Predicting cryptic pocket opening from protein structures using graph neural networks

Michael Ward*
Department of Biochemistry and Molecular Biophysics
Washington University in St. Louis
mdward@wustl.edu

Artur Meller*
Department of Biochemistry and Molecular Biophysics
Washington University in St. Louis
ameller@wustl.edu

Meghana Kshirsagar
AI for Good Research Lab
Microsoft
Meghana.Kshirsagar@microsoft.com

Felipe Oviedo Perhavec
AI for Good Research Lab
Microsoft
Felipe.Oviedo@microsoft.com

Jonathan Borowsky
Department of Biochemistry and Molecular Biophysics
Washington University in St. Louis
borowsky.jonathan@wustl.edu

Geralyn Miller
AI for Good Research Lab
Microsoft
geramill@microsoft.com

Juan Lavista Ferres
AI for Good Research Lab
Microsoft
jlavista@microsoft.com

Gregory R. Bowman
Department of Biochemistry and Molecular Biophysics
Washington University in St. Louis
g.bowman@wustl.edu

Abstract

Proteins undergo structural fluctuations *in vivo* which can lead to the formation of pockets unseen in the native, folded structural state (*i.e.* "cryptic pockets"). Inferring cryptic pockets from experimentally determined protein structures is valuable when developing a drug since ligands typically require a pocket for tight binding. Toward this end, many studies employ molecular dynamics simulations to model protein structural fluctuations, but these simulations often require 100s of GPU hours. We hypothesized that machine learning algorithms that predict sites of cryptic pockets directly from folded structures can speed this up. Here, we adapt a graph neural network architecture, which previously achieved state-of-the-art performance on protein structure learning tasks, to predict sites of cryptic pocket formation from experimental protein structures. We trained this model by re-purposing an existing molecular simulation dataset that was generated to identify cryptic pockets in SARS-CoV-2 proteins. Our model achieves good performance (AUC=0.78) on a held-out test set of protein structures with ligands bound to cryptic sites and requires $<$ 1 second of compute on a single GPU.

*Authors contributed equally

Machine Learning for Structural Biology Workshop, NeurIPS 2021.

1 Introduction

A structure of a protein's native, folded state can reveal potential drug binding sites, but leaves us blind to other potential sites that form as the protein structure fluctuates in solution. There are over 100 confirmed examples of these "other" binding sites where a small molecule binds in a pocket on a protein which was not observable from any previously determined structure of that protein (*i.e.* a "cryptic pocket")[1]. Currently, it is challenging to predict these cryptic pockets from ground state experimental structures, but the ability to do so would come with several benefits. For example, protein structures that lack any obvious binding pockets are often considered undruggable[2], but they may actually prove to be good drug targets if they have a cryptic pocket that can be targeted. Even if a protein structure already reveals a binding pocket that can be targeted with a small molecule (e.g. an active site), it is useful to know if there are cryptic pockets as targeting cryptic pockets may improve specificity (i.e. reduce off-target effects when targeting a family of homologous proteins) or lead to the discovery of allosteric activators [3],[4].

Current methods for identifying cryptic pockets in proteins are either slow or have low accuracy. Molecular dynamics simulations, which use physics-based force fields to model protein structural fluctuations [5], are the primary means to identify and sample structural configurations of cryptic pockets but they often consume 100s of GPU hours per protein. Ideally, one could employ an algorithm that quickly and accurately determines if a protein will form a cryptic pocket, then use this result to determine if resources should be deployed to run a costly simulation or an experimental drug screen. Cryptosite is one such machine learning algorithm that predicts which amino acid residues of a protein will form a cryptic pocket with good performance (AUC=0.83)[1]. However, this method takes $\sim$1 day to run because it relies on simulation data as input to the algorithm. When simulation features are removed, its performance markedly drops (AUC=0.74)[1]. In the current study, we train a graph neural network to accurately determine sites of cryptic pockets from experimental structures. In a previous study, molecular dynamics simulations were performed to identify and sample cryptic pockets across most proteins in the SARS-CoV-2 proteome and $\sim$50 new potential binding sites were uncovered [6]. Across these simulations, there are thousands of events where a cryptic pocket forms. We used these events as training examples to train a graph neural network to classify whether or not a residue is likely to participate in a cryptic pocket given the protein's 3D topology and the chemical environment of its neighborhood.

2 Predicting cryptic pockets with Geometric Vector Perceptrons

We hypothesized that the propensity of an amino acid residue to participate in a cryptic pocket is a function of the 3D topology and the chemical environment in which the amino acid resides. On one hand, if a residue is in a tightly packed region of a protein and has extremely strong attractive interactions with its neighbors, it is unlikely to undergo a structural rearrangement that creates a pocket that a ligand might bind. On the other hand, if a residue is in a loosely packed environment and has weak interactions with its neighbors, it may be more likely to be in a region that forms a cryptic pocket. Given this hypothesis, we sought to train a model that takes a protein structure as input and outputs a value that indicates the likelihood that each amino acid residue will participate in a cryptic pocket.

Previous work has established that graph neural networks are an efficient way to learn complicated tasks from 3D protein structures. Specifically, a graph neural network architecture that employs a Geometric Vector Perceptron (GVP) has been previously shown to accurately evaluate the model quality of predicted protein structures and also successfully predict feasible protein sequences from 3D structures [7]. Briefly, the GVP-based graph neural network takes a set of node (*i.e.* an amino acid residue) and edge features that describe geometric and chemical features describing a residue, i, and its relation to another residue, j (**Fig. 1a**). To learn a representation for a given residue, the network uses message passing in which messages from neighboring residues and edges are used to update the residue representation. Here, we adapt the GVP-based graph neural network to the task of predicting sites of cryptic pockets from a native, folded structure of a protein. As in the original paper, we use node features including: the type of amino acid residue, *sine* and *cosine* transformations of backbone dihedral angles, an imputed unit vector between the α and β carbon, and a forward and reverse unit vector to the $i-1$ and $i+1$ neighboring residues. Edge features include a unit vector between nodes, a distance between nodes, and a *sine* transformation of the distance in the protein's

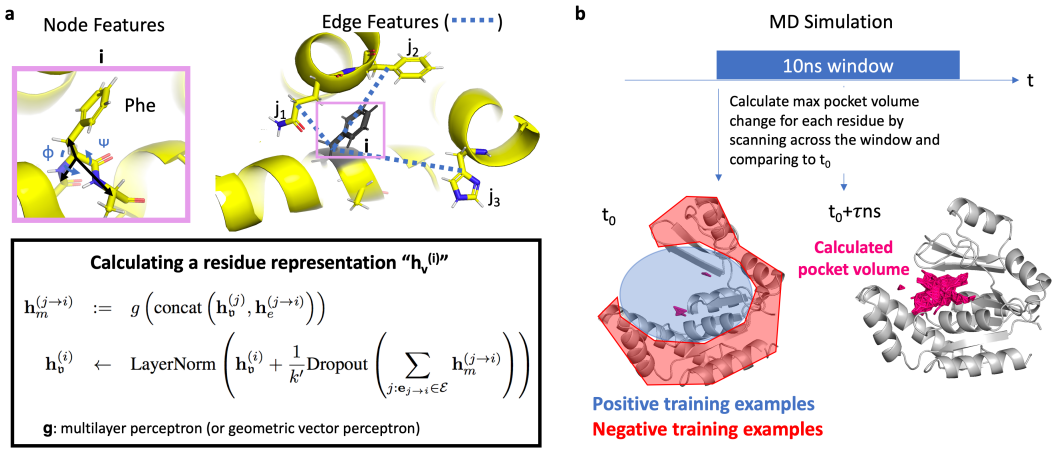

Figure 1: Depiction of how residue level representations are calculated in the graph neural network (a) and how residues are labelled for training (b). Node features include a one-hot encoding of residue type, sine and cosine transformations of 3 different backbone dihedral angles (only 2 shown for clarity), a unit vector capturing the direction of the C-α to C-β bond, and forward and reverse unit vectors (C-α to C-α from preceding and following residues). Edge features include a unit vector in the direction of the neighbor, the distance to the neighbor, and a sine transformation of the distance in sequence space.

primary sequence. We update a node of interest using its 30 nearest neighbors as done in the original work. In the original work, protein-level predictions were made by taking the mean representation of all residues and making a prediction. Importantly, we do not take a mean of residues, but instead predict on residues individually.

The training data for our model comes from time windows from molecular dynamics simulations. Specifically, we select a structure at some time-point in simulation (t_0) and the resulting structures within the next 10 nanoseconds of simulation to calculate the maximum pocket volume increase across the protein structures within that time window using the pocket detection algorithm LIGSITE [8] (**Fig. 1b**). LIGSITE outputs sets of "pocket grid points" that indicate cavities on the protein surface (i.e. points that are surrounded by protein on all sides). For all residues, we calculate how many pocket grid points are within 5 Angstroms of the residue. We label a residue a positive example if at some point in the time window that residue's assigned pocket volume increases by 40 Å^3 (roughly the size of an ADP molecule) relative to its volume at time t_0. We label a residue as a negative example if the change is less than 10 Å^3, and we do not consider residues with intermediate values.

3 Experiments & Results

3.1 Graph neural networks accurately predict residue level pocket volume changes from simulation data

We trained and evaluated a model using a simulation dataset of SARS-CoV-2 proteins (and related human proteins) that consisted of 17 proteins. First, we chose 15 proteins randomly (resulting in 1,160 simulation trajectories) and split the trajectories into training and validation sets in a 90:10 split. Then, we held out all trajectories from 2 proteins as a test set. Importantly, since the simulations provide many different structural configurations of the proteins and because each protein has many residues, the training set contained 1.6M training examples (176K positive, 1.4M negative). Our model effectively learned to classify how the pocket volume around a residue will change over the course of 10 ns of simulation. First, we show that the model trains stably with the training and validation loss flattening around 25 epochs (**Fig. 2a**). We apply the best model (according to validation loss) to the test set of 2 held out proteins and calculate a ROC-AUC of 0.82 (**Fig. 2b**). We also compare with the 3D CNN model from Torng et al.[9] which obtains an ROC-AUC of 0.64. Overlaying the ground truth labels and predicted labels onto a random structure selected from simulation also demonstrates the high accuracy of the model (**Fig. 2d**). Additionally, we plotted a precision-recall curve and

observe good performance (**Fig. 2c**). In both the ROC curve and precision-recall curve, our model substantially outperforms a random baseline, a model that classifies residues based on whether they are polar or nonpolar, and a model that classifies residues based on if they have secondary structure.

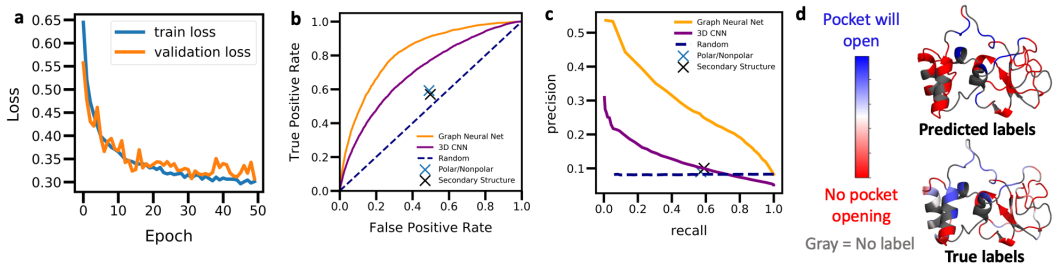

Figure 2: Several metrics evaluating the model relay its ability to accurately predict cryptic pocket opening in simulations. (a) Training and validation loss. (b) ROC curve. (c) Precision-recall curve. (d) Predictions and ground truth labels overlaid on a random structure from simulation.

3.2 Graph neural networks accurately identify cryptic pockets from experimental structures

To evaluate if our model can predict known sites of cryptic pockets without the need for simulations, we applied a trained model to a new test set of experimental protein structures with known cryptic pockets. This model was trained identical to the model from section 3.1 except it also included the 2 prior test set proteins. For the test set, we curated a set of 11 protein structure pairs that include an *apo* protein (no ligand bound) and the corresponding *holo* protein (ligand bound to cryptic site). We applied our model to the 11 *apo* protein structures (which were not part of the training) and found that it accurately identified the known cryptic pockets (**Fig. 3**, ROC-AUC=0.78). This result is slightly better than the reported result for the related algorithm CryptoSite [1] (ROC-AUC=0.74 when not using features extracted from simulations). However, the test sets used in the two studies are different and a follow-up comparison with an identical test set is warranted.

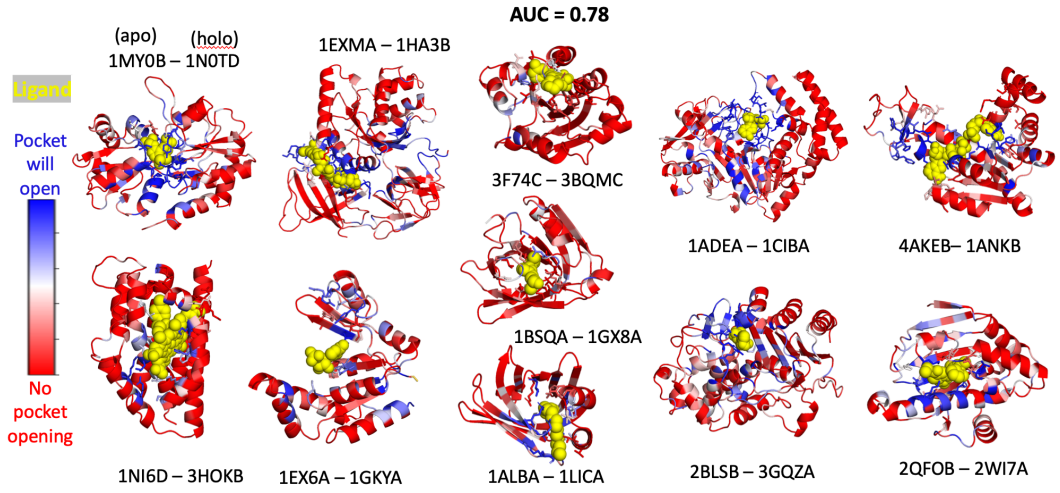

Figure 3: A graph neural network accurately predicts the locations of cryptic sites in 11 crystal structures with known cryptic pockets. *Apo* structures are colored from red to blue depending on the network's prediction of whether pocket opening will occur at that residue. Ligands that bind cryptic sites (shown in yellow) are superposed onto *apo* structures. Residues at the ligand binding site are labelled as positive examples and shown in a stick representation in the *apo* structures. PDB ID's for the *apo* and *holo* structures are provided.

4 Conclusions & Discussion

We have shown that a graph neural network trained on protein simulation data can accurately predict cryptic pocket opening. First, we showed that we could predict whether or not residues in a protein will undergo structural changes that lead to increased pocket volumes over the course of 10 ns of simulation. Next, we showed that this same model could accurately predict sites of cryptic pockets from experimental structures without the need to run molecular dynamics simulations.

While our model can accurately predict cryptic pocket opening, there are some caveats. The highest precision only reaches $\sim$0.5 meaning there is likely to be one false positive for every true positive at the lowest recall value. One explanation is that pocket formation is stochastic across 10 ns simulation windows. Even with the exact same starting structural configuration, a residue can sometimes form a pocket in 10 ns, and sometimes not. Therefore, many of the false positives called by our model may actually be residues that do sometimes form cryptic pockets in other 10 ns windows.

It may be surprising that our model can predict sites of cryptic pocket opening from experimental structures given that the model was not trained to perform this task. Nonetheless, we find that a model trained to predict how pocket volumes change in protein structures over the course of 10 ns of simulation is transferable to the more challenging task of predicting cryptic pocket opening from experimental structures. Given that success on the former task begets success on the latter task, this suggests that 10 ns of simulation time may be a sufficient amount to sample at least partial cryptic pocket openings with molecular dynamics simulations. This is encouraging since, historically, simulations to discover cryptic pockets usually consumed far more resources (100s of GPU hours).

This work represents an encouraging proof-of-concept and should be improved by access to more simulation training data on a larger set of proteins, as well as, better model selection, which can come from a hyperparameter search to find the best model.

Funding

G.R.B. and his lab were supported by funding from Avast, the Center for the Science and Engineering of Living Systems (CSELS), an NSF RAPID award, NSF CAREER Award MCB-1552471, NIH R01 GM124007, NIH RF1 AG067194, a Burroughs Wellcome Fund Career Award at the Scientific Interface and a Packard Fellowship for Science and Engineering.

References

1. Cimermancic P, Weinkam P, Rettenmaier TJ, et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. J Mol Biol. 2016. doi:10.1016/j.jmb.2016.01.029

2. Crews CM. Targeting the Undruggable Proteome: The Small Molecules of My Dreams. Chem Biol. 2010. doi:10.1016/j.chembiol.2010.05.011

3. Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. Cell. 2013. doi:10.1016/j.cell.2013.03.034

4. Hollingsworth SA, Kelly B, Valant C, et al. Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic GPCRs. Nat Commun. 2019. doi:10.1038/s41467-019-11062-7

5. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol. 2002. doi:10.1038/nsb0902-646

6. Zimmerman MI, Porter JR, Ward MD, et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nat Chem. 2021. doi:10.1038/s41557-021-00707-0

7. Jing B, Eismann S, Suriana P, Townshend RJL, Dror R. Learning from Protein Structure with Geometric Vector Perceptrons. 2020:1-18. http://arxiv.org/abs/2009.01411.

8. Hendlich M, Rippmann F, Barnickel G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model. 1997. doi:10.1016/S1093-3263(98)00002-3

9. Torng, Wen, and Russ B. Altman. "High precision protein functional site detection using 3D convolutional neural networks." Bioinformatics 35.9 (2019): 1503-1512.