Residue characterization on AlphaFold2 protein structures using graph neural networks

Nasim Abdollahi[*]
Faculty of Medicine, University of Toronto
Cyclica Inc.
Toronto, ON, Canada
nasim.abdollahi@utoronto.ca

Ali Madani
Cyclica Inc.
Toronto, ON, Canada
ali.madani@cyclicarx.com

Bo Wang[†]
Faculty of Medicine, University of Toronto
Vector Institute
Toronto, ON, Canada
bowang@vectorinstitute.ai

Stephen MacKinnon[‡]
Cyclica Inc.
Toronto, ON, Canada
stephen.mackinnon@cyclicarx.com

Abstract

Three-dimensional structure prediction tools offer a rapid means to approximate the topology of a protein structure for any protein sequence. Recent progress in deep learning-based structure prediction has led to highly accurate predictions that have recently been used to systematically predict 20 whole proteomes by DeepMind's AlphaFold and the EMBL-EBI. While highly convenient, structure prediction tools lack much of the functional context presented by experimental studies, such as binding sites or post-translational modifications. Here, we introduce a machine learning framework to rapidly model any residue-based classification using AlphaFold2 structure-augmented protein representations. Specifically, graphs describing the 3D structure of each protein in the AlphaFold2 human proteome are generated and used as input representations to a Graph Convolutional Network (GCN), which annotates specific regions of interest based on the structural attributes of the amino acid residues, including their local neighbors. We demonstrate the approach using six varied amino acid classification tasks.

1 Introduction

The introduction of deep learning to three-dimensional (3D) protein structure prediction problems has led to a sudden leap in predictive performance as reported in the 2020 protein structure prediction competition, CASP14 (1; 2; 3). Reliable structure prediction can be applied at scale to model full proteomes. In July 2021, DeepMind and the EMBL-EBI announced a partnership aimed at modeling the 3D structures of 100 million sequenced proteins, including an initial release of 20 complete proteomes, including humans (4; 5). In contrast, protein structures determined by experimental means can take months to years to solve. As of September 18th, 2021 the Protein DataBank (PDB) lists 182,176 experimentally determined protein structures (6). Expanding the scope of high-accuracy structure models beyond homology-based approaches could enable functional characterization and drug design applications for countless new systems on demand. For instance, emerging pathogens

[*] nasim.abdollahi@cyclicarx.com

[†] https://wanglab.ml/index.html, https://vectorinstitute.ai/team/bo-wang/, bo.wang@uhnresearch.ca

[‡] https://www.cyclicarx.com/

Machine Learning for Structural Biology Workshop, NeurIPS 2021.

could be targeted for pharmaceutical research programs very rapidly. Typically, genomic sequences of new pathogens are available within weeks of new outbreaks and reliable predicted structure models are rapidly generated through online servers shortly thereafter. Following publication of the novel coronavirus genome in January 2020, reliable 3D protein structure models for the novel coronavirus proteins were automatically generated within days by the top modelling groups globally, including SwissModel (7; 8), ZhangLab (9; 10), AlphaFold (11), and Rosetta (12).

Despite the large recent gains in structure prediction accuracy, using predicted structures effectively for pharmaceutical applications remains a challenge. Modeled protein structures lack valuable structural annotations which are often available through experimental means, such as ligand binding sites, post translational modifications, macromolecular binding surfaces, metal binding sites, solvent binding sites, etc. An automated framework to reliably predict druggable surfaces, surfaces likely to be buried by protein interactions, or other specific regions of functional interest will accelerate future rapid-response structure-based drug design and/or repurposing efforts.

This paper proposes a deep learning model that annotates 3D protein structures with predicted ligand-, DNA-, RNA-, peptide-, or protein- binding sites using Graph Convolutional Networks (GCN). It is a generalized framework to perform any arbitrary protein's amino acid residue (AAR) classification task, using the AlphaFold2 predicted 3D structure representation of a protein and a GCN model. GCNs, which have seen a considerable interest in the last few years, are applicable to the wide range of prediction problems with structure models in different domains including biology (13; 14; 15; 16; 17), chemistry (18; 19), physics (20; 21), natural language processing (22; 23), and social sciences (24; 25; 26). As proteins perform their function through a complex network of AARs, the network structure can naturally be modeled as graphs (27). The graph-based convolutional neural networks are more efficient compared with Convolutional Neural Networks (CNNs) for protein graph-based data representation, especially when working with large-scale datasets as computational cost and memory requirements are relatively insignificant (27).

The GCN technique proposed here is a unified framework that readily models any residue-based dataset and has the flexibility of being rapidly deployed to multiple new problems for predicting any residue-based annotations. While these specific predictions are not new, they are typically performed by individual expert-designed technologies for each separate task, such as reported studies in (28; 29; 30; 31; 32; 33; 34). Modeling multiple tasks with a single, unified framework allows them to simultaneously benefit in parallel from subsequent iterative improvements, such as improved features, model architectures, or innovations in test/train splits. Moreover, unlike predictive engines based on primary sequence alone, the proposed graph-based methodology introduces spatial and local environment context to all predictions, without compromising the scope of the model as a consequence of poor structural coverage of proteins in the dataset.

2 Methodology

Proteins are chains of amino acid residues that fold into a 3D structure that gives them their biochemical functions. The goal of the proposed unified GCN framework is to accept an input protein structure exclusively made up of residue coordinates and annotate specific regions of interest based on the structural attributes of the residue itself and its local neighbors. Fig. 1 provides a schematic representation of the unified GCN framework architecture for residue characterization on AlphaFold2 protein structures as the network input. In the following sections we explain the details of graph data preparation and the GCN framework.

2.1 Graph Representation of Proteins

Protein residue graphs are constructed from each predicted protein structure in the AlphaFold2 human proteome, where nodes correspond to individual residues and edges correspond to inter-residue contacts within pre-set distances.

Constructing Node Features The node features represent the known properties of the residue. Residues (nodes) are annotated with 26 features derived directly from the 3D structure model file, including amino acid (20 features via one-hot encoding), backbone angles (phi, psi, tau, and theta), solvent accessibility of the residue's backbone atoms and solvent accessibility of the residue's side chain atoms. PDB file processing and SASA calculations are performed by Biopython (35; 36)

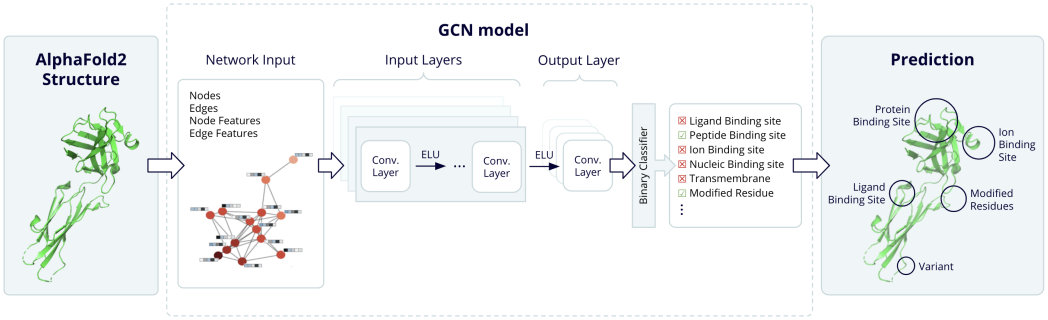


Figure 1: Schematic representation of GCN framework to embed representation of an AlphaFold2 predicted protein structure and perform residue classification tasks.

and FreeSASA (37) respectively. Features are intentionally restricted to residue properties derived exclusively from the structure file to ensure that the GCN model can be applied to future protein structures with a single model file input. Moreover, only structure files derived from the AlphaFold2 database are used for training to avoid source bias.

Constructing Edge Features The neighborhood of a node used in the convolution operator is the set of closest residues as determined by the threshold distance between their c-alpha atoms, considered as the center of the residues. The spatial relationships between residues are represented as features of the edges that connect them. The weighted edge features calculated as the inverse of the square of euclidean distance are found to have significant impact on improving the GCN performance.

Constructing Residue Annotations (Prediction Targets) Residue annotations used as training labels and corresponding to prediction targets can be obtained from any residue-based dataset, mappable to the reference sequences, even if they are not derived from structural studies. For instance, UniProt annotations for 'modified residues' corresponding to post-translational modifications often originate from mass-spectroscopy studies that do not involve solved 3D structure. To demonstrate the utility of this method, we build models for multiple different residue annotations from UniProt including post-translational modifications derived from functional studies, as well as binding site annotations from BioLip, which indexes observed binding sites from macromolecular structures in the PDB (38). Small molecule ligands, nucleic acid, peptide, and inorganic molecule binding sites from BioLip are mapped onto UniProt sequences then used as classification labels. Models lacking any binding annotations are not used for training, as to prevent excessive false negative labels from biasing the dataset.

2.2 Graph Convolutional Network

The GCN model is built using the PyTorch Geometric library (39) to predict labels (annotations) of each node (residue) in the protein graphs. The GCN prediction process is conducted through exploitation of structure features of the target residue and its neighbors in the protein graph. It is performed by applying multiple graph convolutional blocks of different sizes which are connected by ELU activation function (40). The convolutional block in the output layer is followed by a final block of log softmax function (41), which calculates the logarithmic probability of the target of interest being a positive label, after a 50% dropout is applied to prevent overfitting. For the results presented in Section 4, the `GCNConv` convolutional operator is used (42). The Adam optimizer with learning rate of 0.01 is chosen to minimize training loss where the loss is calculated with torch negative log likelihood loss function (43). The GCN framework has the flexibility of performing single-task as well as multi-task learning. The framework can also handle training large datasets by clustering the graphs, which is performed by grouping the proteins to ensure the entire protein graph stays within one cluster.

3 Experiments

For our training experiment only one 3D structure model per protein is included in the dataset. Also, models lacking any binding annotations are not used for training to prevent excessive false negative labels from biasing the dataset, which corresponds to 3630 total proteins consisting of 1,749,863 residues (after filtering missing values). The 3630 total proteins are randomly splitted into 3160 training proteins and 5×100 validation proteins. In order to evaluate the GCN framework, for each train and validation set four separate graphs are built with different distance threshold cutoffs of 5Å, 8Å, 10Å, and 15Å for defining the residues (nodes) connections. For prediction targets that are residue labels in the protein graphs, six residue classification tasks are considered. These tasks include post-translational modifications from UniProt and small molecule binding sites from BioLip, ligands, nucleic acid, peptide and inorganic, that are mapped onto UniProt sequences. For all the experiments performed, the GCN layer structure consists of: four network layers of sizes $26, 16, 16, 8$, and one output layer of size 2.

4 Results and Discussion

In this section, we demonstrate capabilities of our GCN framework for residue characterization across multiple tasks explained in Section 3. We trained the network for 10,000 epochs for all the tasks. After each epoch, to quantify the network performance, the area-under receiver operating characteristic curve (ROC AUC) is averaged on five validation sets for each prediction experiment. These curves suggest about $75\% - 100\%$ for ROC AUC, which are representative of a high-quality residue characterization model (Fig. 2). Transmembrane residue and peptide binding site predictions resulted in the highest and lowest performances of 0.996 and 0.754, respectively, across the tasks. Prediction of transmembrane regions from a 3D structure is nearly trivial given the consistent structural topologies and residue properties in membrane spanning regions. The task was intended as an easy positive control to help in the design of the model. Peptide-binding is considerably more difficult as the structural determinants that differentiate peptide interactions from other ubiquitous forms of protein self-interaction, homomeric or heteromeric interactions may be very subtle.

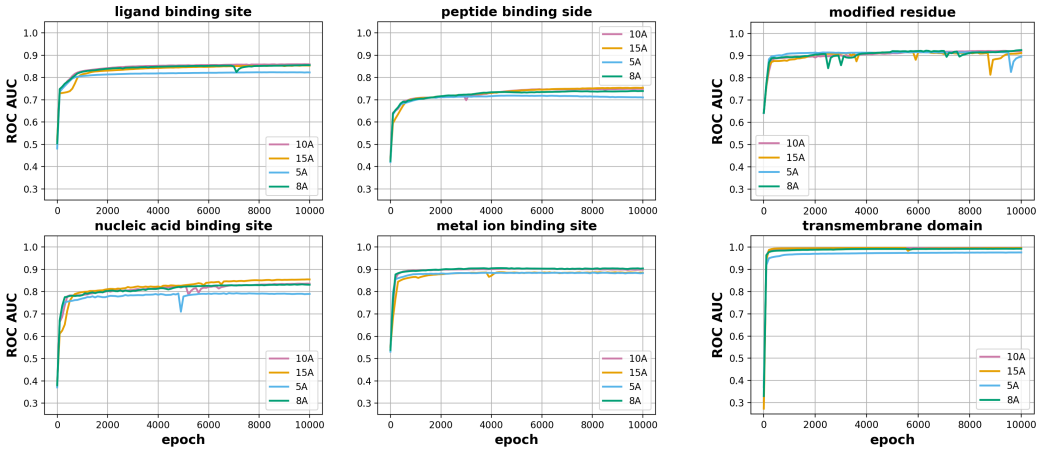
It is observed that changing distance cutoff in graph generation affects the performance across all prediction tasks; however, the effect is not similar. For example, distance cutoff has the maximum effect on performance of nucleic acid binding site prediction while the minimum effect on modified residue prediction task with maximum ROC AUC difference of 0.062 and 0.009, respectively. In the case of nucleic acid binding, the task-dependent preference for larger distance contact networks cutoffs may be a sign that the model is recognizing long-range bulk electrostatic effects. However, more task-dependent preferences and exploratory ablation studies are required before we can confidently link model performance observations with biophysical rationale.

Conclusion

We introduced a GCN framework for residue characterization with AlphaFold2 protein structure as input to the network. The proposed methodology constructs protein graph representations and exploits the structure features of target residue and its spatially adjacent residues in the protein graph by applying multiple layers of graph-based convolutions. It has significant advantages over residue task-specific prediction models by providing a unified framework that can be used for any new residue-based annotation prediction, where it simultaneously benefits from iterative network refinements and inter-task comparisons.

Broader Impact

This study is an active work in progress, requiring additional comparative studies, algorithmic improvements and task analyses. Once complete, it will be documented in an equitable, open access medium and the codebase will be released to the greater community for subsequent academic research and application in the biotechnology, agricultural and pharmaceutical industries.



(a) Binding site annotations (b) UniProt annotations

Figure 2: Proof-of-concept Area Under the Receiver Operating Characteristic Curves (ROC AUC) for six residue classification tasks, each generated by four separate models at varying threshold distances for the protein contact network (5Å, 8Å, 10Å, and 15Å). (a) Four classification tasks obtained by mapping residue annotations from multiple BioLip entries onto the corresponding protein dataset, providing trainable labels, including ligand, peptide, ion, and nucleic acid binding sites. (b) Two classification tasks derived from UniProt annotations, including transmembrane and modified residue.

Acknowledgments and Disclosure of Funding

Authors would like to thank Mitacs and Cyclica Inc. for funding this work.

References

[1] J. Pereira, A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan, and A. N. Lupas, "High-accuracy protein structure prediction in casp14," *Proteins: Structure, Function, and Bioinformatics*, 2021.

[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[3] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021.

[4] DeepMind and EMBL-EBI, "Alphafold protein structure database." `https://alphafold.ebi.ac.uk/`, 2021.

[5] DeepMind and EMBL-EBI, "Deepmind and embl release the most complete database of predicted 3d structures of human proteins." `https://www.ebi.ac.uk/about/news/press-releases/alphafold-database-launch`, 2021.

[6] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The protein data bank nucleic acids res 28: 235–242," 2000.

[7] S. I. of Bioinformatic, "Swiss-model. severe acute respiratory syndrome coronavirus 2." `https://swissmodel.expasy.org/repository/species/2697049b`, 2020.

[8] S. I. of Bioinformatic, "Swiss institute of bioinformatics experts and resources in the fight against coronavirus." `https://www.sib.swiss/about-sib/news/10643-sib-resources-and-tools-in-the-fight-against-coronavirus?utm_source=twitter&utm_medium=social&utm_campaign=organic&utm_content=Coronavirus`, 2020.

[9] ZhangLab, "Modeling of the sars-cov-2 genome using d-i-tasser." https://zhanggroup.org/COVID-19/, 2020.

[10] C. Zhang, W. Zheng, X. Huang, E. W. Bell, X. Zhou, and Y. Zhang, "Protein structure and sequence reanalysis of 2019-ncov genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and hiv-1," *Journal of proteome research*, vol. 19, no. 4, pp. 1351–1360, 2020.

[11] DeepMind, "Computational predictions of protein structures associated with covid-19." https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19, 2020.

[12] I. f. P. D. University of Washington, "Rosetta's role in fighting coronavirus – institute for protein design." https://www.ipd.uw.edu/2020/02/rosettas-role-in-fighting-coronavirus/, 2020.

[13] P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. Bronstein, and B. Correia, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, no. 2, pp. 184–192, 2020.

[14] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," *arXiv preprint arXiv:1509.09292*, 2015.

[15] A. M. Fout, *Protein interface prediction using graph convolutional networks*. PhD thesis, Colorado State University, 2017.

[16] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[17] N. Xu, P. Wang, L. Chen, J. Tao, and J. Zhao, "Mr-gnn: Multi-resolution and dual graph neural network for predicting structured entity interactions," *arXiv preprint arXiv:1905.09558*, 2019.

[18] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, pp. 1263–1272, PMLR, 2017.

[19] K. Do, T. Tran, and S. Venkatesh, "Graph transformation policy network for chemical reaction prediction," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 750–760, 2019.

[20] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia, "Graph networks as learnable physics engines for inference and control," in *International Conference on Machine Learning*, pp. 4470–4479, PMLR, 2018.

[21] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, and K. Kavukcuoglu, "Interaction networks for learning about objects, relations and physics," *arXiv preprint arXiv:1612.00222*, 2016.

[22] D. Beck, G. Haffari, and T. Cohn, "Graph-to-sequence learning using gated graph neural networks," *arXiv preprint arXiv:1806.09835*, 2018.

[23] L. Song, Z. Wang, M. Yu, Y. Zhang, R. Florian, and D. Gildea, "Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks," *arXiv preprint arXiv:1809.02040*, 2018.

[24] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 5165–5175, 2018.

[25] C. Li and D. Goldwasser, "Encoding social information with graph convolutional networks forpolitical perspective detection in news media," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2594–2604, 2019.

[26] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen, "Graph convolutional networks with markov random field reasoning for social spammer detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 1054–1061, 2020.

[27] R. Zamora-Resendiz and S. Crivelli, "Structural learning of proteins using graph convolutional neural networks," *bioRxiv*, p. 610444, 2019.

[28] S. Ahmad and A. Sarai, "Pssm-based prediction of dna binding sites in proteins," *BMC bioinformatics*, vol. 6, no. 1, pp. 1–6, 2005.

[29] C. Yan, M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, "Predicting dna-binding sites of proteins from amino acid sequence," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–10, 2006.

[30] L. Wang and S. J. Brown, "Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W243–W248, 2006.

[31] E. Petsalaki, A. Stark, E. García-Urdiales, and R. B. Russell, "Accurate prediction of peptide binding sites on protein surfaces," *PLoS computational biology*, vol. 5, no. 3, p. e1000335, 2009.

[32] R. Krivák and D. Hoksza, "P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure," *Journal of cheminformatics*, vol. 10, no. 1, pp. 1–12, 2018.

[33] J. S. Sodhi, K. Bryson, L. J. McGuffin, J. J. Ward, L. Wernisch, and D. T. Jones, "Predicting metal-binding site residues in low-resolution structural models," *Journal of molecular biology*, vol. 342, no. 1, pp. 307–320, 2004.

[34] N. Li, Z. Sun, and F. Jiang, "Prediction of protein-protein binding site by using core interface residue and support vector machine," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–13, 2008.

[35] T. Hamelryck and B. Manderick, "Pdb file parser and structure class implemented in python," *Bioinformatics*, vol. 19, no. 17, pp. 2308–2310, 2003.

[36] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[37] S. Mitternacht, "Freesasa: An open source c library for solvent accessible surface area calculations," *F1000Research*, vol. 5, 2016.

[38] J. Yang, A. Roy, and Y. Zhang, "Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions," *Nucleic acids research*, vol. 41, no. D1, pp. D1096–D1103, 2012.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[40] "Elu." `https://pytorch.org/docs/stable/generated/torch.nn.ELU.html`.

[41] "log-softmax." `https://pytorch.org/docs/stable/generated/torch.nn.functional.log_softmax.html`.

[42] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[43] "Nllloss." `https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html`.