Struct2Graph: A graph attention network for structure based predictions of protein-protein interactions

Mayank Baranwal[1*], Abram Magner[2], Jacob Saldinger[3], Emine S. Turali-Emre[4], Shivani Kozarekar[3], Paolo Elvati[5], J. Scott VanEpps[4,6,7], Nicholas A. Kotov[3,4,7,8], Angela Violi[3,5,9], Alfred O. Hero[4,10,11,12,13]

1 Division of Data & Decision Sciences, Tata Consultancy Services Research, Mumbai, Maharashtra, India
2 Department of Computer Science, University of Albany, SUNY, Albany, NY, USA
3 Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, USA
4 Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA
5 Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, USA
6 Department of Emergency Medicine, University of Michigan, Ann Arbor, MI, USA
7 Biointerfaces Institute, University of Michigan, Ann Arbor, MI, USA
8 Department of Materials Science & Engineering, University of Michigan, Ann Arbor, MI, USA
9 Biophysics Program, University of Michigan, Ann Arbor, MI, USA
10 Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI, USA
11 Department of Statistics, University of Michigan, Ann Arbor, MI, USA
12 Program in Applied Interdisciplinary Mathematics (AIM), University of Michigan, Ann Arbor, MI, USA
13 Program in Bioinformatics, Ann Arbor, MI, USA

* baranwal.mayank@tcs.com

Abstract

Development of new methods for analysis of protein-protein interactions (PPIs) at molecular and nanometer scales gives insights into intracellular signaling pathways and will improve understanding of protein functions, as well as other nanoscale structures of biological and abiological origins. Recent advances in computational tools, particularly the ones involving modern deep learning algorithms, have been shown to complement experimental approaches for describing and rationalizing PPIs. However, most of the existing works on PPI predictions use protein-sequence information, and thus have difficulties in accounting for the three-dimensional organization of the protein chains. In this study, we address this problem and describe a PPI analysis method based on a graph attention network, named *Struct2Graph*, for identifying PPIs directly from the structural data of folded protein globules. Our method is capable of predicting the PPI with an accuracy of 98.89% on the balanced set consisting of an equal number of positive and negative pairs. On the unbalanced set with the ratio of 1:10 between positive and negative pairs, Struct2Graph achieves a five-fold cross validation average accuracy of 99.42%. Moreover, *unsupervised* prediction of the interaction sites by Struct2Graph for phenol-soluble modulins are found to be in concordance with the previously reported binding sites for this family.

Author summary

PPIs are the central part of signal transduction, metabolic regulation, environmental sensing, and cellular organization. Despite their success, most strategies to decode PPIs use sequence based approaches do not generalize to broader classes of chemical compounds of similar scale as proteins that are equally capable of forming complexes with proteins that are not based on amino acids, and thus lack of an equivalent sequence-based representation. Here, we address the problem of prediction of PPIs using a first of its kind, 3D structure based graph attention network (available at https://github.com/baranwa2/Struct2Graph). Despite its excellent prediction performance, the novel mutual attention mechanism provides insights into likely interaction sites through its knowledge selection process in a completely *unsupervised* manner.

Introduction

Protein–protein interactions (PPIs) are fundamental to many biological processes. Analysis of the human proteome suggests that the majority of proteins function not alone but rather as part of multi-unit complexes [1]. Indeed, PPIs are the central part of signal transduction, metabolic regulation, environmental sensing, and cellular organization [2]. In these processes, PPIs can alter enzyme kinetics, facilitate substrate channeling, form new binding sites, render a protein inactive, or modify the specificity of a protein with respect to a substrate [3]. Due to the ubiquitous presence of PPIs in living systems, being able to characterize these interactions promises to further our understanding of cellular processes [4] and provide an indispensable tool for disease treatment and drug discovery [5, 6]. PPI and their mathematical description are also essential for creation of protein analogs from other nanoscale building blocks, including but not limited to, lipids [7], sugars [8], polymers [9], nanoscale conjugates [10], and inorganic nanoparticles [11, 12].

A number of strategies have been employed to decode PPIs. Traditionally, high throughput experimental techniques such as two-hybrid screens [13], tandem-affinity purification [14], and mass spectrometry [15] have been applied to create protein interaction networks. Concerns about insufficient accuracy [16], low experimental throughput [17] and high cost [18] of these methods, however, have motivated computational approaches that can complement traditional and robotic experimental protocols. Computational methods can predict whether proteins will interact based on data for the proteins' genetic context, amino acid sequences, or structural information. Genomics analyses consider factors such as gene fusion [19], conservation across common species (phylogenetic profiling) [20], and evolutionary history [21] when determining if a pair of proteins interact.

Typical computational techniques for PPI analysis use the amino acid sequences of the two proteins to determine whether interactions occur. A number of features such as frequency of common sub-sequences [22] and auto-covariance [23] have been proposed to convert sequences of different lengths into a uniformly sized representation. Sequence based methods have recently been able to leverage protein databases and machine learning techniques to make high accuracy predictions. Three-dimensional (3D) structure of protein-protein complexes from sequence can be predicted by CO-threading algorithm, (COTH) that recognizing templates of protein complexes from solved complex structural databases. COTH aligns amino acid chain sequences using scoring functions and structural information [24]. The DeepPPI model [25] predicts interactions using an artificial neural network, which takes as input a feature vector that captures the composition, distribution, and order of the sequence. DeepFE [26] uses natural

language processing algorithms on amino acid sequences to create low dimensional embeddings of the sequence suitable as inputs for neural network analysis. DeepFE, in particular, has been shown to be quite effective, and achieves prediction accuracy of 94.78% and 98.77% on *S. cerevisiae* and human datasets, respectively.

Despite their success, the above sequence based approaches do not generalize to broader classes of chemical compounds of similar scale as proteins that are equally capable of forming complexes with proteins that are not based on amino acids, and thus lack of an equivalent sequence-based representation. While the interaction of proteins with DNA can be accurately predicted [27], the supramolecular complexes with high molecular weight lipids [7], sugars [8], polymers [9], dendrimers [28] and inorganic nanoparticles [11, 12] that receive much attention in nanomedicine and nanodiagnostics, cannot [29–35]. As a consequence, computational approaches that take into account the structure of proteins and their variable counterparts are preferred for interaction prediction tasks, as these methods are not protein-specific. Some methods predicting interactions using the 3D structure of the proteins [36, 37] use a knowledge-based approach to assess the structural similarity of candidate proteins to a template protein complex. As this methodology requires detailed information on the larger complex, template-free docking approaches [38] analyze the unbound protein components and identify the most promising interactions from a large set of potential interaction sites. While docking methods have shown success for some proteins, they face difficulty with proteins undergoing conformational changes during interaction [39]. Many of these structural approaches have also served as the basis for machine learning models. Zhang et al. developed PrePPI [40] which uses amino acid sequence, and phylogenetic features as inputs to a naive Bayes classifier. Northey et al. developed IntPred [41] which segments proteins into a group of patches that incorporates 3D structural information into a feature set to predict interaction with a multi-layer perception network. Recently, [42] integrated 3D structures into a deep learning framework which uses a graph convolutional network to determine which amino acids will be part of an interacting protein interface. However, these models are trained on carefully curated interaction databases consisting of information not only on the binary interactions between proteins, but also on the corresponding interfacing sites or atoms.

In this work, we make the first step toward generalized method to assess supramolecular interactions of proteins with other nanostructures by determining the probability of formation of protein-protein complexes not based on amino-acid amino sequence but rather on nanoscale representation of proteins from crystallographic data. We developed a mutual graph attention network and a corresponding computational tool, *Struct2Graph*, to predict PPIs solely from 3D structural information. Instead of using several protein specific features, such as, hydrophobicity, solvent accessible surface area (SASA), charge, frequency of ngrams, etc., Struct2Graph uses a graph based representation of a protein globule obtained using only the 3D positions of atoms. This graph based interpretation allows for neural message passing [43] for efficient representation learning of proteins. Struct2Graph builds upon our prior work on metabolic pathway prediction [44], where it is shown that an equivalent graph-based structural representation of small molecules and peptides coupled with graph convolutional network, significantly outperforms other classifiers that involve computing various biochemical features as inputs. This approach also leverages generalization of graph theory to describe complex nanoscale assemblies similar to PPI [45].

Beyond its performance, characterized by high accuracy of PPI predictions, Struct2Graph offers a number of advantages. Struct2Graph only requires the 3D structure of individual proteins. In contrast to many other machine learning frameworks, genetic knowledge or sequence information is not required. Furthermore, while in this paper we focus on protein interactions, by using only the positions of

atoms in our analysis, this framework can be generalized to other molecular structures where 3D information is available. Moreover, Struct2Graph is also able to provide insight into the nature of the protein interactions. Through its attention mechanism, the model is able to suggest likely interaction sites. Unlike other models, Struct2Graph is able to produce this data in an unsupervised manner and thus does not require protein complex information which are often unavailable. The key contributions of the proposed work can be summarized as:

- GCN framework for PPI prediction: Struct2Graph uses a multi-layer graph convolutional network (GCN) for PPI prediction from the structural data of folded protein globules. The proposed approach is general and can be applied to other nanoscale structures where 3D information is available.

- Curation of PPI database: A large PPI database comprising of only direct/physical interaction of *non-homologous* protein pairs is curated, along with information on the corresponding PDB files. Special emphasis is based on curation of PDB files based on the length of the chain ID and highest resolution within each PDB file to ensure capturing of the most complete structure information of the protein of interest.

- State-of-the-art prediction performance: Our method is capable of correctly predicting the PPIs with an accuracy of 98.89% on the balanced set consisting of an equal number of positive and negative pairs. On the unbalanced set with the ratio of 1:10 between positive and negative pairs, Struct2Graph achieves a five-fold cross validation average accuracy of 99.42%. Struct2Graph outperforms not only the classical feature-based machine learning approaches, but also other modern deep-learning approaches, such as Deep-PPI and DeepFE-PPI that use sequence information and feature selection for PPI prediction.

- Unsupervised Prediction of interaction sites: The novel mutual attention mechanism provides insights into likely interaction sites through its knowledge selection process in an unsupervised manner. The interaction sites predicted by Struct2Graph for staphylococcal phenol-soluble modulins are found to be in concordance with the previously reported binding sites for this family. It must be noted that none of the constituent proteins (i.e., toll-like receptor 4 (TLR-4), phenol soluble modulin α_1 (PSMα_1), and high mobility group box-1 (HMGB1)) used to validate unsupervised prediction of interaction sites, were included in the training set.

Materials and methods

PPI Database

Struct2Graph focuses on structure-based predictions and interaction sites of the protein pairs. Our PPI database is therefore produced based on only direct/physical interactions of proteins. To build a large physical interaction database, comprising of only *heterologous* pairs, we searched all possible databases available (STRING, BioGRID, IntAct, MINT, BIND, DIP, HPRD, APID, OpenWetWare). Not all PPI databases use the same publications and same ontologies to report the interactions. Consequently, it is not surprising that each database reports PPI differently. Therefore, only up to a 75% concordance between all PPI databases is achieved [46]. For Struct2Graph, two of the largest compiled databases, IntAct [47] and STRING [48] are chosen for further analysis, and results are compared to each other to find the true interactions. Only concordant matches between these two databases are chosen. Struct2Graph database is compiled

from commonly used organisms (*Saccharomyces cerevisiae*, *Homo sapiens*, *Escherichia coli*, *Caenorhabditis elegans* and *Staphylococcus aureus*) PPIs. For these organisms, IntAct provides 427,503 PPIs, and STRING provides 852,327 binding PPIs.

STRING distinguishes the type of interactions as "activation", "binding", "catalysis", "expression", "inhibition" and "reaction". IntAct, on the other hand, describe the type of interactions as "association", "physical association", "direct association/interaction", and "colocalization". Only "direct association/interactions" from IntAct and "binding" from STRING were considered as physical interactions. We only choose concordant pairs of physical interactions from both databases. Therefore, extracting only concordant, physical interaction data from the rest of the interactions reduces the actual number of PPIs to 12,676 pairs for IntAct and 446,548 pairs for STRING. Negative PPI is extracted from the work that derives negative interaction from large-scale two-hybrid experiments [49]. Structure information for Struct2Graph is obtained from PDB files. Hence, we only used the pairs which have associated PDB files. This reduces the total number of pairs to 117,933 pairs (5580 positive and 112,353 negatives). Some proteins are well-studied as they are in the scope of current medical and biotechnological interest. As a result, there is more than one cross-reference to PDB files since various structures are accessible for these proteins. To find the proteins matched with PDB files, all proteins from the database are matched with UniProt accession numbers (UniProt Acc) and mapped with PDB files in UniProt [50]. Unfortunately, not all proteins are crystallized fully in each PDB file, and random choice of PDB file may cause incomplete information of the binding site of the protein. Therefore, we curated the PDB files based on the length of the chain ID and highest resolution within each PDB file to ensure that we capture the most complete structure information of the protein of interest. The chain length and the resolution of each protein's crystal structure were obtained from the RCSB website [51]. The complete set of negative pairs was reduced to 5580 pairs to create a balanced training sample with an equal number of positive and negative pairs. For this curated database consisting of only heterologous pairs, we defined two classes, "0" for non-interacting (negative) pairs and "1" for interacting (positive) pairs.

Mutual graph attention network for protein-protein pairs

We present a novel multi-layer mutual graph attention network (GAN) based architecture for PPI prediction task, summarized in Fig 1. We refer to this architecture as *Struct2Graph*, since the inputs to the proposed GAN are coarse grained structural descriptors of a query protein-protein pair. Struct2Graph outputs the probability of interaction between the query proteins. Struct2Graph uses two graph convolutional networks (GCNs) with weight sharing, and a mutual attention network to extract relevant geometric features related to query protein pairs. These extracted features are then concatenated and fed to a feedforward neural network (FNN) coupled with a `SoftMax` function, which finally outputs a probability of the two classes - '0' (non-interacting pairs) and '1' (interacting pairs). This section first describes the preprocessing and fingerprinting procedure specifying how spatial information on protein pairs are converted into corresponding protein graphs, and then elaborates on different components of the Struct2Graph deep learning architecture.

Protein structure graph

The purpose of the graph construction step is to capture the salient geometry of the proteins in a way that is amenable to further dimensionality reduction by the neural network. There are many possible ways of constructing a graph from spatial coordinates of individual atoms, and each captures a different level of detail about the geometry of

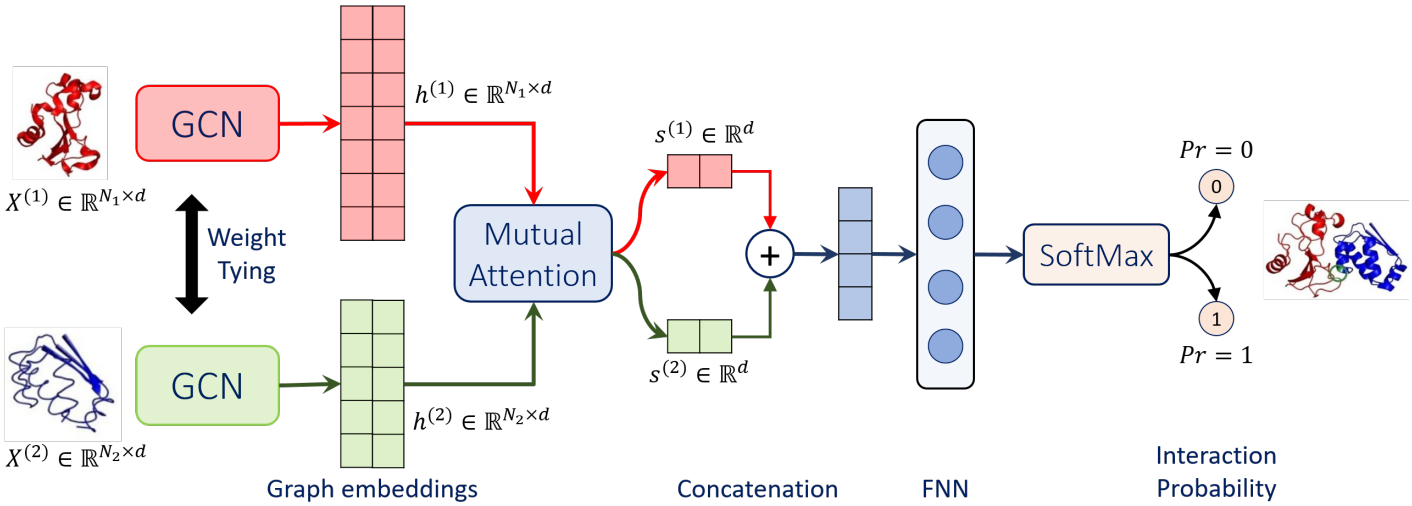


Fig 1. Struct2Graph schematic. Struct2Graph graph convolutional network (GCN) for incorporating mutual attention for PPI prediction. The GCN classifies whether or not a protein pair ($X^{(1)}$ and $X^{(2)}$ on far left) interacts and predicts the interaction sites (on far right).

the protein. We first aggregate atoms into the amino acids that they constitute and define the position of an amino acid to be the average of the positions of its constituent atoms. These amino acids form the vertices of the protein graph. An edge is placed between two vertices if the distance between them is less than some threshold. In this work, we use a threshold of 9.5Å for creating a protein graph from the mean positions of amino acids. This threshold was obtained empirically so as to render the underlying graph fully connected. Note that while we use amino acids as constituent vertices of the protein graphs, the approach can be easily extended to multiresolution representation, where a vertex represents two or more amino acids. The coarse-grained representation opens up new possibilities for studying other nanoscale materials, such as, lipids and polysaccharides, since, lowering the level of representation from all-atom to submolecular can be easily generalized to other non-protein entities. Graphs with greater structural refinement can also be obtained by using functional groups as of amino acids. Moreover, this geometric construction of protein graphs ensures that salient geometric features, such as spatial proximity of non-adjacent amino acids along the polypeptide chain are captured. A sequence based representation of proteins might not capture this geometrical structure as well (see Fig 2).

The graph construction approach converts spatial information associated with a protein into an equivalent protein graph object $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges between them. In the context of protein graph in Fig 2, $v_i \in \mathcal{V}$ is the ith amino acid and $e_{ij} \in \mathcal{E}$ represents an edge between i^{th} and j^{th} amino acids, satisfying their proximity within the specified threshold of 9.5Å. These graph objects must be embedded into real-valued vector space in order to employ our machine learning framework. We use 1-neighborhood subgraphs [44] induced by the neighboring vertices and edges at 1-hop distance from a vertex. A dictionary of all unique subgraphs is constructed by scanning all protein graphs in the training database. Thus, each vertex within a protein is equivalently represented by an element in the dictionary.

Graph convolutional network acting on protein graphs

A graph convolutional network (GCN) maps graphs to real-valued *embedding vectors* in such a way that the geometry of the embedding vectors reflects similarities between the

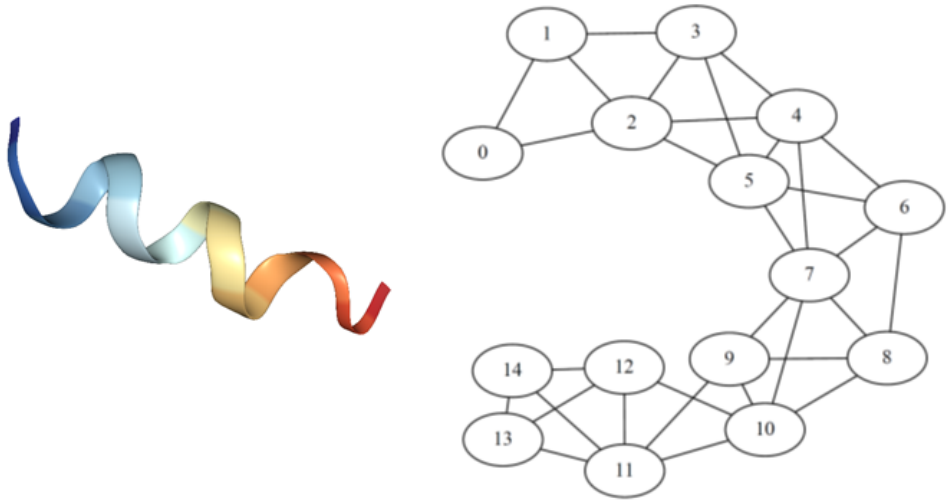


Fig 2. Protein and protein graph. Illustration of extracted protein structure graph (right) from the corresponding PDB description of a peptide segment (left) of the *S. cerevisiae* alpha-factor receptor. The graph is extracted by thresholding the distances between amino acids. The helical structure of the protein (left) gets captured in the corresponding protein graph (right) where, for example, amino acid 4 is linked with amino acid 7.

graphs. The embedding portion of the GCN works as follows. To each vertex $v_i \in \mathcal{V}$, we associate a d-dimensional feature vector, which encodes the 1-neighborhood subgraph induced by the neighboring vertices and edges at 1-hop distance from a vertex. This is in contrast to explicit inclusion of amino acid specific features, such as, hydrophobicity, solvent accessible surface area (SASA), charge, etc. In our encoding, similar to other studies [44, 52], each element of the dictionary of subgraphs is assigned a random unit-norm vector.

Each layer of the GCN updates all vertex features by first replacing each vertex feature by a normalized average over vertex features of all 1-hop neighboring vertices. This is followed by an affine transformation given by the trained weight matrices and bias parameters. In order to impart expressivity to the GCN architecture, each coordinate of the resulting affine transformed embedding vector is passed through a nonlinear activation function, such as, rectified linear unit (ReLU) or sigmoid activations. This process is repeated for all the subsequent layers, and the output of the final layer is the newly transformed embedding (feature) vector that is propagated further to the mutual attention network. Here, the number of layers is a hyperparameter, while the weight matrices are learned from the training data in order to optimize performance of the entire system on the interaction prediction task.

More concisely, given input protein graphs $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$ with adjacency matrices $A^{(1)}, A^{(2)}$ consisting of N_1, N_2 vertices (amino acids), and quantities $X_0^{(1)} \in \mathbb{R}^{N_1 \times d}$, $X_0^{(2)} \in \mathbb{R}^{N_2 \times d}$ representing the d-dimensional embedding of the vertex subgraphs of the query protein-protein pair, respectively, an l-layer GCN updates vertex embeddings using the following update rule:

$$X_{t+1}^{(m)} = \texttt{ReLU}\left(\tilde{A}^{(m)} X_t^{(m)} W_t\right), \quad \text{for all } t \in \{0, \cdots, l-1\}, \tag{1}$$

where $\tilde{A}^{(m)} = \left(\hat{D}^{(m)}\right)^{-\frac{1}{2}} \hat{A}^{(m)} \left(\hat{D}^{(m)}\right)^{-\frac{1}{2}}$ denotes the normalized adjacency matrices, and $m \in \{1, 2\}$. Here, $\hat{A}^{(m)} = A^{(m)} + I$ and $\hat{D}^{(m)}$ is the degree matrix of $\hat{A}^{(m)}$.

Parameters W_t denote the weight matrix associated with the tth-layer of the GCN. The feature embeddings $X_l^{(1)} \in \mathbb{R}^{N_1 \times d}$ and $X_l^{(2)} \in \mathbb{R}^{N_2 \times d}$ produced by the final layer of GCN are fed to a mutual attention network and hereafter denoted as $h^{(1)}$ and $h^{(2)}$, respectively, for notational convenience.

Mutual attention network for PPI prediction

The purpose of the proposed mutual attention network is two fold: (a) extract relevant features for the query protein-protein pair that *mutually* contribute towards prediction of physical interaction of proteins, (b) combine embedding matrices of dissimilar dimensions $N_1 \times d$ and $N_2 \times d$ to produce a representative single output embedding vector of dimension $(2d)$. Attention mechanisms were originally introduced for interpreting sequence-to-sequence translation models by allowing the models to attend differently to different parts of the encoded inputs. Since then, it has been adapted in other fields of deep learning, such as, computer vision [53], and bioinformatics [52].

The mutual attention mechanism proposed in this work computes attention weights $[\alpha_{ij}] \in \mathbb{R}^{N_1 \times N_2}$ and context vectors $s^{(1)} \in \mathbb{R}^d$, $s^{(2)} \in \mathbb{R}^d$ from the GCN-transformed hidden embeddings $h^{(1)}$ and $h^{(2)}$. The attention weights are computed as:

$$\alpha_{ij} = w^{\intercal} \tanh\left(U h_i^{(1)} + V h_i^{(2)}\right), \tag{2}$$

where $U, V \in \mathbb{R}^{d \times d}$ and $w \in \mathbb{R}^d$ are parameters of the mutual attention network that are trained in an end-to-end fashion along with the weights of the GCN. These attention weights are then translated to context vectors $s^{(1)}, s^{(2)}$ using the following knowledge selection procedure:

$$\begin{aligned}
\beta_i^{(1)} &= \frac{1}{N_2} \sum_{j=1}^{N_2} \alpha_{ij}, & \beta_j^{(2)} &= \frac{1}{N_1} \sum_{i=1}^{N_1} \alpha_{ij} \\
p_i^{(1)} &= \frac{\exp\left(\beta_i^{(1)}\right)}{\sum_{k=1}^{N_1} \exp\left(\beta_k^{(1)}\right)}, & p_j^{(2)} &= \frac{\exp\left(\beta_j^{(2)}\right)}{\sum_{k=1}^{N_2} \exp\left(\beta_k^{(2)}\right)}. \\
s^{(1)} &= \sum_{i=1}^{N_1} p_i^{(1)} h_i^{(1)}, & s^{(2)} &= \sum_{j=1}^{N_2} p_j^{(2)} h_j^{(2)}
\end{aligned} \tag{3}$$

Here $p_i^{(1)}$ and $p_j^{(2)}$ denote the relative weights of the amino acids i and j of the query protein-protein pair that contribute towards interaction prediction. Those vertices whose learned attention weights are large are likely to represent potential interaction sites between the query proteins.

The context vectors $s^{(1)}$ and $s^{(2)}$ are then concatenated into a single context vector of dimensions $2d$, which is used as input to a single-layer, fully connected feedforward neural network (FNN) represented by $f(\cdot)$ to produce a two-dimensional output vector. The FNN is parameterized by another weight matrix to be learned in an end-to-end manner. A final `SoftMax` layer is applied to produce a probability, one for each of the possible classes: 0 or 1, as shown in Equation (4). This output represents the classifier's prediction of the probability that the two input proteins interact.

$$y_{\text{out}} = \texttt{SoftMax}\left(f\left(\texttt{concat}\left[s^{(1)}, s^{(2)}\right]\right)\right) \tag{4}$$

Results

As part of our assessment, we compare the performance of Struct2Graph for PPI predictions against a number of recent machine learning models. These methods

include: (a) DeepFE model [26], where we train the natural language processing network on the same database used in the original publication and feed the embeddings into a fully connected feedforward neural network. (b) DeepPPI [25], where we extract 1164 sequence features related to the amino acid composition, distribution, and order. A separate neural network is used for each protein in the protein-protein pair and their outputs are concatenated into a final network for classification. Furthermore, as was done in the original publication [25], we implement these features into a number of traditional machine learning models [54], such as (c) Gaussian naive Bayes (GaussianNB) classifier, (d) Quadratic discriminant analysis (QDA), (e) k-nearest neighbor (k-NN) classifier, (f) Decision tree (DT) classifier, (g) Random forest (RF) classifier, (h) Adaboost classifier, and (i) Support vector classifier (SVC). All models are implemented in Python 3.6.5 on an Intel i7-7700HQ CPU with 2.8GHz x64-based processor. For common machine learning classifiers, such as, GaussianNB, QDA, SVC, RF, DT, k-NN and Adaboost, we use the readily available implementation in the scikit-learn [54] module. Deep learning classifiers, in particular, DeepPPI [55] and DeepFE-PPI [56] are implemented in Keras [57], while Struct2Graph is implemented in PyTorch [58].

For Struct2Graph, the hyperparameters of the models are tuned in order to achieve the reported accuracies. The tuning is obtained by performing grid search over the set of possible hyperparameter settings. The hyperparameters of our Struct2Graph implementation are as follows: `optimizer`: Adam optimizer [59] with learning rate $\lambda = 10^{-3}$ and rate-decay of 0.5 per 10 epochs; `total epochs`: 50; `number of GCN layers`: $l = 2$; `GCN embedding dimension`: $d = 20$; `loss function`: binary cross-entropy. For other competing methods, we use the tuned hyperparameters that are adopted from the original publications.

Performance on balanced database

Table 1 summarizes the comparisons of Struct2Graph and various machine learning models for PPI prediction for a five-fold stratified cross validation study. In the cross validation, the 11360 pairs (5580 positive and 5580 negatives) are randomly partitioned into five subsamples of equal size. Of these five subsamples, a single subsample is retained as the validation data for testing various machine learning models, and the remaining four subsamples are used as training data. In order to reduce the training time with our Struct2Graph model, 800 pairs are randomly sampled with replacement among the 9088 pairs (80%) in each epoch, and the performance on the randomly chosen 800 pairs is used to update the parameters of the neural network. This modification not only reduces the training time considerably, but also injects noise into the training data to avoid any potential overfitting.

The performance is reported for various measures, such as, accuracy, precision, recall, specificity or the true negative rate, Matthews correlation coefficient (MCC), F_1-score, area under the receiver operating characteristic curve (ROC-AUC), and negative predictive value (NPV) (see Tables 1-5). For a balanced training set (Table 1), Struct2Graph outperforms any other existing machine learning models in the literature for all the measures (except for the recall and NPV scores) with an average accuracy and precision of 98.89% and 99.50%, respectively. This is despite the fact that we significantly downsample the number of pairs in each epoch during the training process of the proposed Struct2Graph model.

Note from Table 1 that while QDA outperforms Struct2Graph in terms of recall and NPV scores, it does very poorly in terms of other measures indicating that the QDA classifier largely predicts positive interactions resulting in low false negative counts. Another observation is that the performance of Struct2Graph is only slightly better than that of another deep learning PPI model, DeepFE-PPI for this balanced training

set. However, as discussed below, DeepFE-PPI does not perform as well for unbalanced training set, where positive interactions are underrepresented among all interactions, a case that often arises in practice.

Table 1. Five-fold cross-validation performance analysis of several machine learning methods on balanced dataset (1:1). Note that the proposed Struct2Graph method outperforms all other methods on the majority of metrics.

Method	Performance (%) on Balanced training set - 1:1			
	Accuracy	Precision	Recall	Specificity
GaussianNB	72.14±2.91	98.41±0.51	45.05±6.10	99.24±0.30
QDA	78.66±3.44	70.43±3.41	**99.42±0.40**	57.90±7.12
$k-$NN	94.19±0.56	99.49±0.08	88.83±1.10	99.54±0.07
Decision Trees	96.20±0.43	97.59±0.28	94.75±0.99	97.66±0.29
Random Forest	98.86±0.29	99.45±0.19	98.27±0.49	99.45±0.19
Adaboost	97.85±0.26	98.76±0.18	96.92±0.51	98.78±0.18
SVC	98.49±0.33	99.44±0.18	97.53±0.61	99.45±0.18
DeepPPI	97.22±0.44	98.26±0.82	96.14±0.88	98.29±0.83
DeepFE-PPI	98.64±0.32	99.16±0.28	98.12±0.51	99.17±0.28
Struct2Graph	**98.89±0.24**	**99.50±0.36**	98.37±0.34	**99.45±0.42**
Method	MCC	F1-score	ROC-AUC	NPV
GaussianNB	52.69±4.38	61.53±6.00	72.15±2.91	64.46 ± 2.37
QDA	63.06±5.23	82.40±2.27	78.66±3.43	**99.05**±0.58
$k-$NN	88.89±1.02	93.86±0.63	94.19±0.56	89.92±0.89
Decision Trees	92.45±0.84	96.15±0.46	96.20±0.43	95.23±0.61
Random Forest	97.74±0.58	98.86±0.30	98.86±0.29	98.30±0.46
Adaboost	95.72±0.52	97.83±0.27	97.85±0.26	96.94±0.50
SVC	97.01±0.66	98.48±0.34	98.49±0.33	97.58±0.59
DeepPPI	94.47±0.87	97.19±0.44	99.28±0.11	96.23±0.81
DeepFE-PPI	97.29±0.64	98.64±0.32	99.52±0.09	98.14±0.50
Struct2Graph	**97.79±0.49**	**98.94±0.20**	**99.55±0.16**	98.24±0.42

Performance on unbalanced databases

In most practical scenarios, the number of negative pairs is expected to be larger than positive pairs, since only a small fraction of protein pairs interact within all possible pairs. We thus evaluate the performance of the deep learning models, Deep-PPI and DeepFE-PPI against the proposed Struct2Graph model on various unbalanced training sets, where the number of negative pairs outnumber the positive pairs. These results are summarized in Tables 2-5 for several databases with varying ratios of positive to negative pairs: (a) 1:2 (2790 positive and 5580 negative), (b) 1:3 (1860 positive and 5580 negative), (c) 1:5 (1136 positive and 5580 negative), and (d) 1:10 (558 positive and 5580 negative). Note that the positive pairs for unbalanced databases are selected randomly from the set of curated positive pairs. Struct2Graph again outperforms its deep-learning counterparts consistently for this unbalanced case. Struct2Graph improvement increases when the ratio between positive and negative pairs becomes increasingly skewed. For instance, when the ratio of positive and negative pairs is 1:10, the precision and recall statistics for the Struct2Graph model are 97.54% and 96.43%, respectively, which are higher by 0.98% and 2.14%, respectively than the performance of the next best deep-learning model, DeepFE-PPI.

Table 2. Five-fold cross-validation performance analysis of deep-learning based machine learning methods on unbalanced dataset (1:2).

Method	Performance (%) on Balanced training set - 1:2			
	Accuracy	Precision	Recall	Specificity
DeepPPI	97.40±0.44	98.64±0.61	93.52±1.64	99.35±0.30
DeepFE-PPI	98.91±0.09	99.00±0.32	97.71±0.33	99.51±0.16
Struct2Graph	**99.03±0.24**	**99.13±0.25**	**98.11±0.58**	**99.53±0.13**
Method	MCC	F1-score	ROC-AUC	NPV
DeepPPI	94.16±0.97	96.00±0.72	99.19±0.21	96.85±0.76
DeepFE-PPI	97.54±0.20	98.35±0.13	**99.56±0.08**	**99.86±0.16**
Struct2Graph	**97.87±0.51**	**98.62±0.32**	99.47±0.20	98.97±0.34

Table 3. Five-fold cross-validation performance analysis of deep-learning based machine learning methods on unbalanced dataset (1:3).

Method	Performance (%) on Balanced training set - 1:3			
	Accuracy	Precision	Recall	Specificity
DeepPPI	98.19±0.58	98.73±0.40	93.98±2.43	99.59±0.13
DeepFE-PPI	98.96±0.27	98.30±0.46	97.52±0.88	99.44±0.15
Struct2Graph	**99.30±0.22**	**99.17±0.44**	**98.19±1.09**	**99.71±0.13**
Method	MCC	F1-score	ROC-AUC	NPV
DeepPPI	95.15 ± 1.55	96.28 ± 1.24	99.27±0.14	98.03±0.78
DeepFE-PPI	97.21±0.72	97.90±0.55	**99.51±0.11**	99.18±0.29
Struct2Graph	**98.20±0.56**	**98.67±0.41**	99.49±0.21	**99.33±0.38**

Table 4. Five-fold cross-validation performance analysis of deep-learning based machine learning methods on unbalanced dataset (1:5).

Method	Performance (%) on Balanced training set - 1:5			
	Accuracy	Precision	Recall	Specificity
DeepPPI	97.78±0.45	98.33±0.39	88.20±2.74	**99.70±0.07**
DeepFE-PPI	98.97±0.27	98.19±0.49	95.60±1.52	99.65±0.10
Struct2Graph	**99.13±0.18**	**98.49±0.85**	**96.63±0.93**	99.68±0.19
Method	MCC	F1-score	ROC-AUC	NPV
DeepPPI	91.87±1.68	92.97±1.53	98.69±0.50	97.69±0.52
DeepFE-PPI	96.28±1.00	96.87±0.85	**99.56±0.25**	99.12±0.30
Struct2Graph	**97.03±0.56**	**97.55±0.45**	99.17±0.25	**99.26±0.23**

Discussion

Struct2Graph can predict PPIs solely from 3D structural information and outperforms other existing machine learning models with an average accuracy (98.89%) and precision (99.50%). The success of Struct2Graph is attributed to the use of structural 3D information embedded in the form of a graph, which describes chemical interactions better than sequence-based approaches. In addition to these predictions, Struct2Graph can further identify likely interaction sites of the specific protein-protein complex. This is achieved by considering the probability tuples $\{(p_i, p_j)\}$ of different amino acids

Table 5. Five-fold cross-validation performance analysis of deep-learning based machine learning methods on unbalanced dataset (1:10).

Method	Performance (%) on Balanced training set - 1:10			
	Accuracy	Precision	Recall	Specificity
DeepPPI	98.24±0.49	95.83±2.60	84.33±4.23	99.63±0.23
DeepFE-PPI	99.17±0.33	96.56±1.09	94.19±2.87	99.67±0.10
Struct2Graph	**99.42±0.14**	**97.54±1.28**	**96.43±2.49**	**99.73±0.16**
Method	MCC	F1-score	ROC-AUC	NPV
DeepPPI	88.95±3.14	89.66±2.94	97.18±1.24	98.45±0.41
DeepFE-PPI	94.91±2.07	95.35±1.90	**99.48±0.32**	99.42±0.29
Struct2Graph	**96.65±1.12**	**96.96±1.07**	99.45±0.70	**99.63±0.22**

during the knowledge selection process described in Equation (3). These probabilities capture the relative importance of amino acids and how they interact, and thus reflect the contributions of different amino acids towards interaction prediction. Amino acids with large relative probabilities are likely to participate in the interaction process as potential interaction sites.

We validated the results of Struct2Graph prediction for interaction sites of protein pairs with the literature [60]. *Staphylococcus aureus* is a Gram-positive bacteria and one of the most common causes of human bacterial infections worldwide. Phenol-soluble modulins (PSMs), short, amphipathic, α-helical peptides [61], play a crucial role in *S. aureus* virulence [62]. *S. aureus* has seven PSMs (PSM$\alpha_1-\alpha_4$, PSM$\beta_1-\beta_2$, and δ-toxin) which have multiple functions including, cytolysis, biofilm structuring, and inflammatory activation via cytokine release and chemotaxis. Specifically, PSMs trigger release of high mobility group box-1 protein (HMGB1). Toll-like receptor-4 (TLR4) interacts with HMGB1 activating nuclear factor NF-κB and the production of proinflammatory cytokines [63]. However, *S. aureus* PSMs$\alpha_1-\alpha_3$ significantly inhibit HMGB1-mediated phosphorylation of NF-κB by competing with HMGB1 via interactions with the same surrounding residues of TLR4 domain [60]. As such, the specific interacting residues for these pairs (HMGB1:TLR4 and HMGB1:PSMα_1) have been well described [60].

Struct2Graph identifies 20 residues between sites 82 and 375 on TLR4 (PDB ID: 3FXI) as the likely interaction site of HMGB1 (PDB ID: 2LY4) with 9 of the top 10 predictions falling between residues 328 and 352. While Struct2Graph includes some lower-numbered residues in its prediction of the interaction site of HMGB1 on TLR4, the suggested interaction site shares overlap with the reported domain of residues 336-477, especially with Glu^{336}, Arg^{355}, Phe^{377}. The predicted top 20 interaction residues on HMGB1 match the indicated residues between sites 16 and 88, with 9 of the top 10 predictions falling between residues 36 and 83. Consistent with the competitive binding of HMGB1 and PSMα_1 with TLR4, Struct2Graph reveals PSMα_1 (PDB ID: 5KHB) is interacting within the same domain of HMGB1 as TLR4. Fig 3 shows the residues predicted to be essential and highlights how Struct2Graph predicts a similar site for both interactions. Moreover, Struct2Graph and previously reported MD simulations show PSMα_1 shares interactions at Gly^2 and Val^{10} within the range of indicated residues 2-17. These results underscore that Struct2Graph is capable of identifying residues critical to protein interactions without any training data on the specific nature of these interactions. A complete summary of the residues identified by Struct2Graph learning for these interactions is reported in the supporting information. It must be noted that TLR4, PSMα_1 and HMGB1 were not included in the training set. Struct2Graph not only accurately predicts their (binary) interactions, but also correctly

elucidates the likely interaction sites.

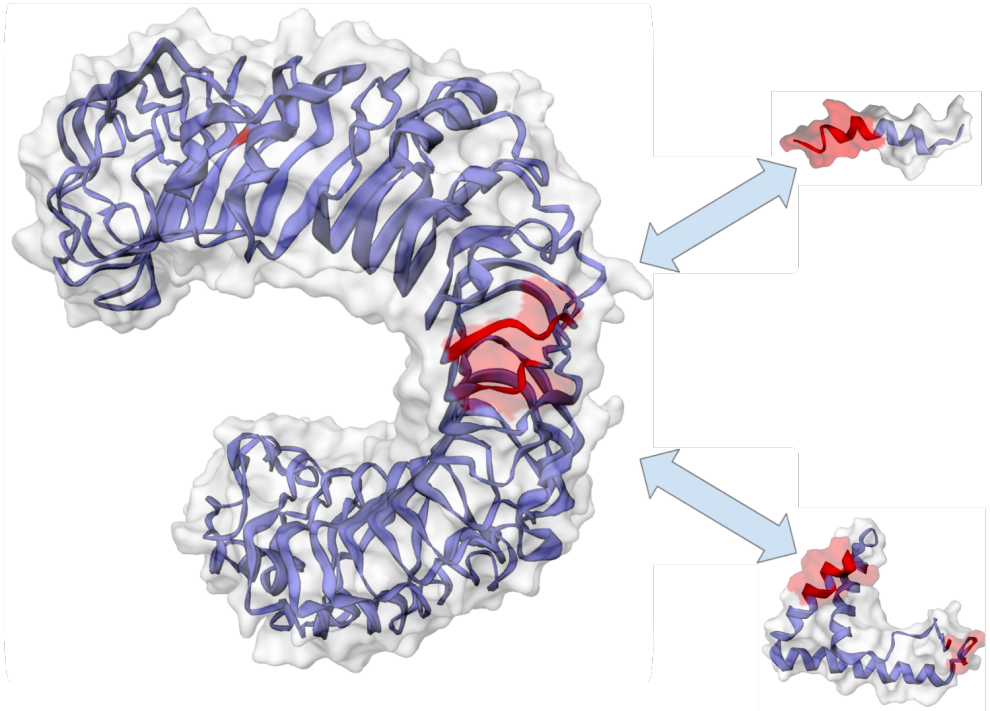


Fig 3. Predicted interaction sites for haemoglobin protein. A comparison of Struct2Graph predicted interaction sites for HMGB1 (left)interacting with TLR4 (top right) and PSM$-\alpha_1$(bottom right) interacting with TLR4. The residues predicted to be most important for interactions are shown in red. Arrows indicate predicted interaction sites.

Conclusion

Struct2Graph, a GCN-based mutual attention classifier, to accurately predict interactions between query proteins exclusively from 3D structural data is proposed. It is worth noting that Struct2Graph does not directly use descriptors, such as sequence information, hydrophobicity, surface charge and solvent accessible surface area, and thus can be generalized to a broader class of nanoscale structures that can be represented in similar fashion. This study demonstrates that a relatively low-dimensional feature embedding learned from graph structures of individual proteins outperforms other modern machine learning classifiers based on global protein features. Our GCN-based classifier achieves state-of-the-art performance on both balanced and unbalanced datasets.

Moreover, the mutual attention mechanism provides insights into likely interaction sites through its knowledge selection process in a completely unsupervised manner. The interaction sites predicted by Struct2Graph for PSMs are in consensus with the previously reported binding sites for this family. This connection between the unsupervised discovery of interaction sites and graph representation of proteins is possible thanks to the somewhat limited type of atoms and bond patterns that commonly occur in such molecules, which makes it possible to characterize properties on local atomistic arrangements. Overall, the proposed framework is quite general and,

while subject to availability of corresponding training data, can be made to predict other kinds of complex sets of collective supramolecular interactions between proteins and nanoscale species of different chemical composition.

Supporting information

S1 File. — Supporting Information.pdf Detailed information on interaction sites. File containing information on Struct2Graph attention-based ranking of the amino acid residues for HMGB1-TLR4 and HMGB1-PSMs-α_1 interactions.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Mayank Baranwal: Methodology, Software, Analysis, Writing - original draft. **Abram Magner:** Methodology, Software, Analysis. **Jacob Saldinger:** Validation of PPI results, Comparison with existing methods. **Emine S. Turali-Emre:** Data curation, Validation, Writing - PPI database and interaction site prediction sections. **Shivani Kozarekar:** Data curation. **Paolo Elvati:** Conceptualization, Writing - review & editing. **J. Scott VanEpps:** Writing - review & editing, Supervision - PPI database. **Nicholas A. Kotov:** Methodology - graph representation of proteins and other nanostructures, Writing - review & editing, Supervision - PPI database. **Angela Violi:** Conceptualization, Writing - review & editing, Supervision. **Alfred O. Hero:** Conceptualization, Writing - review & editing, Supervision.

Acknowledgments

The authors acknowledge the support from the BlueSky Initiative from the University of Michigan College of Engineering and grants from ARO W911NF-19-1-0269, ARO W911NF-14-1-0359 and DARPA HR00111720067. N.A.K expresses gratitude to Vannewar Bush DoD Fellowship ONR N000141812876.

References

1. Berggård T, Linse S, James P. Methods for the Detection and Analysis of Protein–Protein Interactions. PROTEOMICS. 2007;7(16):2833–2842. doi:10.1002/pmic.200700131.

2. Braun P, Gingras AC. History of Protein-Protein Interactions: From Egg-White to Complex Networks. PROTEOMICS. 2012;12(10):1478–1498. doi:10.1002/pmic.201100563.

3. Phizicky EM, Fields S. Protein-Protein Interactions: Methods for Detection and Analysis. Microbiological reviews. 1995;59(1):94–123. doi:10.1128/MMBR.59.1.94-123.1995.

4. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome. Proceedings of the National Academy of Sciences. 2001;98(8):4569–4574. doi:10.1073/pnas.061034498.

5. Fry DC. Protein–Protein Interactions as Targets for Small Molecule Drug Discovery. Biopolymers. 2006;84(6):535–552. doi:10.1002/bip.20608.

6. D Coelho E, P Arrais J, Luis Oliveira J. From Protein-Protein Interactions to Rational Drug Design: Are Computational Methods Up to the Challenge?. Current Topics in Medicinal Chemistry. 2013;13(5):602–618. doi:10.2174/1568026611313050005.

7. Mashaghi S, Jadidi T, Koenderink G, Mashaghi A. Lipid nanotechnology. International journal of molecular sciences. 2013;14(2):4242–4282.

8. Peppas NA, Huang Y. Nanoscale technology of mucoadhesive interactions. Advanced drug delivery reviews. 2004;56(11):1675–1687.

9. Lee SM, Nguyen ST. Smart nanoscale drug delivery platforms from stimuli-responsive polymers and liposomes. Macromolecules. 2013;46(23):9169–9180.

10. Meng H, Nel AE. Use of nano engineered approaches to overcome the stromal barrier in pancreatic cancer. Advanced drug delivery reviews. 2018;130:50–57.

11. Kotov NA. Inorganic nanoparticles as protein mimics. Science. 2010;330(6001):188–189.

12. Bhandari S, Mondal D, Nataraj S, Balakrishna RG. Biomolecule-derived quantum dots for sustainable optoelectronics. Nanoscale Advances. 2019;1(3):913–936.

13. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. A Comprehensive Analysis of Protein–Protein Interactions in Saccharomyces Cerevisiae. Nature. 2000;403(6770):623–627. doi:10.1038/35001009.

14. Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. Nature. 2002;415(6868):141–147. doi:10.1038/415141a.

15. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, et al. Systematic Identifcation of Protein Complexes in Saccharomyces Cerevisiae by Mass Spectrometry. Nature. 2002;415:4.

16. Sprinzak E, Sattath S, Margalit H. How Reliable Are Experimental Protein–Protein Interaction Data? Journal of Molecular Biology. 2003;327(5):919–923. doi:10.1016/S0022-2836(03)00239-0.

17. Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational Prediction of Protein–Protein Interactions. Molecular Biotechnology. 2008;38(1):1–17. doi:10.1007/s12033-007-0069-2.

18. Kaake RM, Wang X, Huang L. Profiling of protein interaction networks of protein complexes using affinity purification and quantitative mass spectrometry. Molecular & cellular proteomics. 2010;9(8):1650–1665.

19. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science. 1999;285(5428):751–753. doi:10.1126/science.285.5428.751.

20. Sun J, Li Y, Zhao Z. Phylogenetic Profiles for the Prediction of Protein–Protein Interactions: How to Select Reference Organisms? Biochemical and Biophysical Research Communications. 2007;353(4):985–991. doi:10.1016/j.bbrc.2006.12.146.

21. Pazos F, Valencia A. Similarity of Phylogenetic Trees as Indicator of Protein–Protein Interaction. Protein Engineering, Design and Selection. 2001;14(9):609–614. doi:10.1093/protein/14.9.609.

22. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting Protein-Protein Interactions Based Only on Sequences Information. Proceedings of the National Academy of Sciences. 2007;104(11):4337–4341. doi:10.1073/pnas.0607879104.

23. Guo Y, Yu L, Wen Z, Li M. Using Support Vector Machine Combined with Auto Covariance to Predict Protein–Protein Interactions from Protein Sequences. Nucleic Acids Research. 2008;36(9):3025–3030. doi:10.1093/nar/gkn159.

24. Mukherjee S, Zhang Y. Protein-protein complex structure predictions by multimeric threading and template recombination. Structure. 2011;19(7):955–966.

25. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks. Journal of Chemical Information and Modeling. 2017;57(6):1499–1510. doi:10.1021/acs.jcim.7b00028.

26. Yao Y, Du X, Diao Y, Zhu H. An Integration of Deep Learning with Feature Embedding for Protein–Protein Interaction Prediction. PeerJ. 2019;7:e7126. doi:10.7717/peerj.7126.

27. Rastogi C, Rube HT, Kribelbauer JF, Crocker J, Loker RE, Martini GD, et al. Accurate and sensitive quantification of protein-DNA binding affinity. Proceedings of the National Academy of Sciences. 2018;115(16):E3692–E3701.

28. Khandare J, Calderon M, Dagia NM, Haag R. Multifunctional dendritic polymers in nanomedicine: opportunities and challenges. Chemical Society Reviews. 2012;41(7):2824–2848.

29. Cha SH, Hong J, McGuffie M, Yeom B, VanEpps JS, Kotov NA. Shape-dependent biomimetic inhibition of enzyme by nanoparticles and their antibacterial activity. ACS nano. 2015;9(9):9097–9105.

30. Kadiyala U, Turali-Emre ES, Bahng JH, Kotov NA, VanEpps JS. Unexpected insights into antibacterial activity of zinc oxide nanoparticles against methicillin resistant Staphylococcus aureus (MRSA). Nanoscale. 2018;10(10):4927–4939.

31. Patra JK, Das G, Fraceto LF, Campos EVR, del Pilar Rodriguez-Torres M, Acosta-Torres LS, et al. Nano based drug delivery systems: recent developments and future prospects. Journal of nanobiotechnology. 2018;16(1):71.

32. Duncan R. Polymer conjugates as anticancer nanomedicines. Nature reviews cancer. 2006;6(9):688–701.

33. Bouffard E, El Cheikh K, Gallud A, Da Silva A, Maynadier M, Basile I, et al. Why Anticancer Nanomedicine Needs Sugars? Current medicinal chemistry. 2015;22(26):3014–3024.

34. Torrice M. Does nanomedicine have a delivery problem?; 2016.

35. Zamboni WC, Torchilin V, Patri AK, Hrkach J, Stern S, Lee R, et al. Best practices in cancer nanotechnology: perspective from NCI nanotechnology alliance. Clinical cancer research. 2012;18(12):3229–3241.

36. Fukuhara N, Kawabata T. HOMCOS: A Server to Predict Interacting Protein Pairs and Interacting Sites by Homology Modeling of Complex Structures. Nucleic Acids Research. 2008;36:185–189. doi:10.1093/nar/gkn218.

37. Ghoorah AW, Devignes MD, Smaïl-Tabbone M, Ritchie DW. Spatial Clustering of Protein Binding Sites for Template Based Protein Docking. Bioinformatics. 2011;27(20):2820–2827. doi:10.1093/bioinformatics/btr493.

38. Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y. MEGADOCK: An All-to-All Protein-Protein Interaction Prediction System Using Tertiary Structure Data. Protein & Peptide Letters. 2013;21(8):766–778. doi:10.2174/09298665113209990050.

39. Szilagyi A, Zhang Y. Template-Based Structure Modeling of Protein–Protein Interactions. Current Opinion in Structural Biology. 2014;24:10–23. doi:10.1016/j.sbi.2013.11.005.

40. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, et al. Structure-Based Prediction of Protein–Protein Interactions on a Genome-Wide Scale. Nature. 2012;490(7421):556–560. doi:10.1038/nature11503.

41. Northey TC, Barešić A, Martin ACR. IntPred: A Structure-Based Predictor of Protein–Protein Interaction Sites. Bioinformatics. 2018;34(2):223–229. doi:10.1093/bioinformatics/btx585.

42. Fout A, Byrd J, Shariat B, Ben-Hur A. Protein Interface Prediction Using Graph Convolutional Networks. Conference on Neural Information Processing Systems. 2017;31:6533–6542.

43. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. p. 1263–1272.

44. Baranwal M, Magner A, Elvati P, Saldinger J, Violi A, Hero AO. A deep learning architecture for metabolic pathway prediction. Bioinformatics. 2019;doi:10.1093/bioinformatics/btz954.

45. Jiang W, Qu Zb, Kumar P, Vecchio D, Wang Y, Ma Y, et al. Emergence of complexity in hierarchically organized chiral particles. Science. 2020;.

46. Lehne B, Schlitt T. Protein-protein interaction databases: keeping up with growing interactomes. Human genomics. 2009;3(3):291.

47. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic acids research. 2014;42(D1):D358–D363.

48. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research. 2019;47(D1):D607–D613.

49. Trabuco LG, Betts MJ, Russell RB. Negative protein–protein interaction datasets derived from large-scale two-hybrid experiments. Methods. 2012;58(4):343–348.

50. Bateman A. UNIPROT: A UNIVERSAL HUB OF PROTEIN KNOWLEDGE. In: PROTEIN SCIENCE. vol. 28. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA; 2019. p. 32–32.

51. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic acids research. 2019;47(D1):D464–D474.

52. Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics. 2018;35(2):309–318.

53. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning; 2015. p. 2048–2057.

54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825–2830.

55. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks; 2017. `https://github.com/gdario/deep_ppi`.

56. Yao Y, Du X, Diao Y, Zhu H. An integration of deep learning with feature embedding for protein–protein interaction prediction; 2019. `https://github.com/xal2019/DeepFE-PPI`.

57. Gulli A, Pal S. Deep learning with Keras. Packt Publishing Ltd; 2017.

58. Ketkar N. Introduction to pytorch. In: Deep learning with python. Springer; 2017. p. 195–208.

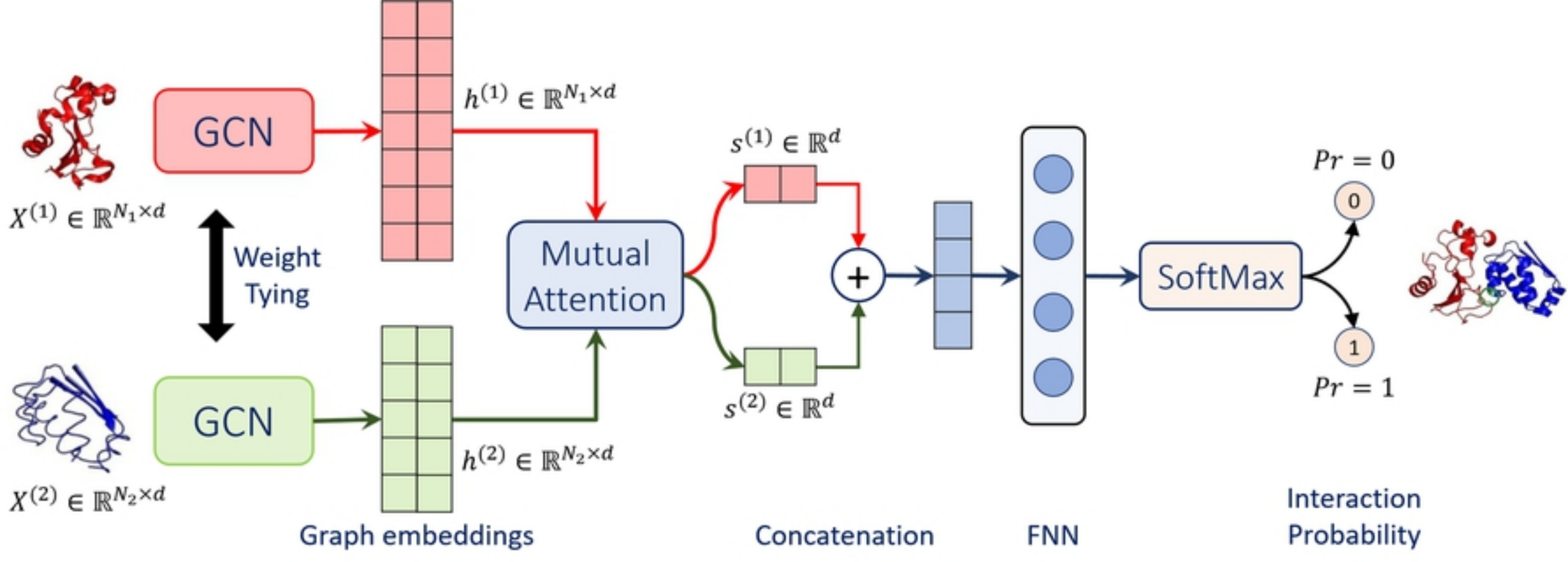
59. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

60. Chu M, Zhou M, Jiang C, Chen X, Guo L, Zhang M, et al. Staphylococcus aureus phenol-soluble modulins $\alpha 1$–$\alpha 3$ act as novel toll-like receptor (TLR) 4 antagonists to inhibit HMGB1/TLR4/NF-κB signaling pathway. Frontiers in Immunology. 2018;9:862.

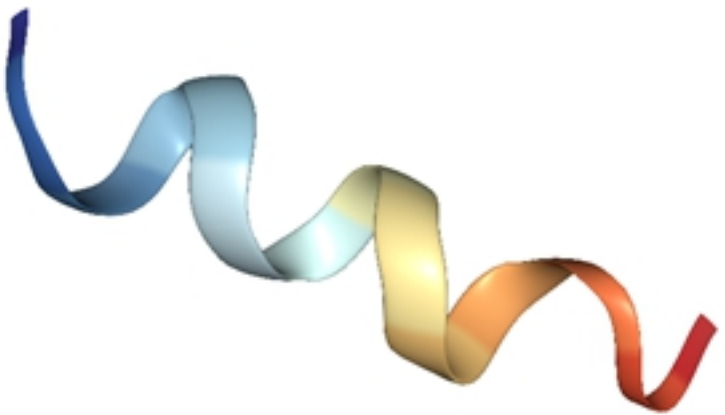
61. Mehlin C, Headley CM, Klebanoff SJ. An inflammatory polypeptide complex from Staphylococcus epidermidis: isolation and characterization. The Journal of experimental medicine. 1999;189(6):907–918.

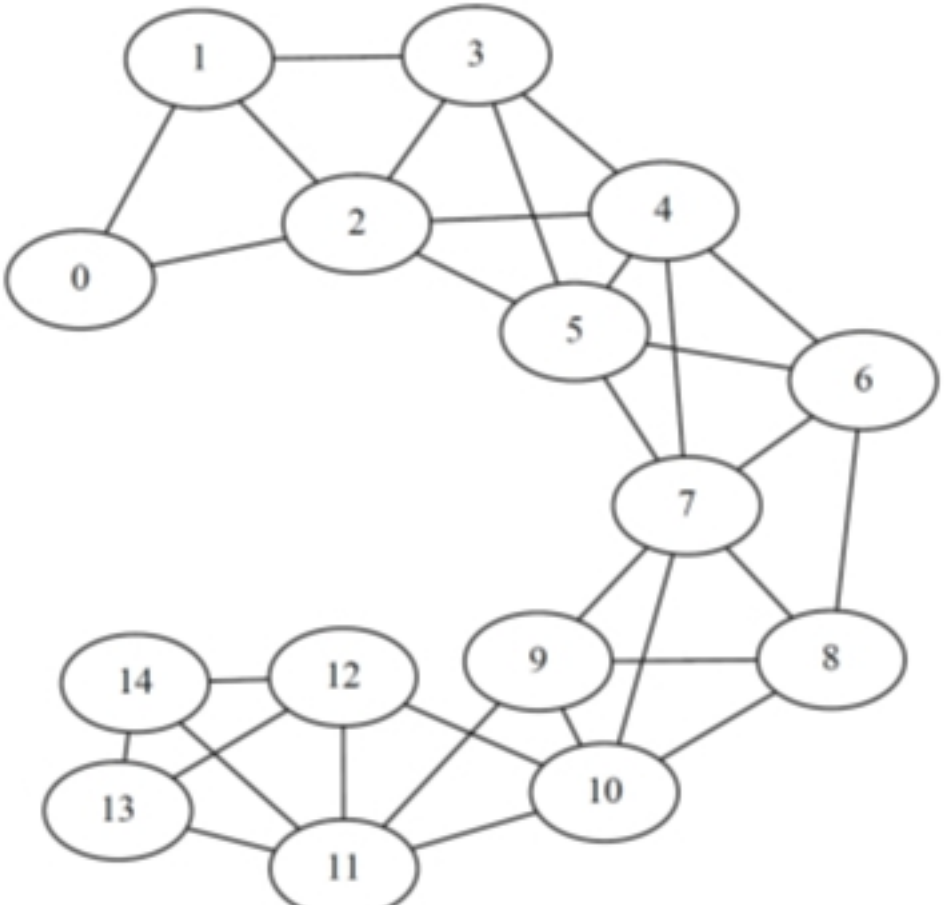
62. Tayeb-Fligelman E, Tabachnikov O, Moshe A, Goldshmidt-Tran O, Sawaya MR, Coquelle N, et al. The cytotoxic Staphylococcus aureus PSMα3 reveals a cross-α amyloid-like fibril. Science. 2017;355(6327):831–833.

63. Wang Y, Weng H, Song JF, Deng YH, Li S, Liu HB. Activation of the HMGB1-TLR4-NF-κB pathway may occur in patients with atopic eczema. Molecular Medicine Reports. 2017;16(3):2714–2720.

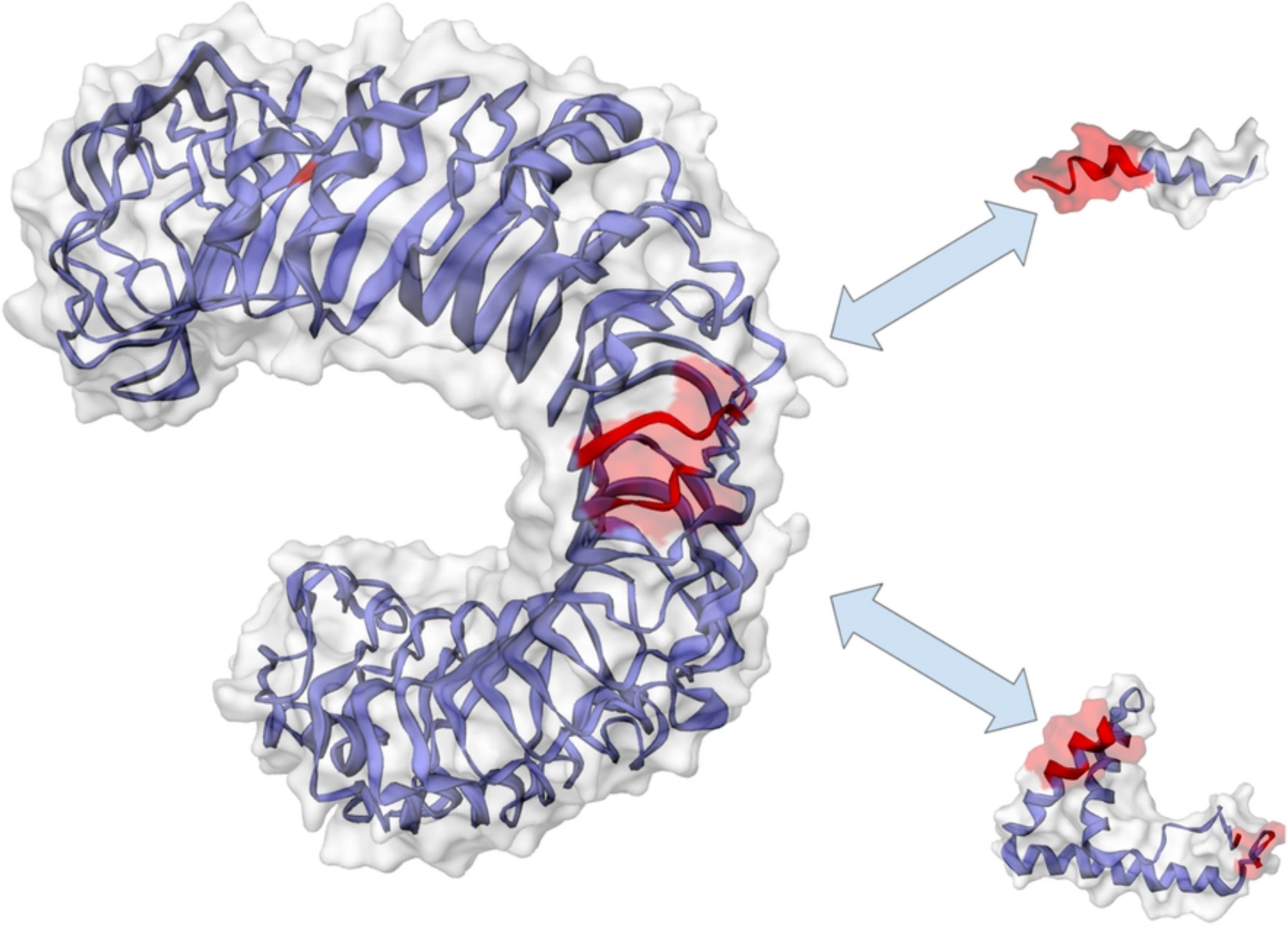

GCN
GCN
$X^{(1)} \in \mathbb{R}^{N_1 \times d}$
$X^{(2)} \in \mathbb{R}^{N_2 \times d}$
Weight Tying
$h^{(1)} \in \mathbb{R}^{N_1 \times d}$
$h^{(2)} \in \mathbb{R}^{N_2 \times d}$
Mutual Attention
$s^{(1)} \in \mathbb{R}^{d}$
$s^{(2)} \in \mathbb{R}^{d}$
+
SoftMax
$Pr = 0$
$Pr = 1$
0
1
Graph embeddings
Concatenation
FNN
Interaction Probability

Figure





Figure



Figure