DeepPCSite: A high-precision enzyme functional site predictor using deep graph convolutional neural networks on atomic point clouds

David Toomer
djtoomer@stanford.edu
Presentation: https://youtu.be/5LXdQsvC7wU

INTRODUCTION

DE NOVO DRUG DESIGN

- *De novo* drug design is the process of generating novel molecules to act as a drug, typically to activate or inactivate an enzyme.
- The first step to this process is determining the best regions on the enzyme to attack, known as **functional sites.**
- DeepPCSite attempts to discover these functional sites in a completely structure-based manner—i.e. without knowing the structure of the drug beforehand!
- Functional sites have common chemical motifs for the network to learn, such as hydrogen bonding and hydrophilicity.

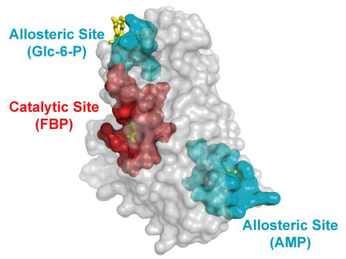


FIGURE: an example protein with three (3) functional sites.

THE EMBEDDING PROBLEM

- One of the most pivotal questions for the task is: **what is the best way to represent proteins to capture this chemical information?**

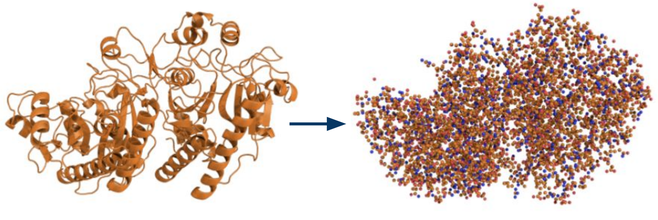
PROTEIN EMBEDDING

PREVIOUS APPROACH

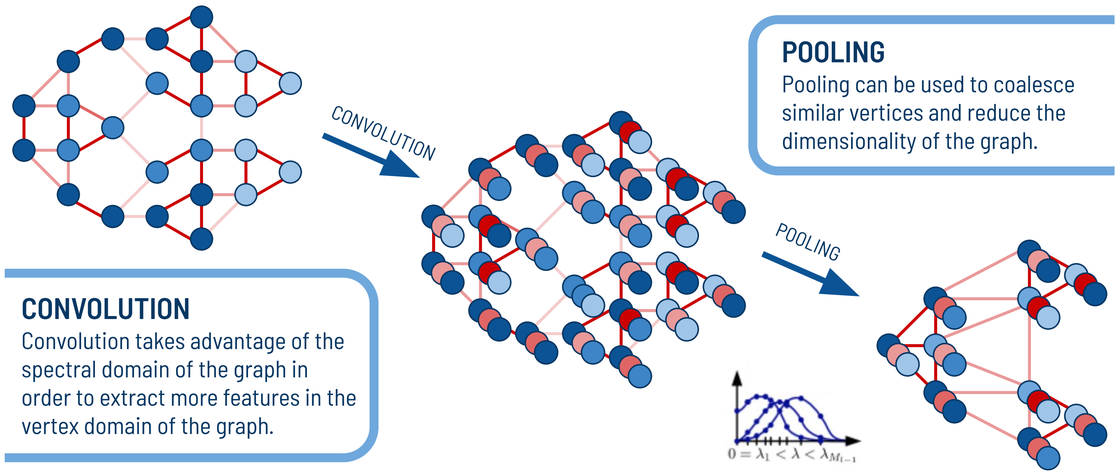
- The most common approach to protein representation is voxelization. However, this format is **sparse** and requires **lots of preprocessing.** Can we do better?

SOLUTION: ATOMIC POINT-CLOUDS

- Using the point-cloud data on each atom is more succinct.
- Interpreting the point-cloud as a k-nearest neighbors graph intrinsically incorporates spatial relationships, and chemical features can be written as signals in the graph vertex domain.
- Graph convolutional neural networks (GCNN) can learn point-clouds.



GRAPH CONVOLUTION



CONVOLUTION

Convolution takes advantage of the spectral domain of the graph in order to extract more features in the vertex domain of the graph.

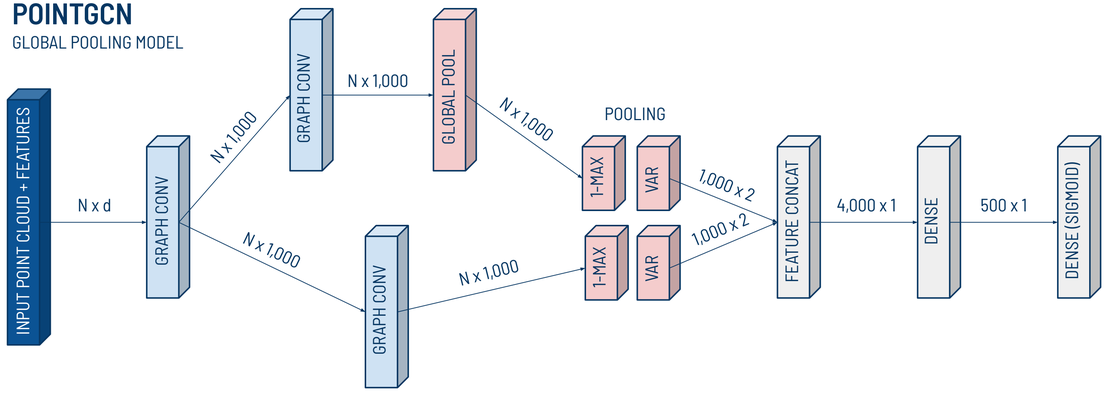
POOLING

Pooling can be used to coalesce similar vertices and reduce the dimensionality of the graph.

NETWORK ARCHITECTURE

POINTGCN

GLOBAL POOLING MODEL



GRAPH CONVOLUTIONAL NEURAL NETWORK (GCNN)

- DeepPCSite combines convolution layers, pooling layers, and more traditional layers, such as fully connected and sigmoid, to create a deep graph convolutional neural network
- Used sc-PDB database to label +/- examples of protein-ligand interactions
- **INPUT:** coordinates (point-cloud) and physicochemical features calculated for each atom of a region in an enzyme
- **OUTPUT:** whether or not the selected region of an enzyme can act as a functional site

NETWORK DETAILS

- Adam optimizer
- Dropout
- ReLU activation
- Sigmoid classification
- L2-regularization

RESULTS

EVALUATION METRICS

The network was evaluated on two scopes:

- site-level: classify whether a subregion can act as a functional site
- protein-level: predict the most likely functional site on an enzyme

SUBREGION CLASSIFICATION

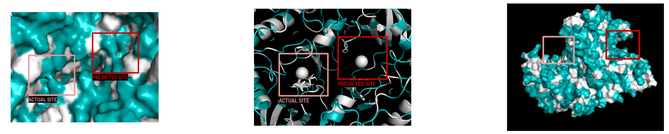
	PREDICTED: 1	PREDICTED: 0
ACTUAL: 1	340,887	23,616
ACTUAL: 0	10,627	1,903,012

- Precision: 0.970
- Recall: 0.935
- Accuracy: 0.985

FUNCTIONAL SITE PREDICTION

- Accuracy: 96.75%
- Average distance from the binding site: 12.7 angstroms

MISCLASSIFIED EXAMPLES REVEAL NON-ACTIVE INHIBITORY ELEMENTS



CONCLUSION

PERFORMANCE

- DeepPCSite performs competitively with other protein functional site predictors, but with less preprocessing and shorter train time
- Point-clouds are effective at learning chemoinformatical properties
- Improvement should be focused on increasing recall

FUTURE WORK

- Find point-cloud data augmentation methods in order to learn more positive samples
- Research the chemical latent space to see which features are most important to the network

REFERENCES

J. Desaphy, G. Bret, D. Rognan, and E. Kellenberger. sc-pdb: a 3d-database of ligandable binding sites—10 years on. Oxford Academic, 2014.
M. Jiang, Z. Li, Y. Bian, and Z. Wei. A novel protein descriptor for the prediction of drug binding sites. BMC Bioinformatics, 20(478):1–13, 2019.
W. Torng and R. Altman. High precision protein functional site detection using 3d convolutional neural networks. International Society for Computational Biology, 35(9):1503–1512, 2019.
R. Townshend, R. Bedi, P. Suriana, and R. Dror. End-to-end learning on protein structure for interface prediction. arXiv, pages 1–8, 2019.
Y. Zhang and M. Rabbat. A graph-cnn for 3d point cloud classification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.