bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

GraphDTA: Predicting drug–target binding affinity with graph neural networks

Thin Nguyen[1], Hang Le[2], Thomas P. Quinn[1], Tri Nguyen[1], Thuc Duy Le[3], Svetha Venkatesh[1]
[1]Applied Artificial Intelligence Institute, Deakin University, Australia
[2]Faculty of Information Technology, Nha Trang University, Viet Nam
[3]School of Information Technology and Mathematical Sciences, University of South Australia, Australia

Abstract

The development of new drugs is costly, time consuming, and often accompanied with safety issues. Drug repurposing can avoid the expensive and lengthy process of drug development by finding new uses for already approved drugs. In order to repurpose drugs effectively, it is useful to know which proteins are targeted by which drugs. Computational models that estimate the interaction strength of new drug--target pairs have the potential to expedite drug repurposing. Several models have been proposed for this task. However, these models represent the drugs as strings, which is not a natural way to represent molecules. We propose a new model called **GraphDTA** that represents drugs as graphs and uses graph neural networks to predict drug--target affinity. We show that graph neural networks not only predict drug--target affinity better than non-deep learning models, but also outperform competing deep learning methods. Our results confirm that deep learning models are appropriate for drug--target binding affinity prediction, and that representing drugs as graphs can lead to further improvements.

Availability of data and materials: The proposed models are implemented in Python. Related data, pre-trained models, and source code are publicly available at https://github.com/thinng/GraphDTA. All scripts and data needed to reproduce the post-hoc statistical analysis are available from https://doi.org/10.5281/zenodo.3603523.

Contact: Thin.Nguyen@deakin.edu.au

Keyword: drug-target, binding affinity, graph neural networks

1 Background

It costs about 2.6 billion US dollars to develop a new drug [1], and can take up to 17 years for FDA approval [2, 3]. Finding new uses for already approved drugs avoids the expensive and lengthy process of drug development [2, 4]. For example, nearly 70 existing FDA-approved drugs are currently being investigated to see if they can be repurposed to treat COVID-19 [5]. In order to

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

repurpose drugs effectively, it is useful to know which proteins are targeted by which drugs. High-throughput screening experiments are used to examine the affinity of a drug toward its targets; however, these experiments are costly and time-consuming [6, 7], and an exhaustive search is infeasible because there are millions of drug-like compounds [8] and hundreds of potential targets [9, 10]. As such, there is a strong motivation to build computational models that can estimate the interaction strength of new drug–target pairs based on previous drug–target experiments.

Several computational approaches have been proposed for drug–target affinity (**DTA**) prediction [11, 12, 13]. One approach is molecular docking, which predicts the stable 3D structure of a drug-target complex via a scoring function [14]. Even though the molecular docking approach is potentially more informative, it require knowledge about the crystallized structure of proteins which may not be available. Another approach uses collaborative filtering. For example, the **SimBoost** model uses the affinity similarities among drugs and among targets to build new features. These features are then used as input in a gradient boosting machine to predict the binding affinity for unknown drug–target pairs [15]. Alternatively, the similarities could come from others sources (rather than the training data affinities). For example, kernel-based methods use kernels built from molecular descriptors of the drugs and targets within a regularized least squares regression (RLS) framework [16, 17]. To speed up model training, the **KronRLS** model computes a pairwise kernel K from the Kronecker product of the drug-by-drug and protein-by-protein kernels [16, 17] (for which any similarity measure can be used). **DTA** prediction may also benefit from adopting methods for predicting drug–target interactions (**DTI**). Approaches in this line of work include DTI-CDF [18], a cascade deep forest model, or DTI-MLCD [19], a multi-label learning supported with community detection.

Another approach uses neural networks trained on 1D representations of the drug and protein sequences. For example, the **DeepDTA** model uses 1D representations and layers of 1D convolutions (with pooling) to capture predictive patterns within the data [20]. The final convolution layers are then concatenated, passed through a number of hidden layers, and regressed with the drug–target affinity scores. The **WideDTA** model is an extension of **DeepDTA** in which the sequences of the drugs and proteins are first summarized as higher-order features [21]. For example, the drugs are represented by the most common sub-structures (the Ligand Maximum Common Substructures (LMCS) [22]), while the proteins are represented by the most conserved sub-sequences (the Protein Domain profiles or Motifs (PDM) from PROSITE [23]). While **WideDTA** [21] and **DeepDTA** [20] learn a latent feature vector for each protein, the **PADME** model [24] uses fixed-rule descriptors to represent proteins, and performs similarly to **DeepDTA** [20].

The deep learning models are among the best performers in DTA prediction [25]. However, these models represent the drugs as strings, which are not a natural way to represent molecules. When using strings, the structural information of the molecule is lost, which could impair the predictive power of a model as well as the functional relevance of the learned latent space. Already,

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

graph convolutional networks have been used in computational drug discovery, including interaction prediction, synthesis prediction, *de novo* molecular design, and quantitative structure prediction [26, 27, 28, 29, 30, 31, 25]. However, graph neural networks have not been used for DTA prediction. Of these, [26, 25, 32] are closest to our work, but look at binary prediction, while our model looks to predict a continuous value of binding affinity. Also, in [25], the input is a drug descriptor (single input), while our model takes as input both a drug descriptor and a sequence (dual input).

In this article, we propose **GraphDTA**, a new neural network architecture capable of directly modelling drugs as molecular graphs, and show that this approach outperforms state-of-the-art deep learning models on two drug–target affinity prediction benchmarks. The approach is based on the solution we submitted to the IDG-DREAM Drug-Kinase Binding Prediction Challenge[1], where we were among the Top Ten Performers from 530 registered participants[2]. In order to better understand how our graph-based model works, we performed a multivariable statistical analysis of the model's latent space. We identified correlations between hidden node activations and domain-specific drug annotations, such as the number of aliphatic OH groups, which suggests that our graph neural network can automatically assign importance to well-defined chemical features without any prior knowledge. We also examine the model's performance and find that a handful of drugs contribute disproportionately to the total prediction error, and that these drugs are inliers (i.e., not outliers) in an ordination of the model's latent space. Taken together, our results suggest that graph neural networks are highly accurate, abstract meaningful concepts, and yet fail in predictable ways. We conclude with a discussion about how these insights can feedback into the research cycle.

2 Methods

2.1 Overview of GraphDTA

We propose a novel deep learning model called **GraphDTA** for drug–target affinity (**DTA**) prediction. We frame the DTA prediction problem as a regression task where the input is a drug–target pair and the output is a continuous measurement of binding affinity for that pair. Existing methods represent the input drugs and proteins as 1D sequences. Our approach is different; we represent the drugs as molecular graphs so that the model can directly capture the bonds among atoms.

2.2 Drug representation

SMILES (Simplified Molecular Input Line Entry System) was invented to represent molecules to be readable by computers [33], enabling several efficient

[1] https://www.synapse.org/#!Synapse:syn15667962/wiki/583305
[2] https://www.synapse.org/#!Synapse:syn15667962/wiki/592145

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

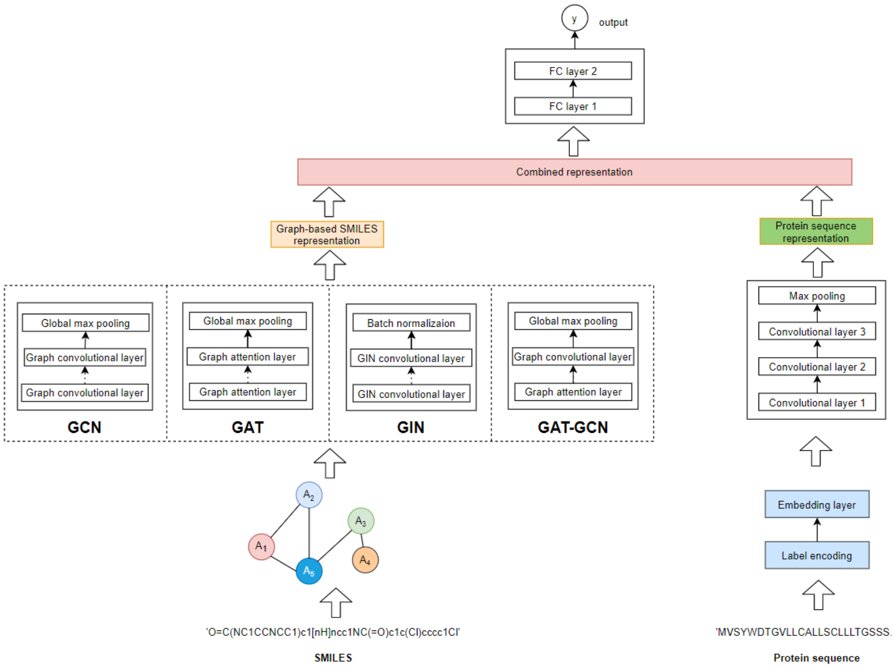

Figure 1: This figure shows the **GraphDTA** architecture. It takes a drug–target pair as the input data, and the pair's affinity as the output data. It works in 3 stages. First, the SMILES code of a drug is converted into a molecular graph, and a deep learning algorithm learns a graph representation. Meanwhile, the protein sequence is encoded and embedded, and several 1D convolutional layers learn a sequence representation. Finally, the two representation vectors are concatenated and passed through several fully connected layers to estimate the output drug–target affinity value.

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

applications, including fast retrieval and substructure searching. From the SMILES code, drug descriptors like the number of heavy atoms or valence electrons can be inferred and readily used as features for affinity prediction. One could also view the SMILES code as a string. Then, one could featurize the strings with natural language processing (NLP) techniques, or use them directly in a convolutional neural network (CNN).

Instead, we view drug compounds as a graph of the interactions between atoms, and build our model around this conceptualization. To describe a node in the graph, we use a set of atomic features adapted from **DeepChem** [34]. Here, each node is a multi-dimensional binary feature vector expressing five pieces of information: the atom symbol, the number of adjacent atoms, the number of adjacent hydrogens, the implicit value of the atom, and whether the atom is in an aromatic structure [34]. We convert the SMILES code to its corresponding molecular graph and extract atomic features using the open-source chemical informatics software **RDKit** [35].

2.3 Protein representation

One-hot encoding has been used in previous works to represent both drugs and proteins, as well as other biological sequences like DNA and RNA. This paper tests the hypothesis that a graph structure could yield a better representation for drugs, and so only drugs were represented as a graph. Although one could also represent proteins as graphs, doing so is more difficult because the tertiary structure is not always available in a reliable form. As such, we elected to use the popular one-hot encoding representation of proteins instead.

For each target in the experimented datasets, a protein sequence is obtained from the UniProt database using the target's gene name. The sequence is a string of ASCII characters which represent amino acids. Each amino acid type is encoded with an integer based on its associated alphabetical symbol (e.g., Alanine (A) is 1, Cystine (C) is 3, Aspartic Acid (D) is 4, and so on), allowing the protein to be represented as an integer sequence. To make it convenient for training, the sequence is cut or padded to a fixed length sequence of 1000 residues. In case a sequence is shorter, it is padded with zero values.

These integer sequences are used as input to the embedding layers which return a 128-dimensional vector representation. Next, three 1D convolutional layers are used to learn different levels of abstract features from the input. Finally, a max pooling layer is applied to get a representation vector of the input protein sequence.

2.4 Deep learning on molecular graphs

Having the drug compounds represented as graphs, the task now is to design an algorithm that learns effectively from graphical data. The recent success of CNN in computer vision, speech recognition, and natural language processing has encouraged research into graph convolution. A number of works have been proposed to handle two main challenges in generalizing CNN to graphs: (1) the

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

formation of receptive fields in graphs whose data points are not arranged as Euclidean grids, and (2) the pooling operation to down-sample a graph. These new models are called graph neural networks.

In this work, we propose a new DTA prediction model based on a combination of graph neural networks and conventional CNN. Figure 1 shows a schematic of the model. For the proteins, we use a string of ASCII characters and apply several 1D CNN layers over the text to learn a sequence representation vector. Specifically, the protein sequence is first categorically encoded, then an embedding layer is added to the sequence where each (encoded) character is represented by a 128-dimensional vector. Next, three 1D convolutional layers are used to learn different levels of abstract features from the input. Finally, a max pooling layer is applied to get a representation vector of the input protein sequence. This approach is similar to the existing baseline models. For the drugs, we use the molecular graphs and trial 4 graph neural network variants, including GCN [36], GAT [37], GIN [38], and a combined GAT-GCN architecture, all of which we describe below.

2.4.1 Variant 1: GCN-based graph representation learning

In this work, we focus on **predicting a continuous value** indicating the level of interaction of a drug and a protein sequence. Each drug is encoded as a graph and each protein is represented as a string of characters. To this aim, we make use of GCN model [36] for learning on graph representation of drugs. Note that, however, the original GCN is designed for semi-supervised node classification problem, i.e., the model learns the **node-level feature vectors**. For our goal, to estimate the drug-protein interaction, a **graph-level representation** of each drug is required. Common techniques to aggregate the whole graph feature from learned node features include Sum, Average, and Max Pooling. In our experiments, the use of Max Pooling layer in GCN-based GraphDTA usually results in better performance compared to that of the remaining.

Formally, denote a graph for a given drug as $G = (V, E)$, where V is the set of N nodes each is represented by a C-dimensional vector and E is the set of edges represented as an adjacency matrix A. A multi-layer graph convolutional network (GCN) takes as input a node feature matrix $X \in \mathbb{R}^{N\times C}(N = |V|$, C: the number of features per node) and an adjacency matrix $A \in \mathbb{R}^{N\times N}$; then produces a node-level output $Z \in \mathbb{R}^{N\times F}(F$: the number of output features per node). A propagation rule can be written in the normalized form for stability, as in [36]:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{1}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix of the undirected graph with added self-connections, $\tilde{D}_{ii} = \sum_i \tilde{A}_{ii}$; $H^{(l)} \in \mathbb{R}^{N\times C}$ is the matrix of activation in the l^{th} layer, $H^{(0)} = X$, σ is an activation function, and W is learnable parameters.

A layer-wise convolution operation can be approximated, as in [36]:

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$Z = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta \tag{2}$$

where $\Theta \in \mathbb{R}^{C\times F}$ (F: the number of filters or feature maps) is the matrix of filter parameters.

Note that, however, the GCN model learns node-level outputs $Z \in \mathbb{R}^{N\times F}$. To make the GCN applicable to the task of learning a representation vector of the whole graph, we add a global max pooling layer right after the last GCN layer. In our GCN-based model, we use three consecutive GCN layers, each activated by a ReLU function. Then a global max pooling layer is added to obtain the graph representation vector.

2.4.2 Variant 2: GAT-based graph representation learning

Unlike graph convolution, the graph attention network (**GAT**) [37] proposes an attention-based architecture to learn hidden representations of nodes in a graph by applying a self-attention mechanism. The building block of a GAT architecture is a *graph attention layer*. The GAT layer takes the set of nodes of a graph as input, and applies a linear transformation to every node by a weigh matrix $\mathbf{W}$. For each input node i in the graph, the *attention coefficients* between i and its first-order neighbors are computed as

$$a(\mathbf{W}\mathbf{x}_i, \mathbf{W}\mathbf{x}_j) \tag{3}$$

This value indicates the importance of node j to node i. These *attention coefficients* are then normalized by applying a soft-max function, then used to compute the output features for nodes as

$$\sigma(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}\mathbf{x}_j) \tag{4}$$

where $\sigma(.)$ is a non-linear activation function and α_{ij} are the normalized *attention coefficients*.

In our model, the GAT-based graph learning architecture includes two GAT layers, activated by a ReLU function, then followed a global max pooling layer to obtain the graph representation vector. For the first GAT layer, *multi-head-attentions* are applied with the number of heads set to 10, and the number of output features set to the number of input features. The number of output features of the second GAT is set to 128.

2.4.3 Variant 3: Graph Isomorphism Network (GIN)

The graph isomorphism network (**GIN**) [38] is newer method that supposedly achieves maximum discriminative power among graph neural networks. Specifically, GIN uses a multi-layer perceptron (MLP) model to update the node features as

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Method	Protein rep.	Compound rep.	CI	MSE
Baseline models				
DeepDTA [20]	Smith-Waterman	Pubchem-Sim	0.790	0.608
DeepDTA [20]	Smith-Waterman	1D	*0.886*	0.420
DeepDTA [20]	1D	Pubchem-Sim	0.835	0.419
KronRLS [16, 17]	Smith-Waterman	Pubchem-Sim	0.871	0.379
SimBoost [15]	Smith-Waterman	Pubchem-Sim	0.872	0.282
DeepDTA [20]	1D	1D	0.878	*0.261*
WideDTA [21]	1D + PDM	1D + LMCS	*0.886*	0.262
Proposed method - **GraphDTA**				
GCN	1D	Graph	0.880	**0.254**
GAT_GCN	1D	Graph	0.881	**0.245**
GAT	1D	Graph	**0.892**	**0.232**
GIN	1D	Graph	**0.893**	**0.229**

Table 1: Prediction performance on the Davis dataset, sorted by MSE. Baseline results are from [20, 21]. We compare 4 graph neural network variants: GIN [38], GAT [37], GCN [36], and combined GAT-GCN [37, 36]. Italics: best for baseline models, bold: better than baselines.

$$MLP((1+\epsilon)\mathbf{x}_i + \sum_{j \in B(i)} \mathbf{x}_i) \tag{5}$$

where ϵ is either a learnable parameter or a fixed scalar, $\mathbf{x}$ is the node feature vector, and $B(i)$ is the set of nodes neighboring i.

In our model, the GIN-based graph neural net consists of five GIN layers, each followed by a batch normalization layer. Finally, a global max pooling layer is added to obtain the graph representation vector.

2.4.4 Variant 4: GAT-GCN combined graph neural network

We also investigate a combined **GAT-GCN** model. Here, the graph neural network begins with a GAT layer that takes the graph as input, then passes a convolved feature matrix to the subsequent GCN layer. Each layer is activated by a ReLU function. The final graph representation vector is then computed by concatenating the global max pooling and global mean pooling layers from the GCN layer output.

2.5 Benchmark

To compare our model with the state-of-the-art **DeepDTA** [20] and **WideDTA** [21] models, we use the same datasets from the [20, 21] benchmarks:

- **Davis** contains the binding affinities for all pairs of 72 drugs and 442 targets, measured as K_d constants and ranging from 5.0 to 10.8 [39].

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Method	Protein rep.	Compound rep.	CI	MSE
Baseline models				
DeepDTA [20]	1D	Pubchem-Sim	0.718	0.571
DeepDTA [20]	Smith-Waterman	Pubchem-Sim	0.710	0.502
KronRLS [16, 17]	Smith-Waterman	Pubchem-Sim	0.782	0.411
SimBoost [15]	Smith-Waterman	Pubchem-Sim	0.836	0.222
DeepDTA [20]	Smith-Waterman	1D	0.854	0.204
DeepDTA [20]	1D	1D	0.863	0.194
WideDTA [21]	1D + PDM	1D + LMCS	*0.875*	*0.179*
Proposed method - **GraphDTA**				
GAT	1D	Graph	0.866	0.179
GIN	1D	Graph	**0.882**	**0.147**
GCN	1D	Graph	**0.889**	**0.139**
GAT_GCN	1D	Graph	**0.891**	**0.139**

Table 2: Prediction performance on the Kiba dataset, sorted by MSE. Baseline results are from [20, 21]. We compare 4 graph neural network variants: GIN [38], GAT [37], GCN [36], and combined GAT-GCN [37, 36]. Italics: best for baseline models, bold: better than baselines.

- **Kiba** contains the binding affinities for 2,116 drugs and 229 targets, measured as KIBA scores and ranging from 0.0 to 17.2 [40].

To make the comparison as fair as possible, we use the same set of training and testing examples from [20, 21], as well as the same performance metrics: Mean Square Error (**MSE**, the smaller the better) and Concordance Index (**CI**, the larger the better). For all baseline methods, we report the performance metrics as originally published in [20, 21]. The hyper-parameters used for our experiments are summarized in Table 3. The hyper-parameters were not tuned, but chosen *a priori* based on our past modelling experience.

Hyper-parameters	Setting
Learning rate	0.0005
Batch size	512
Optimizer	Adam
GCN layers	3
GIN layers	5
GAT layers	2
GAT_GCN layers	2

Table 3: Hyper-parameters for different graph neural network variants used in our experiments.

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.6 Model interpretation

The activation of nodes within layers of a deep neural network are called latent variables, and can be analyzed directly to understand how a model's performance relates to domain knowledge [41]. We obtained the 128 latent variables from the graph neural network layer, and analyzed them directly through a redundancy analysis. This multivariable statistical method allows us to measure the percent of the total variance within the latent variables that can be explained by an external data source. In our case, the external data source is a matrix of 38 molecular JoeLib features/descriptors [42] for each drug (available from ChemMine Tools [43]). We also compare the value of the principal components from these latent variables with the per-drug test set error. Here, the per-drug (or per-protein) error refers to the median of the absolute error between the predicted DTA and the ground-truth DTA for all test set pairs containing that drug (or that protein). For these analyses, we focus on the GIN model [38] (because of its superior performance) and the Kiba dataset [40] (because of its larger drug catalog).

3 Results and Discussion

3.1 Graphical models outperform the state-of-the-art

Table 1 compares the performance of 4 variant **GraphDTA** models with the existing baseline models for the Davis dataset. Here, all 4 variants had the lowest MSE. The best variant had an MSE of 0.229 which is 14.0% lower than the best baseline of 0.261. The improvement is less obvious according to the CI metric, where only 2 of the 4 variants had the highest CI. The best CI for a baseline model was 0.886. By comparison, the GAT and GIN models achieved a CI of 0.892 and 0.893, respectively.

Table 2 compares the performance of the **GraphDTA** models with the existing baseline models for the Kiba dataset. Here, 3 of the 4 variants had the lowest MSE and the highest CI, including GIN, GCN, and GAT-GCN. Of note, the best MSE here is 0.139, which is 28.8% lower than the best baseline of 0.179. Of all variants tested, GIN is the only one that had the best performance for both datasets and for both performance measures. For this reason, we focus on the GIN in all post-hoc statistical analyses.

3.2 Graphical models discover known drug properties

A graph neural network works by abstracting the molecular graph of each drug into a new feature vector of *latent variables*. In our model, there are 128 latent variables which together characterise the structural properties of the drug. Since the latent variables are learned during the DTA prediction task, we assume that they represent graphical features that contribute meaningfully to DTA.

Unfortunately, it is not straightforward to determine the molecular substructures to which each latent variable corresponds. However, we can regress

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

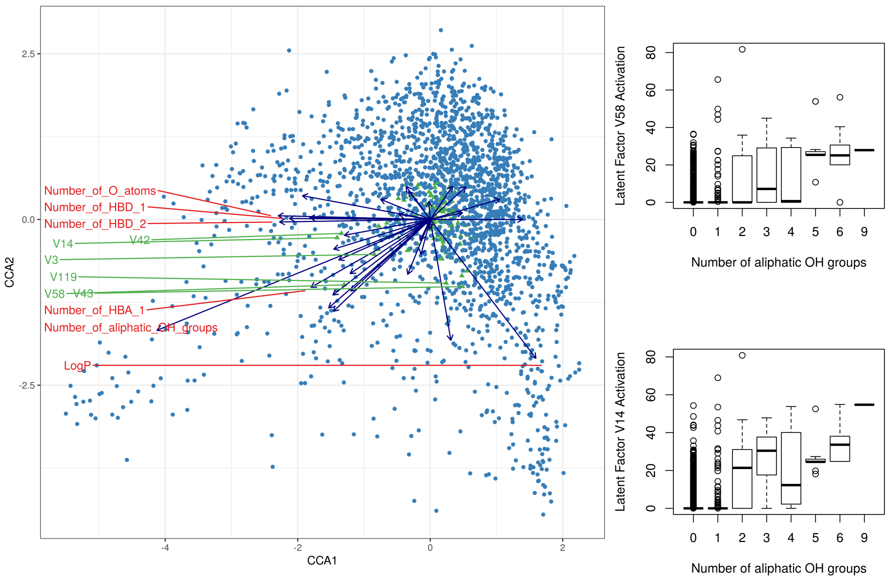

Figure 2: The left panel of the figure shows a redundancy analysis triplot for the 128 drug latent variables regressed with 38 JoeLib molecular descriptors [42]. The blue dots represent drugs, the green dots represent latent variables (the 6 furthest from origin are labelled), and the arrows represent molecular descriptors (the 5 longest are labelled). The right panel of the figure shows the activation of two latent variables plotted against the number of aliphatic OH groups in that drug. These results suggest that the graph convolutional network can abstract known molecular descriptors without any prior knowledge.

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

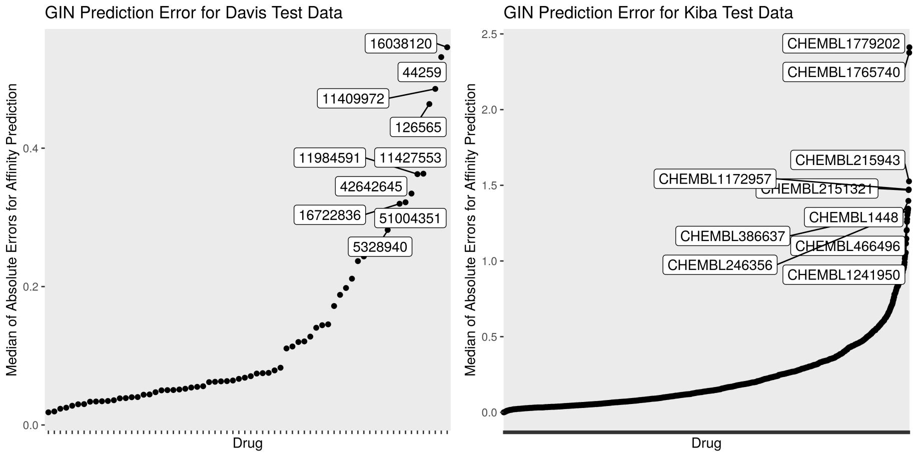

Figure 3: This figure shows the median of the absolute error for each drug, sorted in increasing order, for the Davis and Kiba test sets. Here, we see that the errors are not distributed evenly across the drugs. It is harder to predict the target affinities for some drugs than others.

the learned latent space with a matrix of known molecular descriptors to look for overlap. Figure 2 shows a redundancy analysis of the 128 latent variables regressed with 38 molecular descriptors [42] (available from ChemMine Tools [43]). From this, we find that 20.19% of the latent space is explained by the known descriptors, with the "Number of aliphatic OH groups" contributing most to the explained variance. Indeed, two latent variables correlate strongly with this descriptor: hidden nodes V58 and V14 both tend to have high activation when the number of aliphatic OH groups is large. This finding provides some insight into how the graphical model might "see" the drugs as a set of molecular sub-structures, though most of the latent space is orthogonal to the known molecular descriptors.

3.3 A few drugs contribute disproportionately to total error

Although the **GraphDTA** model outperforms its competitors, we wanted to know more about why its predictions sometimes failed. For this, we averaged the prediction error for each drug (and each protein), for both the Davis and Kiba test sets. Figures 3 and 4 show the median of the absolute error (MAE) for affinity prediction, sorted from smallest to largest. Interestingly, we see that a handful of drugs (and a handful of proteins) contribute disproportionately to the overall error. Of note, CHEMBL1779202 (an ALK inhibitor), CHEMBL1765740 (a PDK1 inhibitor) and the protein CSNK1E all had an MAE above 2.

We examined the latent space with regard to the prediction error, but could

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

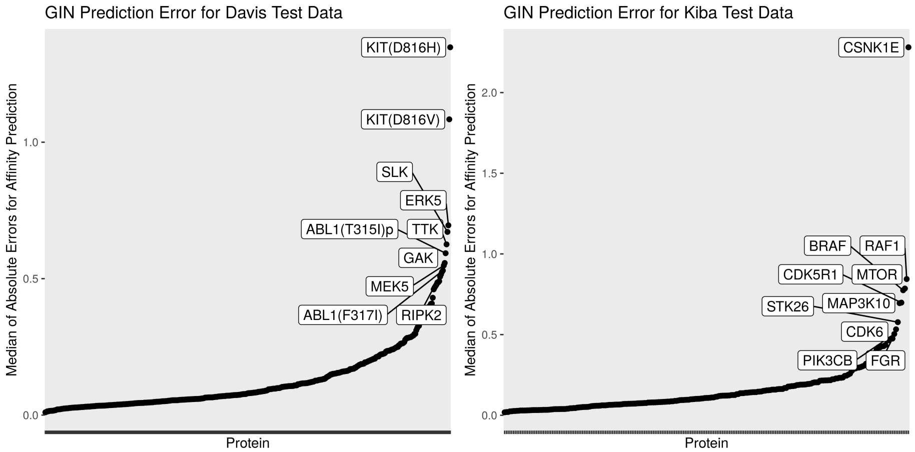

Figure 4: This figure shows the median of the absolute error for each protein, sorted in increasing order, for the Davis and Kiba test sets. Here, we see that the errors are not distributed evenly across the proteins. It is harder to predict the target affinities for some proteins than others.

not find any obvious pattern that separated hard-to-predict drugs from easy-to-predict drugs. The only trend we could find is that the easy-to-predict drugs are more likely to appear as outliers in a PCA of the latent space. Supplemental Figure 1 (https://github.com/thinng/GraphDTA/blob/master/supplement.pdf) shows the median errors plotted against the first six principal components, where we see that the hard-to-predict drugs usually appear close to the origin. We interpret this to mean that drugs with unique molecular sub-structures are always easy to predict. On the other hand, the hard-to-predict drugs tend to lack unique structures, though this is apparently true for many easy-to-predict drugs too.

3.4 Model interpretation and the research cycle

Knowing how a model works and when a model fails can feedback into the research cycle. In the post-hoc statistical analysis of our model, we find that a graph neural network can learn the importance of known molecular descriptors without any prior knowledge. However, most of the learned latent variables remain unexplained by the available descriptors. Yet, the model's performance implies that these learned representations are *useful* in affinity prediction. This suggests that there are both similarities and differences in how machines "see" chemicals versus how human experts see them. Understanding this distinction may further improve model performance or reveal new mechanisms behind drug–target interactions.

Meanwhile, the distribution of the test set errors suggest that there are "problem drugs" (and "problem proteins") for which prediction is especially

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

difficult. One could action this insight either by collecting more training data for these drugs (or proteins), or by using domain-knowledge to engineer features that complement the molecular graphs. Indeed, knowing that the PCA outliers are the easiest to predict suggests that some additional feature input may be needed to differentiate between drugs that lack distinct molecular sub-graphs. Although 2D graphs contain more information than 1D strings, our model still neglects the stereochemistry of the molecules. Future experiments could test whether representing drugs in 3D (or proteins in 2D) further improves model performance.

Interestingly, under-representation of proteins in the training set does seem to be the reason for the "problem proteins". The **Supplement**[3] shows an analysis of the effect of homologous proteins on test set performance. Although we see that test set error varies across clustered protein groups, the training set represents all protein clusters equally well. This suggests that the variation in test set performance is not simply explained by asymmetrical representation of protein groups within the training set.

4 Summary and Future Work

We test **GraphDTA** with four graph neural network variants, including GCN, GAT, GIN, and a combined GAT-GCN architecture, for the task of drug–affinity prediction. We benchmark the performance of these models on the Davis and Kiba datasets. We find **GraphDTA** performs well for two separate benchmark datasets and for two key performance metrics. In a post-hoc statistical analysis of our model, we find that **GraphDTA** can learn the importance of known molecular descriptors without any prior knowledge. We also examine the model's performance and find that a handful of drugs contribute disproportionately to the total prediction error. Although we focus on drug–target affinity prediction, our **GraphDTA** model is a generic solution for any similar problem where either data input can be represented as a graph.

It may be possible to improve performance further by representing proteins as graphs too, for example by a graph of their 3D structure. However, determining the 3D structure of a target protein is very challenging. We chose to use the primary protein sequence because it is readily available. The use of 1D sequences, instead of 3D structures, also reduces the number of parameters that we need to learn, making it less likely that we over-fit our model to the training data. Still, for some problem applications, it may make sense to use structural information, as well as binding site, binding confirmation, and solution environment information, to augment the model.

[3] https://github.com/thinng/GraphDTA/blob/master/supplement.pdf

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Asher Mullard. New drugs cost US $2.6 billion to develop. *Nature Reviews Drug Discovery*, 13(12):877, 2014.

[2] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673, 2004.

[3] Allen D Roses. Pharmacogenetics in drug discovery and development: a translational perspective. *Nature Reviews Drug Discovery*, 7(10):807, 2008.

[4] Stephen M Strittmatter. Overcoming drug development bottlenecks with repurposing: old drugs learn new tricks. *Nature Medicine*, 20(6):590, 2014.

[5] David E. Gordon, Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Matthew J O'Meara, Jeffrey Z. Guo, Danielle L. Swaney, Tia A. Tummino, Ruth Huttenhain, et al. A SARS-CoV-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. *bioRxiv*, 2020. doi: https://doi.org/10.1101/2020.03.22.002386.

[6] Philip Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nature Reviews Drug Discovery*, 1(4):309, 2002.

[7] Martin EM Noble, Jane A Endicott, and Louise N Johnson. Protein kinase inhibitors: Insights into drug design from structure. *Science*, 303(5665):1800–1805, 2004.

[8] Mukund Deshpande, Michihiro Kuramochi, Nikil Wale, and George Karypis. Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering*, 17(8):1036–1050, 2005.

[9] Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[10] Shawn J Stachel, John M Sanders, Darrell A Henze, Mike T Rudd, Hua-Poo Su, Yiwei Li, Kausik K Nanda, Melissa S Egbertson, Peter J Manley, Kristen LG Jones, et al. Maximizing diversity from a kinase screen: identification of novel and selective pan-Trk inhibitors for chronic pain. *Journal of Medicinal Chemistry*, 57(13):5800–5816, 2014.

[11] Sarah L Kinnings, Nina Liu, Peter J Tonge, Richard M Jackson, Lei Xie, and Philip E Bourne. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of Chemical Information and Modeling*, 51(2):408–419, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[12] Murat Iskar, Georg Zeller, Xing-Ming Zhao, Vera van Noort, and Peer Bork. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Current Opinion in Biotechnology*, 23(4):609–616, 2012.

[13] Steven M Corsello, Joshua A Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E Hirschman, Stephen E Johnston, Anita Vrcic, Bang Wong, Mariya Khan, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature Medicine*, 23(4):405, 2017.

[14] Jin Li, Ailing Fu, and Le Zhang. An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, pages 1–9, 2019.

[15] Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, 9(1):24, 2017.

[16] Anna Cichonska, Balaguru Ravikumar, Elina Parri, Sanna Timonen, Tapio Pahikkala, Antti Airola, Krister Wennerberg, Juho Rousu, and Tero Aittokallio. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS Computational Biology*, 13(8):e1005678, 2017.

[17] Anna Cichonska, Tapio Pahikkala, Sandor Szedmak, Heli Julkunen, Antti Airola, Markus Heinonen, Tero Aittokallio, and Juho Rousu. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, 34(13):i509–i518, 2018.

[18] Yanyi Chu, Aman Chandra Kaushik, Xiangeng Wang, Wei Wang, Yufang Zhang, Xiaoqi Shan, Dennis Russell Salahub, Yi Xiong, and Dong-Qing Wei. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Briefings in Bioinformatics*, 12 2019. bbz152.

[19] Yanyi Chu, Xiaoqi Shan, Dennis R Salahub, Yi Xiong, and Dong-Qing Wei. Predicting drug-target interactions using multi-label learning with community detection method (DTI-MLCD). *bioRxiv*, 2020.

[20] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[21] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. WideDTA: prediction of drug-target binding affinity. *arXiv*, (1902.04166), 2019.

[22] Michał Woźniak, Agnieszka Wołos, Urszula Modrzyk, Rafał L Górski, Jan Winkowski, Michał Bajczyk, Sara Szymkuć, Bartosz A Grzybowski, and Maciej Eder. Linguistic measures of chemical diversity and the ‘keywords’ of molecular collections. *Scientific Reports*, 8, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[23] Christian JA Sigrist, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Virginie Bulliard, Amos Bairoch, and Nicolas Hulo. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38(suppl_1):D161–D166, 2009.

[24] Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. PADME: A deep learning-based framework for drug-target interaction prediction. *arXiv*, (arXiv:1807.09741), 2018.

[25] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, 2018.

[26] Wen Torng and Russ B. Altman. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019.

[27] Ke Liu, Xiangyan Sun, Lei Jia, Jun Ma, Haoming Xing, Junqiu Wu, Hua Gao, Yax Sun, Florian Boulnois, and Jie Fan. Chemi-Net: A molecular graph convolutional network for accurate drug property prediction. *International Journal of Molecular Sciences*, 20(14), 2019.

[28] Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato, and Yasubumi Sakakibara. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics*, 19(19):526, 2018.

[29] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, 2016.

[30] Mengying Sun, Sendong Zhao, Coryandar Gilvary, Olivier Elemento, Jiayu Zhou, and Fei Wang. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics*, 06 2019.

[31] Hongyang Gao, Zhengyang Wang, and Shuiwang Ji. Large-scale learnable graph convolutional networks. In *Proc. of SIGKDD Conference*, pages 1416–1424. ACM, 2018.

[32] Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In *Proc. of IJCAI*, pages 3371–3377, 2018.

[33] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

bioRxiv preprint doi: https://doi.org/10.1101/684662; this version posted October 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[34] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More.* O'Reilly Media, 2019.

[35] Greg Landrum. RDKit: Open-source cheminformatics.

[36] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *Proc. of ICLR*, 2017.

[37] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *Proc. of ICLR*, 2018.

[38] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *Proc. of ICLR*, 2019.

[39] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046, 2011.

[40] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

[41] Vuong Le, Thomas P. Quinn, Truyen Tran, and Svetha Venkatesh. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. *BMC Genomics*, 2019.

[42] Jörg K Wegner, Holger Fröhlich, and Andreas Zell. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *Journal of Chemical Information and Computer Sciences*, 44(3):931–939, 2004.

[43] Tyler WH Backman, Yiqun Cao, and Thomas Girke. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Research*, 39(suppl_2):W486–W491, 2011.