

UC Berkeley

UC Berkeley Previously Published Works

Title

PersGNN: Applying Topological Data Analysis and Geometric Deep Learning to Structure-Based Protein Function Prediction

Permalink

<https://escholarship.org/uc/item/598565nb>

Authors

Swenson, Nicolas
Krishnapriyan, Aditi S
Buluc, Aydin
[et al.](#)

Publication Date

2020-10-29

Peer reviewed

PersGNN: Applying Topological Data Analysis and Geometric Deep Learning to Structure-Based Protein Function Prediction

Nicolas Swenson*

Aditi S. Krishnapriyan*

Aydin Buluc

Dmitriy Morozov

Katherine Yelick

Lawrence Berkeley National Laboratory &
Department of Electrical Engineering and Computer Science,
University of California, Berkeley
Berkeley, CA, 94720

Abstract

Understanding protein structure-function relationships is a key challenge in computational biology, with applications across the biotechnology and pharmaceutical industries. While it is known that protein structure directly impacts protein function, many functional prediction tasks use only protein sequence. In this work, we isolate protein structure to make functional annotations for proteins in the Protein Data Bank in order to study the expressiveness of different structure-based prediction schemes. We present PersGNN—an end-to-end trainable deep learning model that combines graph representation learning with topological data analysis to capture a complex set of both local and global structural features. While variations of these techniques have been successfully applied to proteins before, we demonstrate that our hybridized approach, PersGNN, outperforms either method on its own as well as a baseline neural network that learns from the same information. PersGNN achieves a 9.3% boost in area under the precision recall curve (AUPR) compared to the best individual model, as well as high F1 scores across different gene ontology categories, indicating the transferability of this approach.

1 Introduction

Predicting protein function from its raw amino acid sequence is a long standing challenge in computational biology. This is an ideal setting for computational methods because experimental characterization of proteins is costly and time-consuming, and there is an abundance of protein sequence data. Public databases such as Uniprot contain over 100 million protein sequences [1]. The Critical Assessment of protein Function Annotation (CAFA), a recurring challenge that benchmarks different computational approaches, has shown that machine learning and statistical algorithms outperform alignment-based techniques, such as BLAST, or homology transfer methods [2, 3].

However, proteins are not just sequences of amino acids: they fold into complex three-dimensional motifs, which directly impact a protein’s function [4]. Recent work has shown that using protein sequences together with its three-dimensional structural information can lead to better functional predictions, as well as provide scientists with a way to identify the functionally-active areas of a protein [5]. This work is made possible by advances in structural biology, such as X-ray crystallography,

*Equal contribution. {nswenson, akrishnapriyan}@lbl.gov

which have allowed for the structure of many proteins to be determined. For example, the Protein Data Bank (PDB) contains structural information for over 100,000 proteins and other biological molecules [6].

Given the success of these new structure-aware machine learning techniques, in this work, we seek to maximize the usefulness of structure for functional characterization. Since protein structure determination can be laborious and expensive, optimal utility of these structures is of particular importance. We introduce PersGNN, a method that combines topological data analysis (specifically, persistent homology) and graph neural networks to create a more nuanced representation of the protein structure. To the best of our knowledge, this is the first time that these approaches have been combined in this way. Since determining the protein structure via experimental methods is challenging, maximizing the utility of the structural data is key. We show that our hybrid model learns more from structure than either model on its own, and also outperforms a standard neural network that is given the same information. In future work, protein sequence models can also be incorporated in order to create even better functional annotation models.

2 Methods

2.1 Graph Neural Networks

Graph Neural Networks (GNN), which can be motivated through spectral graph convolutions, are a popular form of graph representation learning that provide a framework for extending traditional deep learning techniques to non-Euclidean, graphical data [7, 8]. GNNs learn context-aware node embeddings through successive rounds of local neighborhood aggregation (or "message passing") in a graph. These node embeddings can be combined to create global representations for entire graphs. Previous work has used 3D Convolutional Neural Networks (3D CNNs) to extract useful information from protein structure [9, 10]. However, this technique is both memory and computationally inefficient because proteins are sparse in 3D space and 3D CNNs will perform many convolutions over empty space. More recent work proposed to overcome some of these inefficiencies with a GNN [5]. In this method, 3D protein structures are represented by contact maps, which threshold the pairwise distances between the alpha-carbons of each residue in the protein. These contact maps are then fed into the GNN, which embeds structural information that can be used to make function predictions.

Our GNN model architecture is inspired by the one of *Glorigorijevic et al.*, which uses the graph convolutional layers from *Kipf and Welling*, along with language modeling [5, 7]. In contrast, our work focuses on learning from structure. Thus, we omit the language modeling component and instead opt for a simple one-hot representation of amino acids. Other modeling differences, such as model depth and learning rate, were optimized during the validation stage.

2.2 Persistence Network

We use persistent homology, an area of topological data analysis, to construct a topological representation of the protein structure. In this section, we briefly describe the approach and refer the interested reader to a more thorough survey [11].

We start with the 3D atomic coordinates of a protein structure and take the union of balls around the atoms. We vary the radii of these balls, and track how the topology of this union changes: these are represented as alpha shapes [12]. By doing this and sweeping across all radii, we get an increasing sequence of nested alpha shapes called the *filtration*. By following the changes in the sequence, one can keep track of the appearances and disappearances of topological features in the filtration.

We record the pairs of radii when such topological features appear and disappear, where the appearance of a new feature is the *birth* and the disappearance of this same feature (such as when it gets merged with another feature) is the *death*. The difference between the *birth-death pairs*, i.e. $death - birth$, is called the *persistence* of the feature. The larger the *persistence*, the more prominent the topological feature. The full set of all *birth-death pairs* is called a *persistence diagram*. In this work, we specifically look at the 1-dimensional and 2-dimensional persistence diagrams of each protein structure, where 1-dimensional features correspond to channels ("loops") and 2-dimensional features correspond to voids ("cavities") in the protein structure.

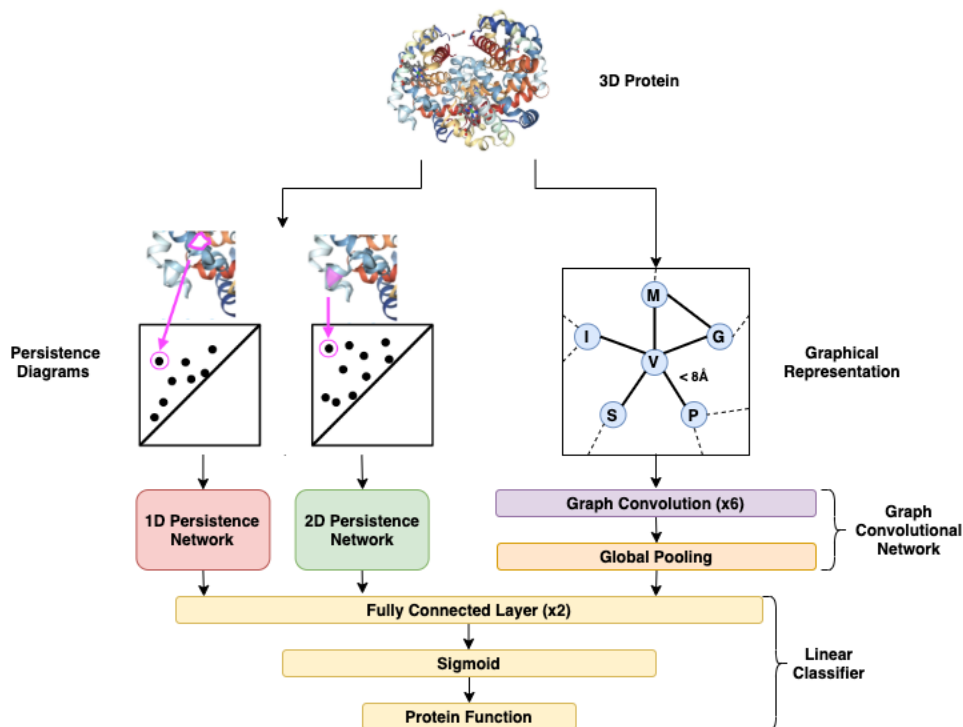


Figure 1: **PersGNN model architecture.** Starting from a protein’s 3D structure, we compute 1D and 2D persistence diagrams (summaries of all of the topological features, i.e. channels and voids, in a protein) and $C_\alpha - C_\alpha$ contact maps. The persistence networks (PersNets) compute a vectorized representation of each persistence diagram. The graph neural network (GNN) combines the contact map with one-hot encoded amino acid labels to learn a separate vectorization for the whole structure. These representations are concatenated together and fed into a two-layer multi-layer perceptron (MLP) to predict molecular function (MF) gene ontology (GO) terms. For the individual models (Sections 2.1 and 2.2), we use only the GNN and PersNet pipelines respectively.

Feeding persistence diagrams, in this case the topological summaries of all the channels and voids in a protein structure, into a neural network architecture requires a transformation of the diagram points. An approach to do this was first proposed by Hofer et al. [13]. Expanding on this, Carrière et al. [14] proposed a layer for neural network architectures based on the DeepSet architecture [15] to encode a vector representation of the persistence diagram through point-wise transformations. We implement a modification of this approach to process both 1D and 2D persistence diagrams of the protein structure, which we call “PersNet” in this work.

The problem of describing protein shape was one of the early motivations behind persistent homology [16]. Persistent homology has also been used previously to construct topological representations to describe protein structures for protein classification problems [17, 18]. However, these approaches turn topological information into feature vectors and rely on the construction of ad hoc handcrafted summaries. More recently, vectorizing persistence diagrams in ways that require minimal processing and retain more of the information, such as to persistence images [19], have been used in machine learning applications for scientific domains [20, 21, 22, 23]. However, these vectorized representations are static — in contrast, the approach outlined here processes the topological information through an automatically differentiable layer, thereby continuing to learn the representation over the training process.

2.3 PersGNN: Hybrid network

Using a GNN in place of a 3D CNN significantly reduces computation time and memory requirements of learning on protein structures, but also requires representing the protein structure with a contact map. This simplification is necessary for the GNN to work but can also lead to a loss of key structural

information. To overcome this, we propose incorporating persistent homology into the learning process by converting the 3D atomic coordinates of the protein structure to persistence diagrams and processing this information through a layer in the neural network architecture, as described in 2.2. We note that previous work has used persistent homology together with 3D CNNs to classify protein folds [23]; however, as mentioned earlier, they used a static vectorization of the persistence diagrams of the protein structure rather than a neural network processing layer that learns the representation through training and did not use GNNs — both things that we found very important in our approach to achieving high accuracy. There has also been previous work combining persistent homology and GNNs [24], but here they incorporated persistent homology into the graph neural network architecture itself by using persistent homology to reweight messages passed between the graph nodes during convolutions. Our method allows us to build on the successes of applying GNNs to protein data (and capitalize on their computational efficiency) while also adding additional structural features that are potentially not captured by a protein’s contact map, such as more global topological information.

Our hybrid model, shown in Figure 1, concatenates the output from the GNN with the output from the PersNet models that process both the 1D and 2D topological features (channels and voids respectively). This hybrid representation is then passed into a Multi-Layer Perceptron (MLP) to perform the classification task. The architecture of the GNN and PersNet models is the same as the ones described in Sections 2.1 and 2.2. PersGNN is trained end-to-end on a protein’s contact maps, amino acid labels, and persistence diagrams and all sub-components are trained simultaneously. This enables PersGNN to learn a complex representation of protein structure efficiently and with minimal manual processing.

2.4 Baseline

In addition to benchmarking our hybrid model against its component models, we trained a simple three-layer MLP to create a more complete picture of how these models are learning from structure. This is distinct from many assessments of protein functional annotations, such as CAFA, which use the alignment-based BLAST as a baseline. Since this investigation is focused on extracting information from protein structure, rather than its amino acid sequence, we decided not to use the sequence-only BLAST (or any other sequence-based methods) as a baseline. The baseline model is given all the same information as the GNN: for each residue, the model is given the amino acid label and the number of residue “contacts,” which are computed from the same contact maps that are fed into the GNN.

2.5 Data and Model Parameters

Our dataset consists of 33,007 proteins taken from PDB that belong to the Arabidopsis, Celegans, E.coli, Fly, Human, Mouse, Yeast and Zebrafish species. This is randomly split into train, validation and test sets at a 75%, 12.5%, 12.5% ratio. Random splitting can sometimes be overly optimistic and not representative of realistic challenges [25]. Previous work, for example, has split their datasets to minimize % sequence identity between training and tests set, which we leave for future work.

For labels, we represent protein function with molecular function (MF) gene ontology (GO) terms. For each term in PDB, we obtain its corresponding MF GO terms using the Structure Integration with Function, Taxonomy and Sequence (SIFTS) database [26, 27]. Consistent with previous studies, we filter out GO terms with less than 25 representative proteins. After the filtration, our label space is made up of 730 MF GO terms. Each protein can also have multiple labels, making this a multi-label multi-classification problem.

Each protein entry in PDB is described by a list of amino acid residues and their 3D atomic coordinates. For the GNN and MLP models, the 3D atomic coordinates are converted to $C_\alpha - C_\alpha$ contact maps. To construct these contact maps, we compute pairwise distances between the α -carbon of each residue and denote a "contact" between residues if their α -carbons are within eight angstroms. The GNN and MLP models also take a one-hot encoded representation of amino acids, which the GNN uses as node labels. We also construct 1D and 2D persistence diagrams from the 3D atomic coordinates of each protein structure, which are then processed through the persistence network layers.

All models are trained using a weighted binary cross-entropy loss function [5], the ADAM optimizer [28], for 300 epochs and with a learning rate of 10^{-5} . To normalize our models, we employ weight normalization and dropout regularization ($p = 0.1$) throughout [29, 30]. For each method, we create

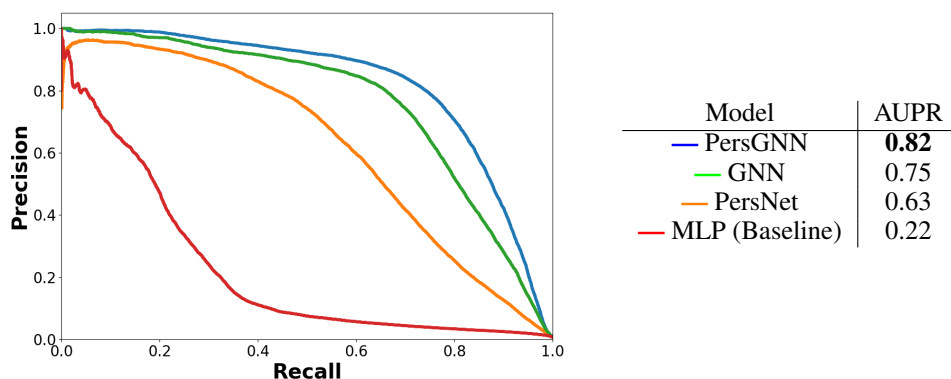


Figure 2: **Precision-recall curves for predicting molecular function (MF) gene ontology (GO) terms.** Precision-Recall curves for four models: the "baseline" model (a multi-layer perceptron) trained on protein contact maps, a persistence network (PersNet) trained on the persistence diagrams created from the 3D atomic coordinates of each protein, a graph neural network (GNN) trained on protein contact maps, and PersGNN, our model that combines PersNet and the GNN. For each method, we independently train ten models and ensemble them by computing the average probability score for each class. Area under the precision-recall curve (AUPR) scores on the test set for each method are shown to the right, where higher scores indicate better accuracy. Our model, PersGNN, achieves the highest AUPR score.

an ensemble by independently training 10 models and taking the average bit-score (per GO term) across the 10 models at evaluation time.

3 Results and Discussion

Our hybrid method, PersGNN, outperforms both GNN and PersNet on their own, and significantly outperforms the baseline MLP that is given the same information as the GNN. The performance of each method, measured in area under the precision-recall curve (AUPR) for molecular function (MF) gene ontology (GO) terms is shown in Figure 2. PersGNN has an AUPR score 9.3% higher than the GNN, the next best model. To focus our study on learning from protein structure, we have not included highly expressive sequence models, such as BLAST or 1D CNNs, nor did we use language models to compute amino acid embeddings. The GNN, however, can learn to embed amino acids through a residue's local neighborhood in the graph structure. PersNet is able to capture further topological information through a protein's 1D and 2D persistence diagrams. When combined, the GNN and PersNet capture complementary information as indicated by the higher AUPR score, thus creating a more complete representation of the protein structure.

In Figure 3, we compute average F1 scores aggregated over different GO categories, which are grouped at various levels of the MF GO hierarchy. F1 scores are a measure of the model accuracy, calculated from the precision and recall, where higher F1 scores indicate that the model was able to successfully classify more proteins. As we see in Figure 3, the PersGNN model has consistently high F1 scores across GO categories, and performs better than the other methods on every GO category and almost every individual GO term.

Figure 4 shows the effects of training set size (the number of times each GO term appears in the training dataset) against model accuracy, again represented via F1 scores. As we see, PersGNN achieves high F1 scores even on GO terms with fewer training examples, while other models like the MLP perform poorly in this regime. The ability of PersGNN to make accurate predictions even with a low training set size is optimistic, as it indicates the model is making good use of the protein structure information. Moreover, while there are millions of raw amino acid sequences, there are far fewer available protein structures, meaning achieving high model accuracy with lower amounts of data is especially important here.

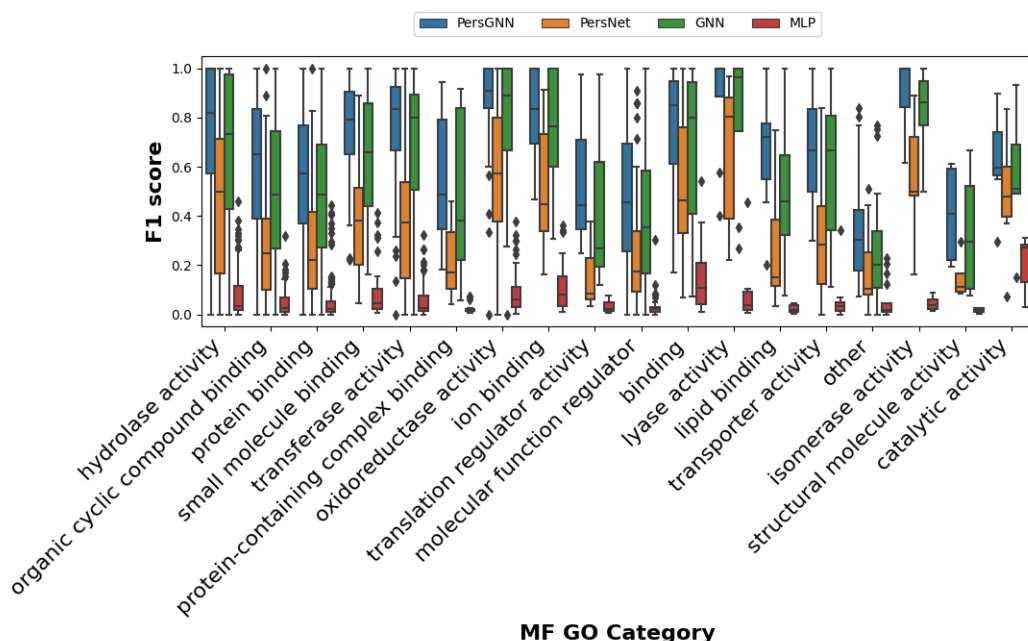


Figure 3: **Model performance by GO term type.** Boxplot showing average F1 scores (higher is better) aggregated over umbrella GO term categories for all four models. PersGNN achieves the highest average F1 scores across all MF categories, showing the transferability of this approach. This indicates that the topological representation from persistent homology is learning complementary structural information to the GNN, such as the ability to capture more global information. As a result, PersGNN is able to learn a more nuanced representation of structure that is important for making accurate functional annotations.

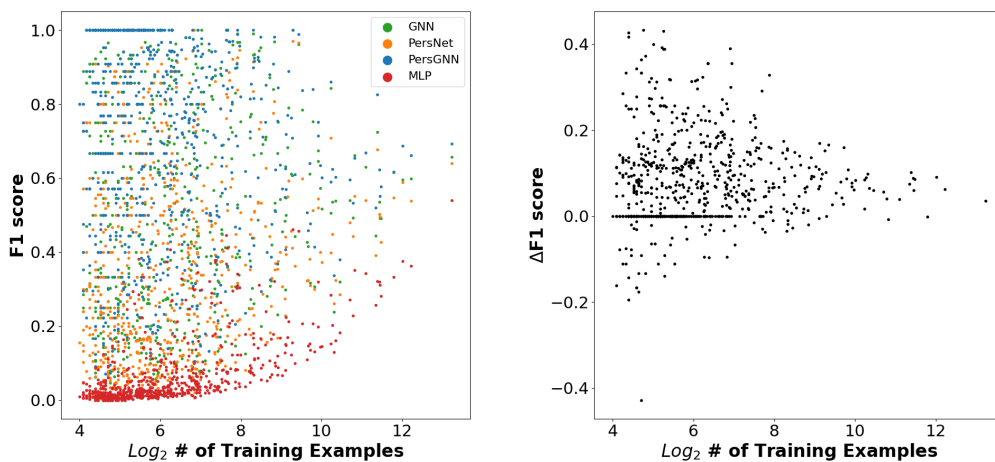


Figure 4: **Effect of number of training examples on model performance.** F1 scores are plotted by MF GO term for all four models, where higher F1 scores indicates more accuracy (left) and the change in F1 score per GO term from PersGNN to GNN, where positive Δ means PersGNN scored higher (right). PersGNN outperforms other models on GO terms with fewer training examples. This suggests that PersGNN is constructing more meaningful structural features, enabling it to learn from fewer training examples.

Our method, PersGNN, more accurately predicts MF GO terms compared to other structure-based methods, including across different categories and with fewer examples. This motivates a further investigation into its performance. Future work in this area should study PersGNN’s performance on all three GO term categories (Biological Process and Cellular Component). In addition, it is known that random splits are often too optimistic, so future work should also evaluate performance on curated splits. We decided to limit the focus of this study to generating representations of protein structure and understanding protein structure–function relationships, despite the limited availability of high quality structural data. Other work has shown that structure-based methods trained on high quality structural data can still perform well when tested with predicted structures [5]. With the continued advances in structure prediction methods (both De Novo and Ab Initio), this trend is likely to persist. As our model achieved high accuracy for protein function predictions using only the structural information of a protein structure (namely only through the 3D atomic coordinates), this confirms the value of incorporating structure into predictive models. Moreover, this approach requires minimal processing of the input data, making it a transferable across different datasets.

Broader Impact

Due the challenges of experimentally characterizing a protein’s function, predicting protein function from sequence and structure is an ongoing challenge in the bioinformatics community (as evidenced by the recurring CAFA challenge). The ability to quickly and accurately annotate proteins using computational methods will have a great impact in metagenomics, drug-discovery, and many other biological applications. Many of the efforts to predict protein function rely on sequence alone, even though it is known that structure directly influences function. This is because sequences are much more abundant, due to rapid improvements in sequencing technology. As more and more protein structures become available, it will be important to incorporate structure into the function prediction pipelines.

Here, we present a new method that combines complementary representations of structure to improve functional annotations, achieving high accuracy by using only the protein structure. This is not a tool that is intended to be used in isolation; rather, it demonstrates that structural information and specifically persistent homology and GNNs should be part of the toolset used for analyzing proteins. Combined with sequence information, we believe this will prove to be a powerful hybrid method for functional annotation.

Acknowledgments and Disclosure of Funding

The authors would like to thank Jiali Chen, Jude Fernandes, Nick Bhattacharya and Andrew Tritt for their help and guidance.

This work was supported by the U.S. Department of Energy under Contract Number DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory. A.S.K. is an Alvarez Fellow in the Computational Research Division at LBNL. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research was also supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. The authors declare they have no competing financial interests.

References

- [1] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.
- [2] Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23, 2019.

- [3] Ronghui You, Shuwei Yao, Yi Xiong, Xiaodi Huang, Fengzhu Sun, Hiroshi Mamitsuka, and Shanfeng Zhu. Netgo: improving large-scale protein function prediction with massive network information. *Nucleic Acids Research*, 47(W1), 2019.
- [4] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. *Biochemistry, 5th edition*. W H Freeman, New York, 2002.
- [5] Vladimir Gligorijevic, P. Douglas Renfrew, Tomasz Kosciolk, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *bioRxiv*, 2020.
- [6] H.M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10, 2003.
- [7] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.
- [8] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs, 2014.
- [9] J. Jiménez, S. Doerr, G. Martínez-Rosell, A.S. Rose, and G. De Fabritiis. Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33:3036–3042, 2017.
- [10] Afshine Amidi, Shervine Amidi, Dimitrios Vlachakis, Vasileios Megalooikonomou, Nikos Paragios, and Evangelia I Zacharaki. Enzyenet: enzyme classification using 3d convolutional neural networks on spatial representation. *PeerJ*, 2018.
- [11] Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.
- [12] Herbert Edelsbrunner and Ernst P Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics (TOG)*, 13(1):43–72, 1994.
- [13] Christoph D Hofer, Roland Kwitt, and Marc Niethammer. Learning representations of persistence barcodes. *Journal of Machine Learning Research*, 20(126):1–45, 2019.
- [14] Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR, 2020.
- [15] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.
- [16] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.
- [17] Zixuan Cang, Lin Mu, Kedi Wu, Kristopher Opron, Kelin Xia, and Guo-Wei Wei. A topological approach for protein classification. *Computational and Mathematical Biophysics*, 1(open-issue), 2015.
- [18] Tamal K Dey and Sayan Mandal. Protein classification with improved topological data analysis. In *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- [19] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research*, 18(1):218–252, 2017.
- [20] Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. Uncovering the topology of time-varying fmri data using cubical persistence. *arXiv preprint arXiv:2006.07882*, 2020.
- [21] Aditi S Krishnapriyan, Maciej Haranczyk, and Dmitriy Morozov. Topological descriptors help predict guest adsorption in nanoporous materials. *The Journal of Physical Chemistry C*, 124(17):9360–9368, 2020.

- [22] Aditi S Krishnapriyan, Joseph Montoya, Jens Hummelshøj, and Dmitriy Morozov. Persistent homology advances interpretable machine learning for nanoporous materials. *arXiv preprint arXiv:2010.00532*, 2020.
- [23] Yechan Hong, Yongyu Deng, Haofan Cui, Jan Segert, and Jianlin Cheng. Classifying protein structures into folds by convolutional neural networks, distance maps, and persistent homology. *BioRxiv*, 2020.
- [24] Qi Zhao, Ze Ye, Chao Chen, and Yusu Wang. Persistence enhanced graph neural network. In *International Conference on Artificial Intelligence and Statistics*, pages 2896–2906, 2020.
- [25] Robert P. Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790, 2013.
- [26] Sameer Velankar, José M. Dana, Julius Jacobsen, Glen van Ginkel, Paul J. Gane, Jie Luo, Thomas J. Oldfield, Claire O’Donovan, Maria-Jesus Martin, and Gerard J. Kleywegt. Sifts: Structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research*, 41(D1):D483–D489, 11 2012.
- [27] Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O’Donovan, Maria Martin, and Sameer Velankar. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research*, 47(D1):D482–D489, 11 2018.
- [28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- [29] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks, 2016.
- [30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1), 2014.