employs deep learning methodology alongside statistical physics with the aim to further protein structure predictive capabilities. Sequence-based deep learning algorithms o er powerful and accurate predictions for localized structural attributes, such as bond angles and atomic distances, while statistical physics allows for large-scale conformation generation and subsequent samplings, evaluations and optimizations that result in viable, structurally sound predicted native protein folds. To maintain consideration for biological processes, the sequential procedures utilized in this approach were deliberately designed to resemble processes that take place under protein folding theory. Results indicate that initial deep learning predictive processes followed by the exploration of a broad conformational space may be utilized to generate viable native protein structures for scientiﬁc research applications.



Proteins are complex molecules which are an essential part of all living organisms. They originate from the synthesis of sequences of amino acids which then interact with one another and the surrounding environment to ultimately transition into their native, lowest energy structural conﬁgurations. Proteins are of main concern in drug design as many are linked to a variety of disease processes. In the case of Covid-19, for example, the spike proteins on the surface of the virus are able to bind to and subsequently hijack a human’s healthy cells, thereby resulting in illness. Knowledge of protein sequences and their structures allows for scientists to explore ways to prevent or interrupt this kind of disease progression.

Scientists have long known how to assess protein amino acid sequences through techniques such as mass spectrometry, however, the ability to evaluate the structure of proteins is a costly, long and arduous process. It is for this reason that availability of amino acid sequences far outweighs the availability of structural data on native proteins. The pursuit of predicting these native structures, then, has signiﬁcant implications, as it would bypass the need for immense resources, time and manpower in order to evaluate structures for innumerous amounts of proteins available in nature. If accurate algorithms could be developed in this pursuit, then

efficacious treatments may be more readily designed and structural evaluation e orts may be focused on only those proteins whose amino acid sequences are signiﬁcantly distinct or complex.

The introduction of Artiﬁcial Intelligence, and speciﬁcally deep learning, came with the realization that accurate predictions for native folds would eventually be possible. The major premise of protein structure prediction is that the conformation of a native protein structure may be predicted given solely the string representation of its amino acid sequence. Over time, the ﬁeld has evolved from simple algorithms, mainly predicting general details of amino acid conﬁgurations, to more sophisticated predictive processes that could surprisingly come close to predicting certain native structures. Some of the most powerful algorithms to date have focused on protein families and their evolution. Similar to most biological organisms, proteins are also subject to evolutionary processes and maintain structural characteristics passed down from their corresponding ancestors. Taking this theory into account, it would follow that we may rely, at least in part, on any known structures that may be related, and use that information to guide ﬁnal native structure predictions for a protein of unknown structure. The activeness, however, would rely on the availability of structural data for relevant ancestral proteins and the ability to accurately predict regions which may not have relational templates. This complication is just one of the motivations for evaluating options that may not be so reliant on relational data as it would lead to better overall reliability and generalization in predicting native folds.

Here, we propose a system of predictive processes that integrates deep learning methodologies with statistical physics to provide an elegant solution to the shortcomings of relation-based protein structure prediction. Deliberate deep learning algorithms may be relied upon for general predictions on protein structure and amino acid arrangements while statistical processes may be utilized to evaluate immense conformational space and select highly probable conﬁgurations for ﬁne-tuning. Furthermore, data engineering and selection measures may be utilized to carry out prediction processes considering only the most reliable and relevant features available.



# Data Preparation

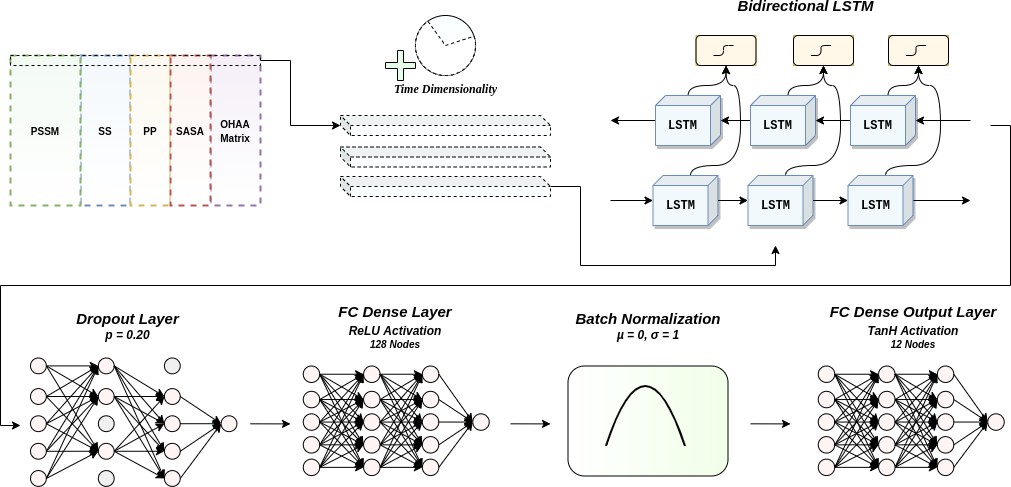
Datasets for prediction processes were prepared by parsing 150,000+ protein structure ﬁles downloaded and stored locally from Protein Data Bank (PDB). An internally developed script was developed to parse and evaluate structures based on completeness, experimental procedure, resolution, R-value and number of outliers in terms of interatomic distances and bond angles. From the ﬁnal viable structures, training data was prepared through the generation of an amino acid matrix, PSSM from PSI-BLAST1, Secondary Structure probabilities from SPOT-1D2, Solvent Accessible Surface Area (SASA) values from VMD3, and physical properties4. SASA values for protein chains with partial transmembrane regions were updated by membrane topology predictions generated from TOPCONS5. Finally, angle and distance targets were calculated from the coordinates of backbone atoms available in PDB ﬁles.

# Baseline Structure Prediction

A Bidirectional LSTM based deep neural network was developed in order to accurately predict structural elements of an unknown target protein. To increase accuracy for prediction, the base model was trained on a dataset of protein chains with highest sequence similarity to the target. Targets that were evaluated to be membrane proteins were predicted from a model trained strictly on a dataset containing only membrane protein chains. All datasets were constructed from clean, high resolution, non-homologous protein chains so as to further improve predictive capabilities. Final target predictions provided an initial baseline structure from which further processing steps were applied.



Figure 1



Deep learning-based algorithm architecture for baseline structure prediction.

# Protein Folding Simulation

For unknown structures, phi and psi dihedral angles were obtained from the prediction realized by the Bidirectional LSTM deep neural network. In addition, omega angle values were obtained through random generation based on a probability distribution from the whole population of protein structures, simulating a variation of the Random Energy Model (REM)6 in protein folding theory. The random omega generation process was utilized to simulate realistic folding intermediates with a conformational sampling space of 107 unique folds. Also, to account for inherent error in the LSTM prediction of dihedral angles phi and psi, random noise was added to evaluate the possibility for small augmentations that may have resulted in a structure with lower potential energy. Discrete conformational modiﬁcations also served to overcome potential energy barriers between local minima and move further toward the native fold. The angle-based NERF7 (Natural Extension Reference Frame) algorithm was used to

realize the Cartesian coordinates of fold samples in phi, psi, omega dihedral space for structural evaluation. Potential energy, radius of gyration and van der Waals radii were used as ﬁltering criteria. Viable structures left over from the ﬁltering step were then subjected to clustering via MUFold-CL8 for more isolated comparisons. Folds with perceived minimal potential energy values with the highest structural integrity were selected as prospects for ﬁne tuning and evaluation.

# Fine-Tuning, Minimizing Potential

Side chains for all residues in each remaining candidate fold were generated by the psfgen tool which utilizes the CHARMM369 force ﬁeld for accurate structure creation. Structures were ranked based on an orientation-dependent atomic potential calculated using calRW+10. The highest ranking folds were selected for structural optimization through Molecular Dynamics (MD) simulation. MD simulation was used in order to further minimize potential energy through the calculation of interatomic forces with the added presence and inﬂuence of water molecules using the NAMD11 package. The application of either explicit or implicit solvent calculations was decided taking into account chain size and computational capacity. Final results from molecular simulation provided further insights into the quality of each fold candidate.

# Distance Based Corroboration

Final folded structures were evaluated through a conﬁrmation process based on corroborating candidate interatomic distance values with predicted interatomic distances from DeepMetaPSICOV12. In this step, discernment based on intuition of inter-residue interactions and experience in protein folding theory from human intervention was utilized.

# Template Based Modeling

In the case that a prediction target’s amino acid sequence achieved at least 80% similarity with a sequence of a protein with known structure, the template based modeling algorithm HHpred13 was used. When signiﬁcant, independent regions of a protein chain were unmatched or evaluated to have structures that did not corroborate with expert intuition, the LSTM based structural prediction process was utilized to predict these localized structures.



1. Altschul, S. F., Madden, T. L., Schä er, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.
2. Hanson, J., Paliwal, K., Litﬁn, T., Yang, Y., & Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, *35*(14), 2403-2410.



3.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-38.



4.

Zhou, Y., & Faraggi, E. (2010). Prediction of one-dimensional structural properties of proteins by integrated neural networks. *Introduction to protein structure prediction: methods and algorithms*, 45-74.



5.

Tsirigos, K. D., Peters, C., Shu, N., Käll, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic acids research*, *43*(W1), W401-W407.



6.

Bryngelson, J. D. and Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding), *J. Phys. Chem.* 93 (19), 6902–6915.

7. Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., & Strauss, C. E. (2005). Practical conversion from torsion space to Cartesian space for in silico protein synthesis. *Journal of computational chemistry*, *26*(10), 1063-1068.



8.

Zhang, J., & Xu, D. (2013). Fast algorithm for population‐based protein structural model analysis. *Proteomics*, *13*(2), 221-229.



9.

Huang, J., & MacKerell Jr, A. D. (2013). CHARMM36 all‐atom additive protein force ﬁeld: Validation based on comparison to NMR data. *Journal of computational chemistry*, *34*(25), 2135-2145.

1. Zhang, J., & Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one*, *5*(10), e15386.
2. Phillips, J. C., Hardy, D. J., Maia, J. D., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., ... & McGreevy, R. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*, *153*(4), 044130.
3. Kandathil, S. M., Greener, J. G., & Jones, D. T. (2019). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1092-1099.



13.

Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*. W244–W248.