scientific reports

Check for updates

OPEN

Decoding the protein–ligand interactions using parallel graph neural networks

Carter Knutson, Mridula Bontha, Jenna A. Bilbrey & Neeraj Kumar✉

Protein–ligand interactions (PLIs) are essential for biochemical functionality and their identification is crucial for estimating biophysical properties for rational therapeutic design. Currently, experimental characterization of these properties is the most accurate method, however, this is very time-consuming and labor-intensive. A number of computational methods have been developed in this context but most of the existing PLI prediction heavily depends on 2D protein sequence data. Here, we present a novel parallel graph neural network (GNN) to integrate knowledge representation and reasoning for PLI prediction to perform deep learning guided by expert knowledge and informed by 3D structural data. We develop two distinct GNN architectures: GNN_F is the base implementation that employs distinct featurization to enhance domain-awareness, while GNN_P is a novel implementation that can predict with no prior knowledge of the intermolecular interactions. The comprehensive evaluation demonstrated that GNN can successfully capture the binary interactions between ligand and protein's 3D structure with 0.979 test accuracy for GNN_F and 0.958 for GNN_P for predicting activity of a protein–ligand complex. These models are further adapted for regression tasks to predict experimental binding affinities and pIC_{50} crucial for compound's potency and efficacy. We achieve a Pearson correlation coefficient of 0.66 and 0.65 on experimental affinity and 0.50 and 0.51 on pIC_{50} with GNN_F and GNN_P, respectively, outperforming similar 2D sequence based models. Our method can serve as an interpretable and explainable artificial intelligence (AI) tool for predicted activity, potency, and biophysical properties of lead candidates. To this end, we show the utility of GNN_P on SARS-Cov-2 protein targets by screening a large compound library and comparing the prediction with the experimentally measured data.

Accurate prediction of protein–ligand interactions (PLI) is a critical step in therapeutic design and discovery. These interactions influence various molecular-level properties, such as substrate binding, product release, regio-selectivity, target protein function, and ability to facilitate potential hit identification, which is the first step in finding novel candidates for drug discovery[1]. With increases in computing power, code scalability, and advancement of theoretical methods, physics-based computational tools such as molecular dynamics and molecular/quantum mechanics can be used for the reliable representation of PLI and prediction of accurate binding free energies[2–4]. However, these methods are computationally expensive and are limited to a number of protein–ligand complexes[5]. This limits their routine use in high-throughput virtual screening[3,4] for the discovery of novel hit candidates and lead optimization[6] for a given protein target[7]. Molecular docking has been used to predict binding affinity and estimate interactions with reasonable computational cost[2,8–15]; however, its accuracy is relatively low as it uses heuristic rules to evaluate the scoring function.

The use of deep learning has revolutionized the healthcare system in recent years[16]. There has been significant effort to develop deep learning models that predict PLI[17,18], and other biophysical properties[19,20] that are critical for therapeutic design but cannot be predicted through physics based modeling[21]. The greater understanding of PLI enabled by deep learning can help in the estimation of properties such as activity, potency and binding affinity[22]. However, several technical challenges limit the use of deep learning for modeling protein–ligand complexes and accurate prediction of properties. The first challenge relates to the limited availability of protein–ligand 3D data and the second challenge focuses on the appropriate representation of the data (domain knowledge), specifically in terms of the comprehensive 3D geometry representation. Structure-based methods have the advantage of producing results that can be interpreted but are limited by the number of available samples[21]. Circular fingerprints, generated by encoding localized structural and geometric information, have long

Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99352, USA. ✉email: neeraj.kumar@pnnl.gov

been a cornerstone of cheminformatics[23]. The flexibility of fingerprints has created new avenues for molecular computational research, including the increased implementation of graph-based representations to include domain awareness[24,25]. Molecular graph representations provide a way to model and simulate the 3D chemical space while retaining a wider range of structural information. In that context, Ragoza et al. implemented a deep convolutional neural network (CNN) that operates directly on 3D molecular graph input, similar to the AtomNet model previously implemented by Walloch et al.[26,27]. Other approaches such as Graph-CNN developed by Torng et al. use unsupervised autoencoders to leverage sequence-based data that is more abundant but also costly in terms of structural accuracy[28].

Graph-based representations extend the learning of chemical data to graph neural networks (GNNs). Monti et al. designed a mixture model network (MoNet) that enables non-Euclidean data, such as graphs, to be learned by CNNs[29]. That approach has been generalized and improved through the formulation of graph attention (GAT) networks[30]. GAT architectures operate on the importance of a given node, leading to improved computational efficiency and accuracy[31]. Lim et al. showed the implementation of such an architecture to capture PLI, which provides a baseline for the development of more robust models[32]. Chen et al. proposed a bidirectional attention-driven, end-to-end GNN to predict PLI and enable biochemical insights through attention weight visualization[1]. Predicting the activity of a protein–ligand complex is a binary classification problem. Reshaping the problem to focus on affinity creates a regression problem of heightened complexity. The existing deep learning models that predict the binding affinity or other biophysical and biochemical properties such as IC_{50}, K_i, K_d, and EC_{50} [22,21,33]. However, most of the methods use sequence-based data for proteins and SMILES representations for interacting ligands. For example, DeepDTA[22] and DeepAffinity[21] use SMILES strings of the ligands and amino-acid sequences of the target proteins to predict the affinity. MONN[33] is a multi-objective sequence-based neural network model that first predicts the non-covalent interaction between the ligand and the residues of the interacting target and then the binding affinities in terms of IC_{50}, K_i, and K_d. Such methods are accessible due to the abundant availability of sequence-based data, but do not capture 3D structural information in the PLI and predicting regression properties. Binding is best understood when the 3D pocket of the target is known, and in situ, the protein–ligand complex is formed due to changes in the conformation of the 3D structure of the protein and ligand post-translation.

In this contribution, we formulated two GNNs based on the GAT architecture by incorporating domain-specific featurization of the protein and ligand atoms (GNN_F) and by implementing parallel GAT layers such that GNN_P uniquely learns the interaction with limited prior knowledge. The inclusion of different features on the protein and ligand atoms enables our models to be more physics informed. The implementation of GAT layers combined with our featurization enables the model to learn the representation and the chemical space of the training data. We further use these models to predict experimental binding affinity and pIC_{50} of the protein–ligand complex. This allows us to leverage the 3D structures of the target protein, ligands, and the interaction between them which is crucial both for the activity and affinity prediction.

Methods

Network architecture. The goal in this work is to define a GNN architecture that predicts characteristics of a protein–ligand pair by learning features of the protein and ligand that may not be obvious to the human observer. Our molecular graph structure is defined as G{V,E,A}, where V is the atomic node set, E is the corresponding edge set, and A is the adjacency matrix. Given the diverse structural properties of protein–ligand complexes, we include additional biomolecular domain-aware features to previous GAT architectures[30,32] by defining distinct featurizations for the protein and ligand components, as shown in Table 1, denoted GNN_F, and by removing the dependency on prior knowledge of the protein–ligand interaction through the implementation of parallel GAT layers, denoted GNN_P.

The GNN_F and GNN_P models differ in the architecture of the attention head as seen in Fig. 1, a schematic of the prediction logic implemented in our models. In GNN_F, the protein and ligand adjacency matrices are combined into a single matrix, and edges are added between protein and ligand nodes based on the distance matrix obtained from docking simulations. The GNN_F attention head uses a joined feature matrix for the ligand and target protein, which is passed into one GAT layer that learns attention based on the PLI adjacency matrix and a second GAT layer that learns attention based on the ligand adjacency matrix. The output of these two GAT layers are subtracted in the final step of each attention head.

In the absence of a co-crystal structure of a protein–ligand complex, docking simulations are typically performed to model the PLI. In GNN_P, the 3D structures of the protein and ligand are initially embedded separately based on their adjacency matrices, which represent internal bonding interactions. The GNN_P attention head passes separate features for the protein and ligand to individual GAT layers that learn attention based on the respective adjacency matrix. The outputs of the GAT layers are concatenated in the final step of the attention head. Separation of the ligand and protein in parallel GAT layers preformed by GNN_P removes prior information about the interactions, providing a foundation to remove the need for docked structural information. This discrete representation enables us to enter the protein and ligand directly into the GNN without knowing the prior protein–ligand interaction, which otherwise needed to be computed using physics-based simulations.

In both models, each node is given a set of features F, described in Table 1, which are engineered with an emphasis toward biochemical information, using the molecular Python package RDKit. When input into a GAT layer along with the corresponding adjacency matrix, each node feature is transformed by a learned weight matrix $\boldsymbol{W} \in \mathbb{R}^{F \times F}$, where F is the dimensions of the node features attributed to input represented by $\mathrm{h} = \{\hat{\mathrm{h}}_1, \hat{\mathrm{h}}_2, \ldots, \hat{\mathrm{h}}_N\}$ where $\hat{\mathrm{h}}_i \in \mathbb{R}^F$ and N is the number of atoms. The attention coefficient e_{ij} for interacting atoms i and j is calculated as a summation of the importance of the i-th node interaction with the j-th node and vice versa. For node i given an input feature matrix $\hat{h}_i$ at convolution layer l, the attention coefficient e_{ij} is given as:

Feature	Graph-CNN Torng and Altman[28]		GNN Lim et al.[32]		GNN_F & GNN_P (Current work)	
	Protein	Ligand	Protein	Ligand	Protein	Ligand
Atom type	x	x	x	x	x	x
Atom degree	x	x	x	x		x
N hydrogen atoms	x	x	x	x		x
Implicit valence	x	x	x	x		x
Aromaticity	x	x	x	x		x
Atom in ring						x
Residue type					x	
Hybridization						x
Formal charge						x
Single bond[a]		x				x[b]
Double bond[a]		x				x[b]
Triple bond[a]		x				x[b]
Bond aromaticity[a]		x				x[b]
Conjugation[a]		x				x[b]
Bond in ring[a]		x				x[b]

Table 1. Features associated with each atom in the protein and ligand for the Graph-CNN model described by Torng et al.[28], the GNN described by Lim et al.[32], and GNN_F and GNN_P described in this contribution. Features are associated with the atom unless otherwise noted. [a]Bond feature; [b]bond feature is indirectly considered by a corresponding atom-level feature that captures the same physical property.

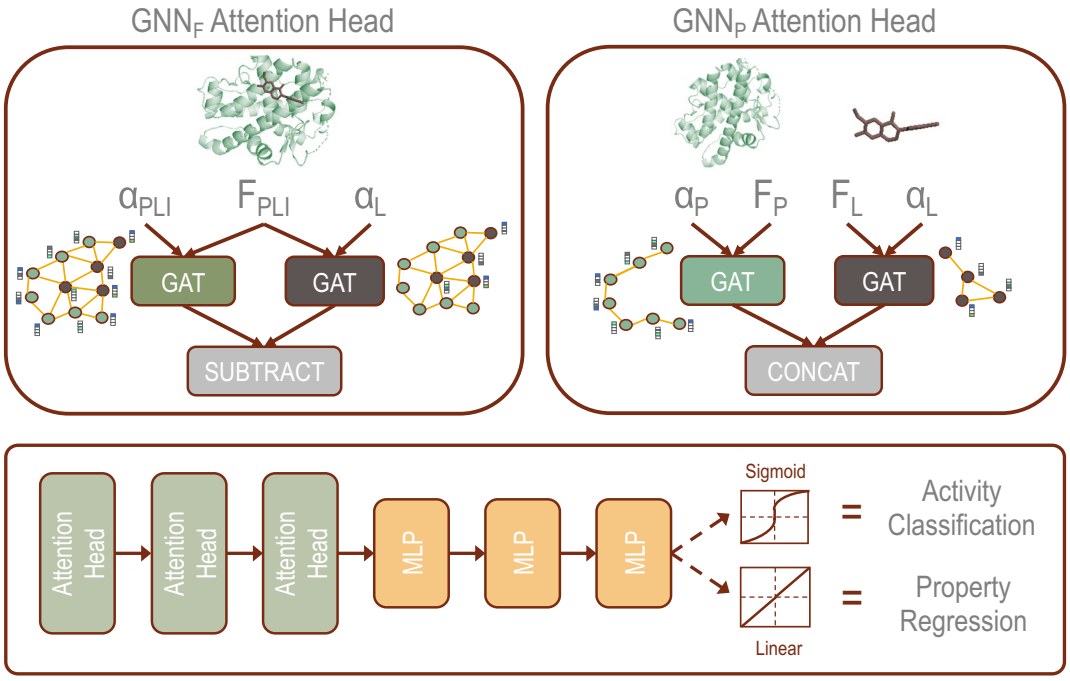

Figure 1. Schematic showing the prediction logic implemented in our GNN models. The two models differ based on the applied attention head. GNN_F uses the PLI obtained from docking simulations to create a combined feature and adjacency matrix. In GNN_P, the features for the ligand and target protein are coded separately alongside their corresponding adjacency matrices. The output from the attention head is passed through a series of MLP that can be tuned for activity classification through application of the sigmoid activation function and binary cross-entropy loss function or property regression through application of the linear activation function and mean squared error loss function.

Dataset	Targets	Total actives	Total in-actives
PDBbind2018	991	1418	4804
DUD-E	96	46,145	16,996,568
SARS-CoV-2	7	56,191	56,191

Table 2. Number of active complexes, inactive complexes, and protein targets in the PDBbind, DUD-E, IBS, and SARS-CoV-2 datasets used to create the training and test sets in this work.

$$e_{ij} = \hat{h}_i' \boldsymbol{W} \hat{h}_j' + \hat{h}_j' \boldsymbol{W} \hat{h}_i' \tag{1}$$

Using the softmax activation function, the attention coefficients are normalized across neighbors and multiplied by the adjacency matrix A_{ij}, which gives higher importance to node pairs closer in distance, reflecting the physical principle that the strength of an intermolecular bond decreases as the bond distance increases. The normalized attention coefficient a_{ij} is given by:

$$a_{ij} = softmax(e_{ij}) * A_{ij} \tag{2}$$

Each feature is then updated as a linear combination of neighboring node features $\hat{h}_i''$. Finally, the gated attention mechanism is employed to give the transformed set of node features: $\hat{h}_i'$:

$$\hat{h}_i' = \phi(\boldsymbol{U}(\hat{h}_i || \hat{h}_i'') + b)\hat{h}_i + (1 - \phi(\boldsymbol{U}(\hat{h}_i || \hat{h}_i'') + b))\hat{h}_i'' \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{2F \times 1}$ is a learned vector, b is a learned scalar, and ϕ is the activation function. These features are passed through a multilayer perceptron (MLP). For binary classification of the activity of the protein–ligand complex, the sigmoid activation function is applied, and the binary cross-entropy loss function is used. For regression, the ReLU activation function is applied, and the mean squared error loss function is used.

The distinct featurization of our GNNs reduces the feature size of the protein, which can contain a large number of atoms, and enhances the feature size of the ligand, which is typically a small molecule. This shift reduces redundancy in the protein representation and focuses computational resources on the improved atomic representation of the ligand. While the Graph-CNN developed by Torng et al. also involved a large number of ligand features, more than half were at the bond level[28]. Because our GNNs include only atom-level features, we examined the physical properties behind the chosen bond-level features and chose atom-level features that impart the same physical information. For example, Torng et al. included an encoding for whether a bond was in a ring; similarly, we included an encoding for whether an atom was in a ring. Torng et al. also had three different features corresponding to bond type (single bond, double bond, and triple bond); we consolidated these features into a single atom-level feature, hybridization, which describes the bonding properties of an atom.

Dataset preparation

Classification model datasets. Machine learning models for training PLI prediction require data on target proteins, ligands/compounds, and interactions between them. In this work, our goal is to improve the degree to which the graph-based model can be generalized while also maintaining accuracy. We accomplish this by enlarging the dataset used to train our model and including a variety of targets. We collected and curated protein–ligand complexes from two public datasets, DUD-E[34] and PDBbind[35,36], which are described in further detail below. The datasets consist of protein–ligand complexes and their docking affinities, while some samples include experimentally derived binding affinities. Prediction is based on accessing ligands as active (active-interact molecules or positive) or inactive (set of decoys or negative) depending on whether the ligand is able to bind with the protein, as described in more detail below. Table 2 shows counts of the targets, ligands, and their various complexes for both datasets.

To confirm that these two datasets offer diversity in terms of both target protein and different functionality of the ligand, we computed pair similarities between the targets and ligands, which can be found in the supporting information Figure SS1. Target similarities were determined by computing the homology between each pair, while ligand similarities were taken as the Dice similarity coefficient of the Morgan fingerprints (diameter = 6) of each pair. In both datasets, the target similarity is centered around 40%, with DUD-E having a more diverse target set than PDBbind. The opposite is true for ligand similarity. While both datasets have mainly dissimilar ligands, the PDBbind dataset offers more diversity in ligand structure. Combining these datasets for training leads to a highly diverse training set in terms of both target proteins and ligand molecules.

DUD-E. The DUD-E dataset consists of pairs of experimentally verified active complexes and property-matched inactive pairs, called decoys[34]. The dataset was originally designed to test benchmark molecular docking programs by providing challenging decoys, but some have noted that the dataset suffers from limited chemical space and biases[37,38]. In their analysis of the DUD-E dataset, Chen et al. generated a docked subset[38], which we use here. Ligands from the experimentally verified ChEMBL dataset were designated as positive, while the generated decoys and their docked structures were considered negative. These files were parsed into individual standard database formats (SDF) files for each ligand and corresponding docking pose.

Ligands with a molecular weight above 500 Da were removed, and the docked structures were converted to a machine-readable format using RDKit[39]. The target proteins were pulled from the original DUD-E set, cleaned of water molecules, and matched with their designated ligands. Targets were screened with a homology test to understand the diversity of targets present in the dataset, and none had a similarity greater than 90%. Each complex was processed to crop the target protein and retain atoms within a distance of 8 Å from the ligand. Because the DUD-E dataset is biased towards inactive complexes due to the large number of decoys, equal counts of active and inactive samples were collected in regards of a target with no consideration of a specific complex. A complete 1:1 active-to-inactive ratio was not always achieved; however, the imbalance was found to be minimal and to have no affect on performance. Samples from the experimentally verified ChEMBL dataset were labeled as active, while decoy samples were denoted with the key word ZINC.

PDBbind. The PDBbind dataset contains experimentally verified protein–ligand complexes from the Protein Data Bank[35,36]. Binding poses for a refined set of the protein–ligand complexes were generated by docking calculations[40]. A 90% homology test was run on this set, which resulted in a small number of proteins being removed because of a high level of similarity. As with the DUD-E dataset, the docked structures were converted to a machine-readable format using RDKit. The root-mean-square distance (RMSD) was used to label ligands as positive if they maintained an RMSD less than or equal to 2Åcompared to the original crystal structure, and negative if the RMSD was greater than or equal to 4Å. Molecules with RMSDs between these thresholds were removed. The viable molecules and their target proteins were processed into a dataset of protein atoms cropped within a distance of 8Å. As PDBbind data were significantly limited, all available samples were used. This resulted in a greater inclusion of negative samples than positive samples.

SARS-CoV-2 dataset. The SARS-CoV-2 dataset is composed of seven protein targets: M^{Pro}_6WQF, NSP15_6XDH, M^{Pro}_6LU7, PL^{Pro}_6W9C,PL^{Pro}_6WRH, ADRPNSP3_6W02, and NSP10-16_6W61 with three main protease (M^{Pro}, two papain-like cysteine protease, one open reading frame, and three non-structural proteins. Large ligand libraries composed from FDA[41] and manually curated antiviral data were used to generate the docked complexes with each target. Ten docking poses were calculated for each complex with the qvina docking program[42] through a custom non-covalent pipeline. The docking data was parsed and cropped to the 8 Å threshold in the same manner as the DUD-E and PBBind datasets. Similar to PDBbind, positive and negative samples are determined with RMSD. Protease data were largely directed into the training set while the other targets, with the exception of on non-structural protein, were directed to the test set. This variety allows the inclusion of the critical targets of SARS-CoV-2 viral life cycle in both the training and the test sets.

IBS dataset. The IBS dataset is created using 486,232 synthetic compounds from the IB screening database (www.ibscreen.com). All ligands were fed as an input to the our GNN models paired against M^{Pro} and NSP15 target. Because the IBS dataset only contains ligands and we did not dock the IBS molecules with their corresponding receptors, we used our GNN_P model to evaluate these complexes. We chose SARS-CoV-2 (M^{Pro}) and SARS-CoV-2 non-structural protein endoribonuclease (NSP15) as target proteins to study the performance of our model. These are exactly the same targets we have used for our creating our SARS-CoV-2 dataset. Our team has been actively workin1g on the development of covalent electrophiles and non-covalent inhibitor candidates against the viral proteases, this gave us ready access to the protein pockets and active compounds that were binding with these targets. In addition, we performed homology tests on M^{Pro} and NSP15 against the targets in the DUD-E and PDBbind training sets to quantify the similarity of these new targets to our larger dataset. M^{Pro} showed an average similarity of 40% with the DUD-E training targets and 48% with the PDBbind training targets. NSP15 showed an average similarity of 55.72% with DUD-E targets and 52.26% with PDBbind targets.

Regression model datasets. In this section, we discuss the datasets used for the regression models noting that part of these datasets are created from the same sources used for the classification models. Our regression models are composed of the same attention heads as in the classification models and, therefore, include the same domain-level information from the protein and ligand atoms. The main difference in our classification and regression models is the final activation layer, as shown in Fig. 1. We perform two regression experiments referred as Experimental Binding Affinity (EBA)and pIC_{50} prediction. It is important to highlight that some of these properties such as pIC_{50} cannot be accurately modeled through physics based modeling methods. Table 3 and 4 gives a summary of the target and ligands distribution in various regression datasets.

For Experimental Affinity experiments, we consider three data sources: (1) PDBbind2016, (2) PDBbind2018, and (3) PDBbind2019[36]. We consider just crystal poses from the PDBbind2016 general, refined, and core datasets. A PDBbind general set refers to the main body set of samples provided. Refined refers to the higher quality samples extracted from the general set, and the core set is a small sample set of the highest quality complexes. For the PDBbind2018 dataset, we used both the docked and crystal structures as two independent datasets for two independent experiments. For the docked dataset, we used the same targets and split as used for classification experiments. For the crystal-only dataset, we used all the available targets without any homology-similarity screening. For the PDBbind2019 dataset, we considered just a handful of crystal structures as a part of independent test set. The PDBbind2019 dataset is a structure-based evaluation set comprises targets that have been added to PDBbind2019 after the year 2016 and thus are not a part of PDBbind2016 and have a GDC similarity less than 65% to the training targets.

For the PDBbind2016 dataset, we prepared various train-test splits, which are summarized in Table SS5. The experimental binding affinity is estimated in terms of K_i and K_d, which refer to the inhibition and dissociation constants, respectively. These properties collectively determine the binding affinity of a molecule

Dataset	Total targets	Total ligands	Train targets	Test targets	Train samples	Test samples	Total samples
PDBbind2018-EBA (with docked poses)	1278	1278	1023	255	6485	1557	8042
PDBbind2018-EBA (crystal only)	10,375	10,375	8300	2075	8300	2075	10,375
PDBbind2016-EBA (crystal only)	11,674	11,674	[a]	[a]	[a]	[a]	11,654
PDBbind2019-EBA (crystal only)	190	190	–	–	–	–	190

Table 3. Summary of data used for training and testing EBA regression models. [a]Refer to Table SS5 in the supporting information for detailed dataset splits.

Dataset	Total samples	Train samples	Test samples
PDBbind2016-pIC_{50}	4576	3676	900
DUD-E-pIC_{50}-only-top-docked-pose	3706	3105	601

Table 4. Table summarizing the dataset used for pIC_{50} regression models.

PDB-ID	Target	Inhibitor
7LTJ[2]	M^{Pro}	MCULE-5948770040
7L0D[43]	M^{Pro}	ML188
7LME[44]	M^{Pro}	ML300
7L11[45]	M^{Pro}	Compound 5[aa]
7L12[45]	M^{Pro}	Compound 14[a]
7L13[45]	M^{Pro}	Compound 21[a]
7L14[45]	M^{Pro}	Compound 26[a]

Table 5. SARS-CoV-2 targets with experimentally assessed non-covalent inhibitors used for testing pIC_{50} models. [a]As defined by Zhang et al.[45].

towards a receptor. The PDBbind repository provides a database of protein–ligand complexes along with their experimentally measured data. Here, we define the experimental binding affinity as $-log(\frac{K_i}{K_d})$, which is used as the label in the regression model. All decoys from the PDBbind2018 dataset were labeled with the experimental affinity of the corresponding crystal structure.

We focus on the pIC_{50}(the inverse log of the half maxi-mal inhibitory concentration, IC_{50}), which is an experimentally measured property that captures the potency of a therapeutic candidate towards a protein target where higher values indicate exponentially more potent inhibitors. For the pIC_{50} data, we used a combination of the DUD-E and PDBbind2016 datasets. From DUD-E, we included active ligands for 65 of the DUD-E targets, as only active protein–ligand pairs have an associated experimental affinity in the ChEMBL repository. We considered the top-scoring docked pose for each protein–ligand complex in the DUD-E dataset because there were no crystal structures available.

For the PDBbind2016 data, all crystal poses with an experimentally measured IC_{50} were used. We retained the same 80:20 split for the PDBbind2016 dataset as used for the experimental affinity dataset. Table 4 gives details of the dataset used for pIC_{50} regression.

We also considered an independent dataset associated with SARS-CoV-2 targets. To prepare this dataset, we used crystal structures of SARS-CoV-2 main protease (M^{Pro}) bound to non-covalent inhibitors that have an associated experimental IC_{50}. The complex 7LTJ[2] is obtained as a result of non-covalent inhibition of MCULE-5948770040 compound with M^{Pro} (PDB-ID: 7JUN) discovered using our previous high throughput virtual screening as a part of the U.S. Department of Energy National Virtual Biotechnology Laboratory (NVBL) project[2]. When a protein–ligand complex had multiple measured IC_{50}, we used an average of the values as the label. The PDB-IDs of the other targets with the constituting protein and ligand are listed in Table 5.

Hyperparameter optimization. Hyperparameters such as network depth, layer dimension, and learning rate can have a large effect on model training and the weights in the final realized model. Therefore, we performed a number of trainings to examine combinations of learning rate, number of attention heads, and layer dimension. The hyperparameters that performed best were a learning rate of 0.0001, two attention heads, and a dimension of 70. These parameters resulted in an average test AUROC of 0.864. The combinations of these parameters are

Method	Test accuracy	Sensitivity	Specificity	References
GNN_F Overlap	**0.979**	0.840	**0.970**	Current work
GNN_P Overlap	0.958	**0.870**	0.910	Current work
GNN_P Distinct	0.855	0.590	0.910	Current work
GNN_F Distinct	0.951	0.690	**0.970**	Current work
GNN_F Base	0.934	0.660	**0.970**	Current work
GNN_P Base	0.845	0.580	0.900	Current work
Docking	0.591			Current work
GNN	0.968	0.830	**0.970**	Lim et al.[32]
CNN	0.904			Gonczarek et al.[46]
CNN	0.868			Ragoza et al.[26]
CNN	0.855			Wallach et al.[27]
Graph-CNN	0.886			Torng and Altman[28]

Table 6. Comparison of test dataset results for our GNN_F and GNN_P models on the various test sets described in the text, along with representative examples from the literature. The top scores are shown in bold.

summarized in Table SS1, and all successfully completed hyperparameter experiments can be found in Table SS2 along with further explanation of the trials.

Results and discussion

GNN classification models. Our primary goal in this work is to develop a model that has high accuracy and that can be generally applied for predicting PLI and activity using distinct atom- and bond-level features (domain awareness) for the protein and ligand. To this end, we would like to understand the effects of the number of protein targets and the number of protein–ligand complexes per target; therefore, we train the GNN_F and GNN_P models on a variety of datasets. The datasets consisted of either 17, 79, or 96 targets from the DUD-E dataset and all available data from the PDBbind dataset, which are summarized in Figure SS3 in the SI. In all cases, the accuracy greatly increases from 0.723 to 0.879 when the number of targets is increased from 17 to 79, and then decreases to 0.842 when the number of targets is further increased to 96. The accuracy of each experiment is shown in Table SS3. In each training set, we examined the effects of having 1000 complexes per target or 2000 complexes per target. When using 17 DUD-E targets, including more complexes per target did not improve the accuracy, while in both the 79 and 96 DUD-E target sets, the accuracy was improved to over 90% when the dataset consisted of 2000 complexes per target. Notably, the improvement was greater for the 96 DUD-E target set. It is important to note that all training sets had an equal distribution of positive and negative samples and complexes were randomly divided into training and test sets with an 80:20 split.

In addition to making sure the model can be generally applied, we are interested in developing a model that does not require the docked structure to be known before inferences can be made. We performed the same experiment using varying numbers of DUD-E targets and complexes on our GNN_P model, which does not require advance knowledge of the protein–ligand interaction to predict activity. The same overall trends were observed for GNN_P as for GNN_F but with reduced accuracy, as shown in Table SS3. The highest scoring GNN_P model was that trained on 79 DUD-E targets with 2000 complexes per target with an accuracy of 0.880%, which is only 3.2% lower than the GNN_F model trained on the same dataset. Though decreased accuracy was observed with GNN_P, its advantage relies on knowledge of only the separated protein and ligand structures, greatly reducing required preprocessing steps and increasing the throughput of the trained model.

To assess the ability of our GNN models to be generally applied, we produced three test sets of varying similarity to the training set (Table 6). Our base dataset consists of 79 targets from DUD-E and 991 targets from PDBbind randomly distributed into the training and test sets, with roughly 2000 samples per target and equal distribution of positive and negative samples. We also considered a dataset with overlapping targets but with novel complexes distributed into the training and test sets. Approximately 2000 additional samples that were withheld from the initial training were collected for each training target to show the effect of target overlap between training and test samples. Furthermore, we created a distinct dataset comprised of samples not used for training for the same target distribution in the base set. Table 6 shows the results of these tests in terms of test set accuracy, sensitivity, and specificity, along with some representative examples from the literature. The best performance is attained for the GNN_F model for the overlapping target dataset with an accuracy of 0.979, followed by the GNN_P model on the same dataset with an accuracy of 0.958. Additionally, the models show improved performance on the distinct sample set as compared with the base sample set. The accuracy increases from 0.934 to 0.951 for the GNN_F model and from 0.845 to 0.855 for the GNN_P model. The close similarity of the base, distinct, and overlap test set accuracies of the GNN_F model indicates that this model could be generally applied. GNN_P, however, showed reduced accuracy for the base and distinct test sets as compared with the overlap test set, indicating decreased capability to be generally applied. Overall, each implementation shows significantly improved performance in terms of prediction compared to docking. We can see that the overlap set produces a slight increase in AUROC, decrease in specificity, and significant increase in sensitivity.

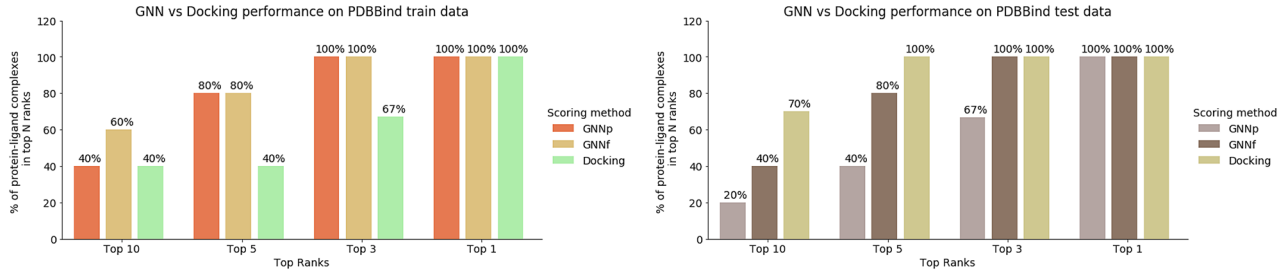

Figure 2. Comparison of the GNN_P and GNN_F models with docking. Each bar corresponds to the percentage of protein–ligand complexes identified in top-N ranks which have an RMSD less than 2 Å from the crystal structure.

Top-N ranks. We then assessed the model's ability to identify top-scoring 3D poses in each protein–ligand complex from our PDBbind repository described in the dataset preparation section. The 'best' docked pose is quantified as having an RMSD of less than 2 Å with respect to the crystal structure. In this analysis, we measure not only the ability of the model to identify an active protein–ligand pose but also its ability to identify the best pose among multiple docked poses. Figure 2 shows the percent of complexes found in the top-N ranks.

In each rank, both the GNN_F and GNN_P models outperformed or matched the performance of docking when using complexes from the training data. However, on the test data, docking showed better performance for certain ranks, while both docking and GNN_F identified 100% of the protein–ligand complexes. For the top rank, all three methods showed 100% identification. If we consider the percent targets with at least one pose in the top-N ranks, all three models show equivalent performance (see Fig. SS4).

In addition, we tested our models on datasets that include molecules from ChEMBL for DUD-E targets and SARS-CoV-2-target-specific data. Additional ChEMBL data were collected for a small subset of DUD-E targets extracted from the initial test set and implemented without docking for the GNN_P model. These molecules were prepared and matched with a pocket from the designated target. Roughly 2000 samples were prepared for two SARS-CoV-2 targets, M^{Pro} and NSP15, in a relatively balanced split of positive and negative samples. The GNN_F and GNN_P models trained on DUD-E and PDBbind then were used for inference on this set. Among all the experiments, GNN_P performed best with the ChEMBL data, showing an ROC of 0.596. The models performed relatively low on the SARS-CoV-2 targets, showing very low ROCs of 0.415 and 0.281 for the GNN_P and GNN_F models, respectively. This can be attributed to the fact that the SARS-CoV-2 target (M^{Pro}) used in this dataset has an average similarity of 40% with DUD-E training targets and 48% similarity to the PDBbind training data and therefore represents extrapolation rather than interpolation. Notably, graph neural networks are known to perform poorly on nonlinear extrapolation tasks far from the training data[47].

We next investigated the performance of the GNN_P model on IBS molecules with M^{Pro} and NSP15 protein targets. Their binding probability distributions are shown Fig. 3. The predicted activity distributions for NSP15 actives and IBS molecules are similar, suggesting that GNN_P can identify a majority of compounds from IBS with potential to bind to the NSP15 receptor. The predicted probability distribution of M^{Pro} target actives shows that GNN_P can identify a high percentage of molecules with potential binding affinity against the receptor, though it struggled to identify many of the IBS molecules. Theoretically, compounds with lower molecular weight (smaller size) possess a greater tendency to bind against a target. Most of the IBS compounds have a molecular weight between 250 and 700 Dalton (see Fig. SS6). For the M^{Pro} target, our results indicate that the binding probability is centered near zero and the majority binding probability is under 0.5. For the NSP15 target, the binding probability is centered closer to 0.1 with a small number of compounds having probabilities greater than 0.5. We observe that GNN_P consistently performed better in relation to the NSP15 target. Contributing factors to this increased performance include a larger number of available active compound samples for the NSP15 targets; that is, 2157 samples as opposed to 307 samples available for the M^{Pro} target. The NSP15 target has a disproportionately larger binding pocket than that of M^{Pro}, which also attributes to the improved performance.

GNN as a regression model. Predicting the binding affinity of a protein–ligand complex plays a critical role in identifying a lead molecule that binds with the protein target; however, the experimental measurement of protein–ligand binding affinity is laborious and time-consuming, which is one of the greatest bottlenecks in drug discovery. On the other hand, half maximal inhibitory concentration (IC_{50}) provides a quantitative measure of the potency of a candidate to inhibit a protein target and is typically estimated from experiments. If we can predict affinity and potency of a specific ligand to a target protein quickly and predict their interactions accurately, the efficiency of in silico drug discovery would be significantly improved. To this end, we modified both GNN models to perform regression in order to predict the biophysical and biochemical properties (affinity and potency) of protein–ligand complexes. Our regression models are composed of the same attention heads as in the classification models and, therefore, include the same domain-level information for the protein and ligand atoms. We evaluated these models on each of the datasets discussed in the Methods section.

Experimental binding affinity (EBA) regression experiments. To predict binding affinity and assess the performance of our model, we performed three experiments using three different datasets: (1) PDBbind2018-docked, (2) PDBbind2018 crystal, and (3) PDBbind2016 crystal structure dataset. Since most of the deep learning models that we compare in this section are trained on complexes from the PDBbind2016

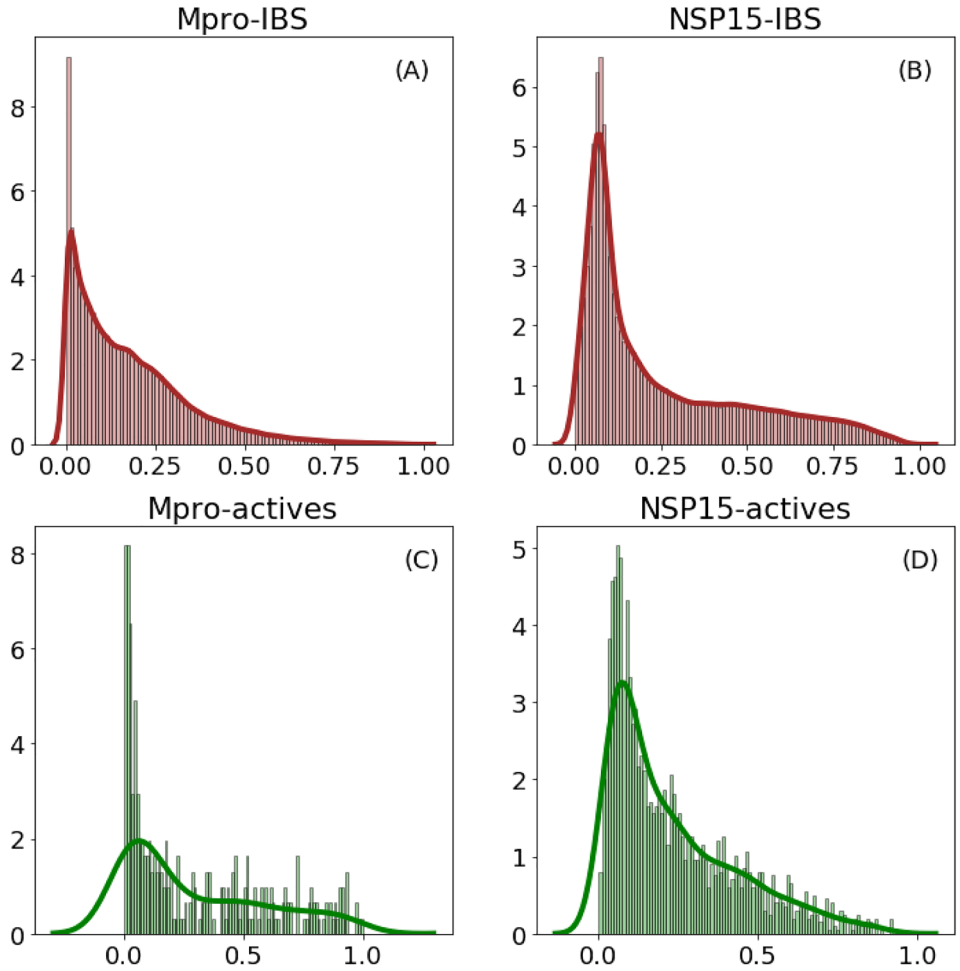

Figure 3. Binding probability distribution for IBS molecules with M^{Pro} and NSP15 as targets. (**A**,**B**) correspond to the predicted binding probability for NSP15 and M^{Pro} targets against IBS molecules. (**C**,**D**) correspond to the predicted binding probability on active molecules for M^{Pro} and NSP15 respectively (for each plot, the x-axis denotes the predicted probability and y-axis denotes the density of molecules).

database, training our models on the same dataset helps to obtain a comparison with the previous models[48–50]. We also tested our models on the PDBbind2016 core set, which is a refined subset filtered on the basis of protein-sequence similarity.

First, we assessed the performance of the GNN models trained using PDBbind2018 data docked and crystal-only data as shown in Table 7. Notably, the addition of more unique protein–ligand complexes improves the performance, while the addition of multiple docking poses for fewer complexes decreases performance, as the Pearson and Spearman correlations are low on the docked dataset as compared to the crystal-only dataset. We compared our methods with Pafnucy[49], which is a 3D CNN for protein–ligand affinity prediction that combines the 3D voxelization with atom-level features. Among all the results, the best performance is achieved by the GNN_P-EBA model, and overall, both GNN_P and GNN_F outperform Pafnucy on both the docking and crystal-only datasets.

In addition, for non-docking experiments, we used the PDBbind2016 general and refined datasets for training and the PDBbind2016 core set for testing. We assessed the performance of our methods for predicting the affinity through comparison with the K_{DEEP}[50] and FAST[48] models. Our analysis on the PDBbind2016 core set shows that the GNN_P-EBA model performs similarly to the FAST model[48], while K_{DEEP} shows the highest performance (see Table SS4 for detailed comparison). We also report the performance of our model on the dataset used for training and evaluating the FAST model (see Table SS6 for details) using the same train-validation-test splits as provided by the authors over the PDBbind2016 general (G) and refined (R) sets[48]. Out of all the combinations of the general and refined sets, we achieved the best performance with the GNN_P model trained on the general set.

Our results on the PDBbind2018 and PDBbind2016 datasets suggests that with our proposed GNN frameworks, we achieve top performance compared with prior deep learning methods for binding affinity prediction while preserving the spatial orientation of the protein and ligand. While K_{DEEP} is a purely 3D CNN-based network for protein–ligand affinity predictions, the FAST model uses a graph-based network, but utilizes the 2D graph representation which does not include the non-covalent interactions. This structural information is critical for understanding PLI and its impact on estimating affinity. To tackle these issues, we devised our

MODEL	RMSE	MAE	Pearson r	Spearman r	r^2
GNN_F-EBA-docking	1.74	1.41	0.40	0.39	0.16
GNN_P-EBA-docking	1.61	1.32	0.49	0.49	0.24
GNN_F-EBA-crystal-only	1.68	1.32	0.45	0.46	0.21
GNN_P-EBA-crystal-only	**1.48**	**1.16**	**0.59**	**0.59**	**0.35**
Pafnucy[49] (docking)[a]	1.60	1.34	0.52	0.50	0.27
Pafnucy[49] (crystal-only)[b]	1.86	1.50	0.38	0.37	0.14

Table 7. Performance comparison of our GNN models in predicting experimental affinity on the PDBbind dataset. The top scores are shown in bold. [a]255 test targets from our PDBbind2018-EBA-docking dataset using all docked poses for evaluation; [b]PDBBind2018-EBA-crystal-only test dataset.

MODEL	RMSE	MAE	Pearson r	Spearman r
GNN_F-EBA	1.39	**1.10**	0.49	0.50
GNN_P-EBA	1.52	1.22	0.42	0.46
Pafnucy[49]	**1.38**	1.11	**0.52**	**0.52**
FAST[48]	1.48	1.21	0.42	0.40
K_{DEEP}[50]	1.42	1.13	0.48	0.47

Table 8. Performance comparison of our GNN models in predicting experimental affinity on the PDBbind2019 structure-based evaluation dataset. The results for the FAST method are reported for its 3D CNN model. The top scores are shown in bold.

graph-based models to specifically include all the atom- and bond-level information while using the 3D structures of the protein and ligand. The graph representation not only encloses atom- and bond-level information, but also retains the spatial information associated with the protein–ligand complexes, thus enabling us to include all necessary information associated with the natural binding state of the protein and ligand. Our GNN_F model accounts for intermolecular interactions between the protein and ligand, which are not captured in the FAST model.

To investigate the generalizability of our GNN models in predicting the binding affinity of unseen and novel targets, we compare the performance of our GNN and various models on the PDBbind2019 structure-based evaluation dataset (Table 8). The PDBbind2019 structure-based evaluation dataset is composed of targets that are novel from the PDBbind2016 in terms of their addition to the database as well as sequence similarity. The prediction results of the GNN models are better than those of previous models, as shown in Table 8, and most importantly, our GNN_F-EBA model performs as accurate as Pafnucy. Our results demonstrate that even with limited 3D structural data in terms of the size typically needed to train deep learning models, we achieved relatively more accurate generalization with the GNN model. In addition, our GNN outperforms FAST and K_{DEEP} in terms of generalizability given the scarcity of available structural data.

Finally, to expand the scope of our experiment, we also trained our model on a physics-based docking affinity score. We refer to these models as Docking Binding Affinity (DBA) models. We compare our model's performance against the physics-based docking on the two DBA datasets. Our results suggest that, with a correlation score of 0.79, the GNN_F model was able to reproduce some correlation between the actual and predicted affinity to an extent (see Tables SS7 and SS8 for detailed dataset description and results). This indicates that, with our GNN framework, we are not only capturing the details needed for predicting the binding affinity but also achieving the capability to differentiate between distinct docked poses of a protein–ligand complex and associate it with its docking score.

pIC_{50} *regression experiments.* As a next step, we tailored our GNN models to predicting pIC_{50}. pIC_{50} provides a quantitative measure of the potency of a candidate to inhibit a protein target, which is typically estimated from experiments. A number of methods have been developed to approximate pIC_{50}, so we compared our pIC_{50} prediction with existing deep-learning methods, such as DeepAffinity[21], DeepDTA[22] and MONN[33]. While these methods were trained purely on the protein sequence and 2D SMILES representation of the ligands, our model is novel in that it considers the 3D structures of the protein and ligand to predict pIC_{50}, which is key to defining the inhibition rate while identifying and optimizing hits in early-stage drug discovery. To the best of our knowledge, 3D protein–ligand complexes have not been used to predict pIC_{50} before.

The overall performance of the GNN model is relatively low compared to the existing methods listed in the Table 9. This could be attributed to the smaller size of the dataset containing both pIC_{50} and the corresponding crystal structures. DeepAffinity[21], DeepDTA[22], and MONN[33] were trained on BindingDB data, which has nearly 10 times the amount of data than that available for our dataset. In addition, we observed improvement in the Pearson correlation coefficients in predicting pIC_{50} from 0.45 to 0.51 when using weights from the GNN-EBA model. The improvement from the baseline to the transfer learning model suggests that our model can achieve

Model	RMSE	Pearson r
GNN_P-pIC_{50}	1.24	0.45
GNN_F-pIC_{50}	1.26	0.44
GNN_P-pIC_{50} best GNN_F-EBA weights	1.21	0.51
GNN_F-pIC_{50} best GNN_P-EBA weights	1.24	0.51
DeepAffinity[21]	**0.74**	0.84
DeepDTA[22]	0.78	0.85
MONN[33]	0.76	**0.86**

Table 9. Performance comparison of deep learning models in predicting pIC_{50}. Our models were trained and tested on PDBbind2016 + DUD-E targets, whose IC_{50} was curated from the PDBbind and ChEMBL repositories, respectively. DeepAffinity, DeepDTA, and MONN were trained and tested on BindingDB data. The top scores are given in bold.

PDB-ID	GNN_P-pIC_{50}	GNN_P-pIC_{50} best EBA	GNN_F-pIC_{50}	GNN_F-pIC_{50} best EBA	Experimental pIC_{50}
7TLJ	6.20	7.12	6.51	6.90	5.37
7L0D	7.56	7.08	7.33	7.43	5.6
7LME	7.61	7.07	7.46	7.59	5.3
7L11	7.90	7.13	7.74	7.59	6.8
7L12	7.96	7.36	8.26	7.61	7.74
7L13	8.11	7.42	8.42	7.56	6.89
7L14	8.06	7.16	7.87	7.67	6.76

Table 10. Performance of GNN on SARS-CoV-2 M^{Pro} targets and some of the potential inhibitors whose IC_{50} has been experimentally measured.

better performance if a larger dataset is used, as the GNN-EBA models are trained on comparatively larger datasets.

To assess how critical the learned protein and ligand representations are for predicting pIC_{50}, we predicted pIC_{50} of a few inhibitors that were recently designed for SARS-CoV-2 Mpro, where the co-crystal structures have been solved and their IC_{50} have been experimentally measured. From the perspective of deep learning and 3D protein–ligand complex representation, this is a much more difficult regression problem compared to the classification problem above. Our ultimate goal was to quantify the error in IC_{50} prediction relative to the experimental measurement that could used for iterative design of potential inhibitors or lead optimization.

Our extensive analysis on the SARS-CoV-2 M^{Pro} data demonstrates that both GNN models overestimate IC_{50} by 0.92% as compared to experimental values, as shown in Table 10. The pIC_{50} of our recently designed MCULE-5948770040 compound with M^{Pro} 7TLJ complex with GNN_P model predicted to be 6.20, which is comparable to the measured experimental value of 5.37[2]. Interestingly, GNN_P proved to be the best model with an average error of 0.42 Molar. It is important to highlight that GNN_P gives an advantage in predicting pIC_{50} even when the experimentally bound structure is not known. This can help rank potent candidates while screening potential libraries against a protein target or possible protein targets of a given disease, which can then be utilized for experimental testing as summarized in Fig. 4.

Conclusion

In this work, we devised graph-based deep learning models, GNN_P and GNN_F, by integrating knowledge representation, 3D structural information and reasoning for PLI prediction through classification and regression properties of protein–ligand complexes. The parallelization of the GNN_P model provides a basis for novel implementation of structural analysis that requires no docking input but instead separate protein and ligand 3D structures. The basic strategy of GNN_P is to learn embedding vectors of the ligand graph and protein graph separately and combine the two embedding vectors for prediction. The featurization of GNN_F provided a baseline for our implementation of domain-aware capabilities enhanced through feature engineering to identify significant nodes and differentiate the contribution of each interaction to the affinity. In GNN_F, the embedding vectors are learned simultaneously for the protein and ligand complex as an early embedding strategy.

These implementations enable us to leverage the vast amount of 3D structural data of both the target protein and ligands, and interactions between them which is crucial for activity, potency, and affinity prediction to accelerate in silico hit identification during early stages of drug design. The goal of our extensive study is to generalize the graph-based models by incorporating domain-aware information, features, and biophysical properties and by utilizing a large amount of data including a variety of targets. The test accuracy for GNN_F reached 0.951 on a distinct sample set. We achieved top GNN_F performance with the target overlap sample set, resulting in a test accuracy of 0.979 (0.958 for GNN_P), providing a basis that further generalizing our model can produce top classification performance. In addition, we used the GNN_P model to evaluate the performance on SARS-CoV-2

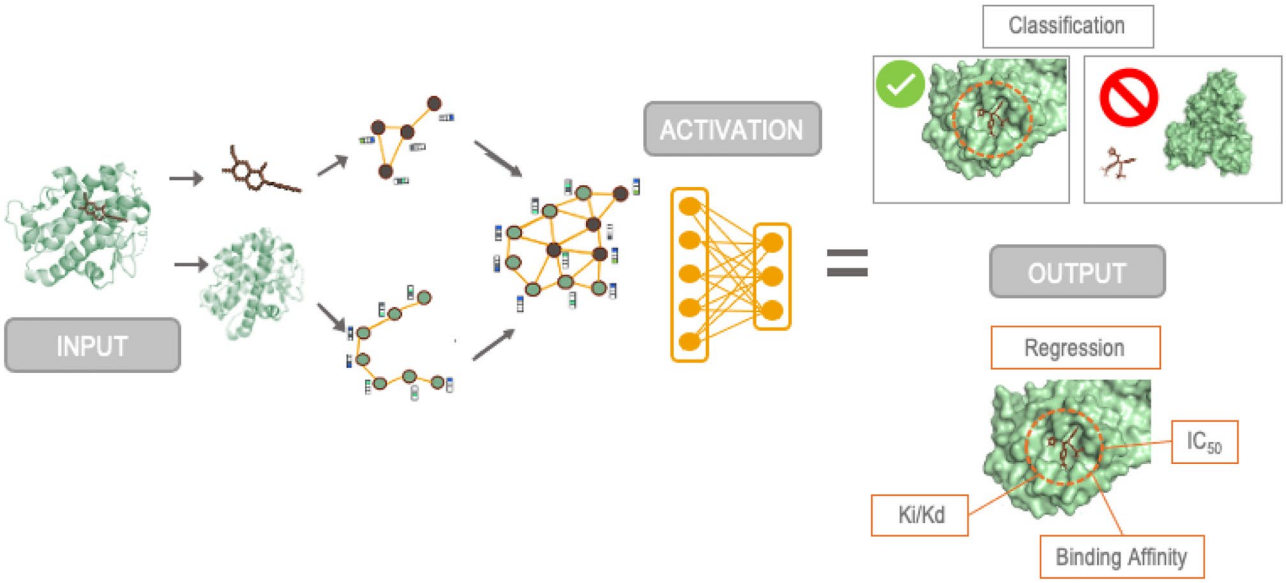

Figure 4. Schematic showing the results produced by each method in the both regression and classification model. The inclusion of 3D structural data provides numerous advantages and ability to produce such a range of prediction both activity and biophysical properties, their relationship with protein–ligand interactions.

(M^{Pro}) and NSP15 as target proteins. The predicted probability distribution of target actives shows that GNN_P can identify a high percentage of molecules with potential binding affinity against the receptor.

Our GNN models were further modified for regression tasks to predict binding affinities and pIC_{50} in comparison with experimentally measured values. Experimentation with regression problems such as pIC_{50}, experimental affinity and docking binding affinity shows that our graph-based featurization of protein and ligands not only captures the binding probability but is efficient enough to learn other important factors associated with PLI. In terms of prediction, our GNN model outperform existing models[48–50] with highest prediction correlation coefficients. Using PDBbind2016 data, we achieved Pearson correlation coefficients of 0.66 and 0.65 on experimental affinity prediction and 0.50 and 0.51 on pIC_{50} prediction using GNN_F and GNN_P, respectively. Even with limited 3D structural data for the pIC_{50} dataset, we achieved comparable performance to existing methods that were trained on relatively larger 2D sequence datasets. With the availability of pIC_{50} data and corresponding protein structure predicted from Alphafold[51], the GNN model performance can be further improved.

Our model is unique and novel in that it considers the 3D structures of the protein and ligand to predict affinity and pIC_{50}, which is key to provide a quantitative measure of the potency and selectivity of a candidate to inhibit a specific protein target. To accelerate in silico hit identification and lead optimization in the early stage of drug design, our GNN_P model can be used to screen a large ligand library to predict either biophysical properties or activities against a given protein target or set of targets for specific disease.

Supporting Information

The Supporting information is available with Table (S1–S8) and Figures (S1–S7)) detailing data, models, hyper-parameter optimization, Performance Metrics for Classification and Regression GNN Models, IBS molecule properties, and Docking Binding affinity Data and Results used in this study. Description about hyper-parameters and the trained GNN_P and GNN_f models and the code to reproduce this study is available at https://github.com/PNNL-CompBio/pf-gnn_pli.

Data and code availability

We collected and curated protein–ligand complexes from two public datasets, DUD-E[34] and PDBbind[35,36] which are described in further detail in Dataset Preparation subsection. SARS-CoV-2 and other dataset is generated in using high throughput docking simulations and the IBS dataset is created using known synthetic compounds from the IB screening database (www.ibscreen.com). The trained GNN_P and GNN_f models and the code to reproduce this study is available at https://github.com/PNNL-CompBio/pf-gnn_pli.

Received: 9 December 2021; Accepted: 6 April 2022
Published online: 10 May 2022

References

1. Chen, W., Chen, G., Zhao, L. & Chen, C.Y.-C. Predicting drug-target interactions with deep-embedding learning of graphs and sequences. *J. Phys. Chem. A* **125**, 5633–5642 (2021).
2. Others, et al. High throughput virtual screening and validation of a SARS-CoV-2 main protease non-covalent inhibitor. *bioRxiv* (2021).

3. Wang, L. *et al.* Accurate modeling of scaffold hopping transformations in drug discovery. *J. Chem. Theory Comput.* **13**, 42–54 (2017).
4. Beierlein, F. R., Michel, J. & Essex, J. W. A simple QM/MM approach for capturing polarization effects in protein–ligand binding free energy calculations. *J. Phys. Chem. B* **115**, 4911–4926 (2011).
5. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395 (2014).
6. Boniolo, F. *et al.* Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opin. Drug Discov.* **16**, 991–1007 (2021).
7. Others, et al. Evolution of sequence-based bioinformatics tools for protein–protein interaction prediction. *Curr. Genom.* **21**, 454–463 (2020).
8. Venkatachalam, C. M., Jiang, X., Oldfield, T. & Waldman, M. LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.* **21**, 289–307 (2003).
9. Allen, W. J. *et al.* DOCK 6: Impact of new features and current docking performance. *J. Comput. Chem.* **36**, 1132–1156 (2015).
10. Ruiz-Carmona, S. *et al.* rDock: A fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.* **10**, e1003571 (2014).
11. Zhao, H. & Caflisch, A. Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. *Bioorg. Med. Chem. Lett.* **23**, 5721–5726 (2013).
12. Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **46**, 499–511 (2003).
13. Jones, G., Willett, P., Glen, R., Leach, A. & Taylor, R. *Development and Validation of a Genetic Algorithm for Flexible Ligand Docking* 154-COMP. Abstracts of Papers of the American Chemical Society (1997).
14. Friesner, R. A. *et al.* Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
15. Vina, A. Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading Trott, Oleg; Olson, Arthur J. *J. Comput. Chem.* **31**, 455–461 (2010).
16. Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T. & Alinejad-Rokny, H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *J. Biomed. Inform.* **113**, 103627 (2021).
17. Others, et al. Artificial intelligence in drug design. *Sci. China Life Sci.* **61**, 1191–1204 (2018).
18. Wen, M. *et al.* Deep-learning-based drug-target interaction prediction. *J. Proteome Res.* **16**, 1401–1409 (2017).
19. Tsukiyama, S., Hasan, M. M., Fujii, S. & Kurata, H. LSTM-PHV: Prediction of human-virus protein–protein interactions by LSTM with word2vec. *Briefings Bioinform.* **22**, bbab228 (2021).
20. Khatun, S., Alam, A., Shoombuatong, W., Mollah, M. N. H., Kurata, H. & Hasan, M. M. Recent development of bioinformatics tools for microRNA target prediction. *Curr. Med. Chem.* **2021**.
21. Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**, 3329–3338 (2019).
22. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: Deep drug-target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).
23. Glen, R. C. *et al.* Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **9**, 199 (2006).
24. Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **28**, 2224–2232 (2015).
25. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
26. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
27. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855 (2015).
28. Torng, W. & Altman, R. B. Graph convolutional neural networks for predicting drug-target interactions. *J. Chem. Inf. Model.* **59**, 4131–4149 (2019).
29. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J. & Bronstein, M. M. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 5115–5124 (2017).
30. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. & Bengio, Y. Graph attention networks. arXiv preprint arXiv:1710.10903 (2017).
31. Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493 (2015).
32. Lim, J. *et al.* Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. *J. Chem. Inf. Model.* **59**, 3981–3988 (2019).
33. Li, S. *et al.* MONN: A multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Syst.* **10**, 308–322 (2020).
34. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
35. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* **47**, 2977–2980 (2004).
36. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).
37. Smusz, S., Kurczab, R. & Bojarski, A. J. The influence of the inactives subset generation on the performance of machine learning methods. *J. Cheminform.* **5**, 17 (2013).
38. Chen, L. *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE* **14**, e0220113 (2019).
39. Landrum, G. *RDKit: Open-Source Cheminformatics*. https://www.rdkit.org/ Q3 (2016).
40. Meli, R. *RMeli/PDBbind-Docking: PDBbind19 Refined Docking (Version 0.1.0)*. https://github.com/RMeli/PDBbind-docking (2020).
41. Zinc database. http://zinc.docking.org/substances/subsets/fda/?page=1 (accessed 30 August 2020).
42. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2009).
43. Lockbaum, G. J. *et al.* Crystal structure of SARS-CoV-2 main protease in complex with the non-covalent inhibitor ML188. *Viruses* **13**, 174 (2021).
44. Others, et al. Structure-based optimization of ML300-derived, noncovalent inhibitors targeting the Severe Acute Respiratory Syndrome Coronavirus 3CL Protease (SARS-CoV-2 3CLpro). *J. Med. Chem.* (2021).
45. Others, et al. Potent noncovalent inhibitors of the main protease of SARS-CoV-2 from molecular sculpting of the drug perampanel guided by free energy perturbation calculations. ACS Cent. Sci. **7**, 467–475 (2021)
46. Gonczarek, A., Tomczak, J. M., Zaręba, S., Kaczmar, J., Dąbrowski, P. & Walczak, M. J. Learning deep architectures for interaction prediction in structure-based virtual screening. arXiv preprint arXiv:1610.07187 (2016).
47. Xu, K., Zhang, M., Li, J., Du, S. S., Kawarabayashi, K, Jegelka, S. How neural networks extrapolate: From feedforward to graph neural networks. arXiv preprint arXiv:2009.11848 (2020).

48. Jones, D. *et al.* Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* **61**, 1583–1592 (2021).
49. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).
50. Jiménez, J., Skalic, M., Martinez-Rosell, G. & De Fabritiis, G. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **58**, 287–296 (2018).
51. Others, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 1–11 (2021).

Acknowledgements
This research was supported by Laboratory Directed Research and Development Program and Mathematics for Artificial Reasoning for Scientific Discovery investment at the Pacific Northwest National Laboratory, which is operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RLO. We thank Rajendra Joshi, Darin Hauner, and Andrew McNaughton at PNNL for discussion on the protein–ligand interactions and docking simulations. We extend our thank to Sutanay Choudhury, Khushbu Agarwal at PNNL and Aman Ahuja at Virginia Tech for extensive discussion on graph based models for accurate representation of 3D structural coordinates of protein–ligand complexes.

Author contributions
N.K. and J.A.B. contributed concept and implementation. C.K. M.B, J.A.B and N.K. co-designed experiments and were responsible for programming. All authors contributed to the interpretation of results and wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10418-2.

Correspondence and requests for materials should be addressed to N.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.