MODELLING DRUG-TARGET BINDING AFFINITY USING A BERT BASED GRAPH NEURAL NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the interactions between novel drugs and target proteins is fundamentally important in disease research as discovering drug-protein interactions can be an exceptionally time-consuming and expensive process. Alternatively, this process can be simulated using modern deep learning methods that have the potential of utilising vast quantities of data to reduce the cost and time required to provide accurate predictions. In this paper, we seek to leverage a set of BERT-style models that have been pre-trained on vast quantities of both protein and drug data. The encodings produced by each model are then utilised as node representations for a graph convolutional neural network, which in turn models the interactions without the need to simultaneously fine-tune both protein and drug BERT models to the task. We evaluate the performance of our approach on two drug-target interaction datasets that were previously used as benchmarks in recent work. Our results significantly improve upon a vanilla BERT baseline approach as well as the former state-of-the-art methods for each task dataset. Our approach builds upon past work in two key areas; firstly, we take full advantage of two large pre-trained BERT models that provide improved representations of task-relevant properties of both drugs and proteins. Secondly, inspired by work in natural language processing that investigates how linguistic structure is represented in such models, we perform interpretability analyses that allow us to locate functionally-relevant areas of interest within each drug and protein. By modelling the drug-target interactions as a graph as opposed to a set of isolated interactions, we demonstrate the benefits of combining large pre-trained models and a graph neural network to make state-of-the-art predictions on drug-target binding affinity.

1 INTRODUCTION

Developing personalised medicine has been at the forefront of recent disease research, which has been accelerated with vast quantities of data being generated and refined in laboratories across the world. As new drugs and proteins are regularly being produced and discovered, it is becoming ever more challenging to utilise this data correctly and gain an understanding of the biological systems that operate within complex diseases. Modern disease research and drug discovery require new methods that can capitalise on the information that is available within these vast resources and in turn, channel this knowledge towards improving drug-target interaction simulations.

Deep learning has the potential to address the concerns of modelling such complex heterogeneous data. In recent years, deep learning has been used to model Drug-Target Interactions (DTIs) as it is ideally suited to handle large datasets without requiring feature engineering. By using deep learning to map out the drug-target landscape, one can quickly identify the proteins that are targeted by each drug – thereby accelerating drug discovery during clinical trails (Santos et al., 2017). Initial applications of machine learning models posed this as a classification problem due to the variability between each interaction pair (Bleakley & Yamanishi, 2009; Cao et al., 2014; Öztürk et al., 2016). However, these early approaches do not provide enough information about the actual binding affinity value, which is troublesome when one seeks to learn the potency of a particular drug-target pair. Deep learning now plays an important role in determining patterns in complex drug-target systems. Applications of deep learning are becoming ubiquitous in drug-drug interaction modelling (Ryu et al., 2018; Liu et al., 2016), as well as forming predictions for protein-protein interactions (Sun et al.,

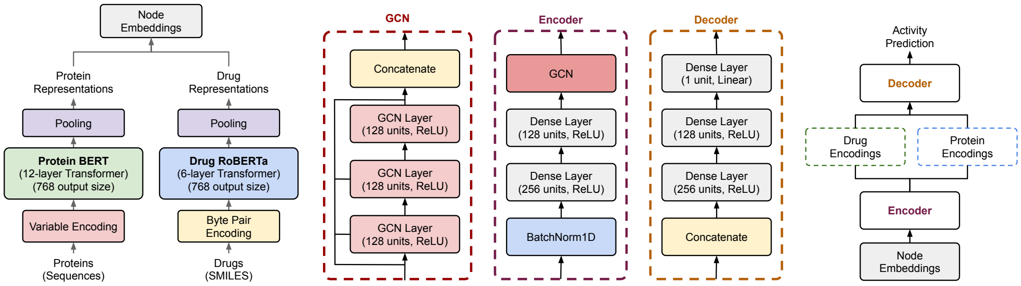

Node Embeddings
Protein Representations
Drug Representations
Pooling
Pooling
Protein BERT (12-layer Transformer) (768 output size)
Drug RoBERTa (6-layer Transformer) (768 output size)
Variable Encoding
Byte Pair Encoding
Proteins (Sequences)
Drugs (SMILES)
GCN
Concatenate
GCN Layer (128 units, ReLU)
GCN Layer (128 units, ReLU)
GCN Layer (128 units, ReLU)
Encoder
GCN
Dense Layer (128 units, ReLU)
Dense Layer (256 units, ReLU)
BatchNorm1D
Decoder
Dense Layer (1 unit, Linear)
Dense Layer (128 units, ReLU)
Dense Layer (256 units, ReLU)
Concatenate
Activity Prediction
Decoder
Drug Encodings
Protein Encodings
Encoder
Node Embeddings

Figure 1: Overview of the BERT-GCN Approach.

2017; Peng & Lu, 2017; Wang et al., 2017), and the identification novel drug-target interactions (Wen et al., 2017; Xie et al., 2018; Feng et al., 2018; He et al., 2017; Öztürk et al., 2018).

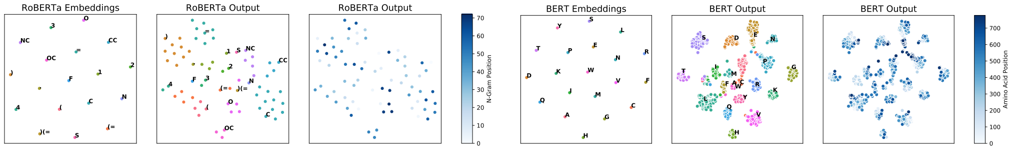
In recent work, the focus has moved away from developing classification models. Instead, the drug-target identification problem has been formulated as a regression task that requires the model to predict the binding affinity value directly. Building a regression model has the potential to rank therapeutic drugs, which limits the scope at which large compound libraries are analysed during drug discovery studies. These measured affinity values may include measurements such as dissociation constant (K_d), inhibition constant (K_i) or the half-maximal inhibitory concentration (IC_{50}).

In this paper, we will consider a BERT (Devlin et al., 2018; Rao et al., 2019) model and a RoBERTa (Liu et al., 2019) model that have been pre-trained on a large corpus of protein and drug data respectfully. During training, both models are used to provide node embeddings to the graph convolutional neural network (GCN) that is applied to model the interactions between the drug-target pairs. Our approach improves upon past work in two key areas; firstly, our method capitalises on two pre-trained BERT style networks, which provide robust embeddings for each drug and protein. These models can also be used to visualise the critical areas of interest within each drug-target pair. Such insights will benefit the field of computational biology as it becomes easier for the end-user to distil the knowledge from these models. Secondly, our method implements a GCN to model the interaction between individual pairs as opposed to past work that use a simple multiply layer perceptron (MLP) to produce a prediction for the binding affinity value of each interaction.

In most cases, the total number of unique drugs and proteins tested during these experiments is limited, which does not provide a complete depiction of how a particular drug or target protein might operate under the same experimental conditions. We seek to address these issues through the use of pre-training and graph neural networks. End-to-end, our approach will be able to encode any drug-protein even it is not present within the original datasets. By implementing this style of modelling our approach will be able to analyse and determine the essential features within each protein and drug sequence without causing the models to overfit to the limited number of labelled examples observed during training.

2 Related Work

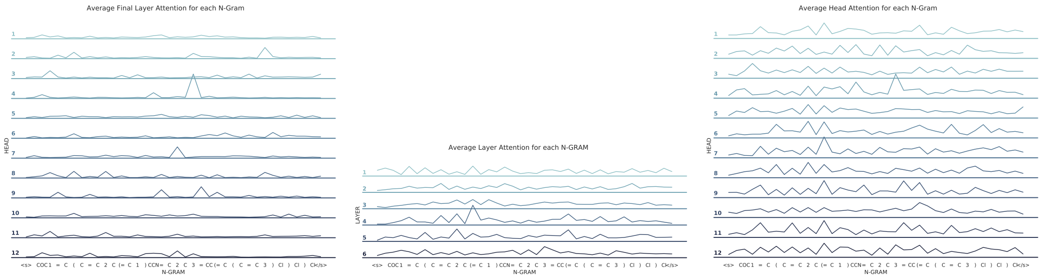
Previous machine learning methods relied on scoring functions and a series of feature-engineered steps to transform the original drug-protein pair before producing a final prediction (Ballester & Mitchell, 2010; Li et al., 2015; Shar et al., 2016). These approaches did not generalise well, as the machine learning models were optimised on a feature set engineered from only on a few observations. This limited scope provided less information about the raw interaction between the drug and the protein (Gabel et al., 2014). Examples include Kronecker Regularised Least Squares (Kron-RLS) (Pahikkala et al., 2014) algorithm that utilised drug similarity information and a Smith-Waterman similarity representation (Smith et al., 1981) of each target protein to model interaction values by formulating it as a regression problem (Smith et al., 1981). This kernel approach performed well, given its lack of complexity, which in turn stopped the model from overfitting during training. Later, He et al. (2017) designed a gradient boosting machine learning model (Chen &



(a) Drug RoBERTa Embeddings

(b) Protein BERT Embeddings

Figure 2: A set of t-SNE plots showcasing the initial embeddings and final encodings for an example drug and protein (see text for details).



(a) Final Layer

(b) Individual Layers

(c) Individual Heads

Figure 3: Average Attentions within the Drug RoBERTa model.

Guestrin, 2016) that was trained using network-based features from the observed drugs, targets and drug-target interactions from each dataset. The training data was based on the drugs and targets, which formed the nodes of the graph, while the binding affinity values represented the edges. The Simboost algorithm was a significant jump from the Kron-RLS as it included a far more extensive and rigorous feature set, while also utilising a more sophisticated machine learning algorithm. However, both methods share the same constraint as they are only capable of modelling a summarisation of the raw data available for the drug-target interaction.

Deep learning is now a prominent application of machine learning, as these models do not require the same level of feature engineering. Recent work in the subfields of natural language processing (NLP) (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018), and computer vision (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017), has produced state-of-the-art results using deep learning approaches, and have recently revealed the value of pre-training these model as they typically outperform the original baselines. However, such methods have yet to be rigorously tested within the context of drug-protein modelling. One of the main drawbacks to applying deep learning is the sacrifice of interpretability, as it becomes increasingly challenging to distil the knowledge of the model. Öztürk et al. (2018) designed a deep learning model for a set of DTI regression tasks that aimed to predict the binding affinity scores by utilising a set of convolutional neural networks (CNN). The proposed model was comprised of two individual three-layer CNNs that were adopted to encode the drug (i.e. SMILES strings) and target (i.e. protein sequences) respectively. The final features produced by each CNN were then max pooled and concatenated together before finally being passed through a multi-layer perceptron (MLP) to form a prediction for the binding affinity. Since the DeepDTA model incorporates CNN models to encode both the drug and the protein, it could only capture local dependencies within the SMILES strings and protein sequences.

Later, Shin et al. (2019) improved upon the DeepDTA model by replacing the drug CNN component with a pre-trained character transformer, that unified both transformer and convolutional neural networks. The drug transformer was pre-trained using the PubChem database and was a clear improvement over a CNN as it was better at capturing long-range dependencies in the drug sequence. Such a characteristic is vital to model intermolecular interactions properly, as a deep learning model should be able to incorporate all the information about the structure of both the drug and the protein. However, the Shin et al. (2019) model still included a set of convolutional layers designed to extract features from the protein, and therefore suffered at accurately modelling the complete sequence of amino acids. It should be noted that the transformer within the Shin et al. (2019) model required further fine-tuning to each interaction dataset before it could produce better results over the DeepDTA model.

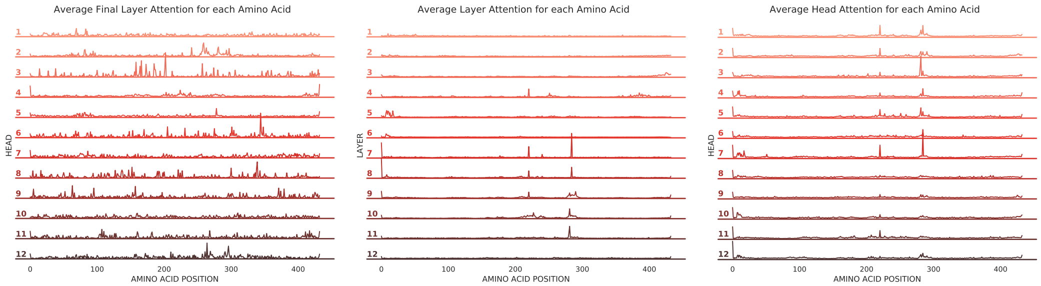
Graph convolutional neural networks (GCNs) have also been used to encode the molecular graph, whereby the atoms are the nodes, and the bonds are the edges of the graph. (Duvenaud et al., 2015) implemented a GCN model to replicate circular fingerprints, which could extract relevant molecular features. (Kearnes et al., 2016) likewise presented a molecular graph convolutional model for learning small molecules. (Coley et al., 2019) use a GCN based approach to model the interactions between organic reactions to predict the products. As an example of white-box deep learning, Kearnes et al. (2016) were able to gain insight and derive knowledge from the model's predictions, which was later validated by experts. Although (Coley et al., 2019) model was tailored to modelling drug reactions, it could be modified to integrate target protein information so that the most active drugs can be determined. These applications display the potential for future applications of deep learning within the virtual screening process.

3 Materials and Methods

The modelling scheme used in this investigation is broken down into two parts, as outlined in Figure 1. Firstly each drug-protein pair is encoded into a vector representation by the pre-trained BERT and RoBERTa models respectfully. Our work improves upon past work by employing both of these state-of-the-art pre-trained models to provide robust representations for each drug and protein. The protein BERT model (Devlin et al., 2018; Rao et al., 2019) was trained with masked-token prediction based on a large corpus of protein sequences collected from the recently curated Pfam database (El-Gebali et al., 2018), which has the proteins organised into clusters that share evolutionary-related groups (i.e. families). This corpus holds approximately thirty-one million protein domains that have been extensively used in bioinformatics. It now forms the corpus used to train large sequence models, such as TAPE (Rao et al., 2019). The protein BERT model consisted of twelve-layers with a hidden size of 512 units and 12 attention heads. Each protein is encoded using a standard variable encoding scheme with the complete vocabulary containing a total of thirty characters, including the special characters. In post hoc model interpretation of the TAPE transformer, Vig et al. (2020) explored how this particular model was capable of discerning structural and functional properties about proteins. As the BERT transformer was able to model long-range dependencies within the protein, it was able to deduce information about the protein based on the folding structure, target binding sites, and additional complex biophysical properties. Vig et al. (2020) concluded that the specific heads within the model attended to individual amino acids, as the attention similarity matrix was highly correlated to the expected substitution scores (i.e. BLOSUM62) for each amino acid. They also noted that the deeper layers of the BERT model focused relatively more attention on binding sites and contacts (i.e. high-level concepts). In contrast, information about the secondary structure (i.e. low- to mid-level concepts) within the protein was targeted evenly across each of the layers.

The drug RoBERTa model was trained on 250,000 Simplified Molecular Input Line Entry System (SMILES) strings from the ZINC15 database of drug-like molecules (Irwin & Shoichet, 2005), again using masked-token prediction. From the raw SMILES strings, the drugs were tokenised using a Byte-Pair Encoder (BPE), via the Huggingface tokeniser library (Wolf et al., 2019), which is one of the most commonly applied subword encoding algorithms in natural language processing. Subword algorithms such as BPE can decompose rare words into frequently occurring subwords, which allows DNNs to model large vocabularies without hindering the model's performance with out-of-vocabulary words. In our context, the subword encoding algorithm breaks the SMILES string into commonly occurring subsequences, and it is then able to find the most optimal vocabulary by iteratively merging symbols within the original SMILES string until the best segmentations was determined (Shibata et al., 1999). A set of additional tokens were also included within this vocabulary (i.e. to denote special tokens for unknown characters, padding, separation and masked characters) such as to avoid unknown tokens during our pre-training stage.

In Figure 2, we have visualised both the original embeddings and final encodings for an example drug and protein using both respective models. In Figure 2a, we observe how the drugs encodings naturally group together as the carbon-based molecules cluster towards the carbon atom encodings. Unsurprisingly in Figure 2b, we see each amino acid forming their own individual cluster as the final



(a) Final Layer

(b) Individual Layers

(c) Individual Heads

Figure 4: Average Attentions within the Protein BERT model.

layer of the model is tailored for predicting the masked tokens. The amino acids with hydrophobic side chains appear to gather in the centre of the plot while the amino acids that can be either electrically charged or uncharged are clustered at the edges of the plot. In Figures 3a-4a, we visualise the average attention across the final layer of each model for another example drug-target pair. From these plots, we can see that no two heads of the final layer attend to the same information within either drug or protein. By taking an average across each layer and head within the models in Figures 3b-3c and 4b-4c respectfully, we can uncover what part of each drug and protein is most influential to its final vector representation. From Figures 3b-3c, one can see that RoBERTa drug model pays special attention to the positioning of the brackets within the in SMILES string and the number of carbons included in each branch. This is to be expected as the model progressively learns the overall structure of the drug from the sequence of n-grams. When considering Figures 4b-4c, the protein BERT model appears to pay more attention to specific parts of the protein sequence in the later layers (e.g. binding sites and contacts), which is in line with the finds from Vig et al. (2020).

3.1 Modelling

As in previously work, we will model the drug-target interaction as a regression task, whereby the model must produce predictions of the binding affinity scores. As mentioned, our approach adopts a pair of pre-trained BERT models that contain six and twelve layers for the drug and protein model respectfully, as outlined in Figure 1. In both models, these layers are then followed by a final pooling layer to produce a vector representation for each drug and protein. The BERT networks are the more suitable approach to learn from either sequence since the architecture has attention blocks that allow effective learning from arbitrary sequence lengths. In addition to the pre-trained BERT models, our approach includes graph convolutional layers (GCN) layers. The most powerful feature of GCN models is their ability to capture the local neighbourhood information about each drug-protein as these final representations are then utilised to predict the binding affinity scores. Unlike past examples, we do not truncate either drug or protein when producing the final encodings. This will allow the graph neural network to learn a complete representation of each drug and protein, and thereby avoid training our algorithm with excessive amounts of padding. Once the model has collected local features based on the original BERT embeddings followed by the output of each of the GCN layers, these final features are then concatenated once more to form the drug-protein interaction pairs. Residual connections are employed between each GCN layer to improve training and the overall performance of the model. These interaction features are then passed through a set of dense layers to reduce the final interaction features into a prediction for the binding affinity scores. Mean squared error (MSE) is used as a loss function as we optimise our model via the Adam optimisation algorithm (Kingma & Ba, 2014), with the default learning rate of 0.001. Once the network has been suitably trained, it can then encode each drug-protein pair and analyse how the observed interactions dictate the affinity values.

3.2 Tasks

Following citetozturk2018deepdta, our approach was evaluated on two separate benchmark datasets, the Davis kinase dataset (Davis et al., 2011) and the KIBA dataset (Tang et al., 2014), as summarised in Table 1. For both datasets, the drugs were represented in the SMILES string format and

Table 1: Summary of the two downstream tasks (Öztürk et al., 2018).

	Proteins	Drugs	Interactions
Davis (K_d)	442	68	$30,056$
KIBA	229	2111	$118,254$

Table 2: Results for the two downstream tasks.

Dataset	Method	MSE (std)	CI (std)	r_m^2 (std)	AUPR (std)
Davis	GCN-BERT (Ours)	**0.199 (0.003)**	**0.896 (0.002)**	**0.741 (0.002)**	**0.806 (0.007)**
	MLP-BERT (Ours)	0.311 (0.009)	0.862 (0.004)	0.589 (0.022)	0.721 (0.009)
	MT-DTI (Shin et al., 2019)	0.245 (n/a)	0.887 (0.003)	0.665 (0.014)	0.730 (0.014)
	DeepDTA (Öztürk et al., 2018)	0.261 (n/a)	0.878 (0.004)	0.630 (0.017)	0.714 (0.010)
	SimBoost (He et al., 2017)	0.282 (n/a)	0.872 (0.002)	0.644 (0.006)	0.709 (0.008)
	KronRLS (Pahikkala et al., 2015)	0.379 (n/a)	0.871 (0.001)	0.407 (0.005)	0.661 (0.010)
Kiba	GCN-BERT (Ours)	**0.149 (0.001)**	**0.888 (0.001)**	**0.761 (0.009)**	**0.838 (0.003)**
	MLP-BERT (Ours)	0.282 (0.005)	0.803 (0.002)	0.580 (0.008)	0.748 (0.008)
	MT-DTI (Shin et al., 2019)	0.152 (n/a)	0.882 (0.001)	0.738 (0.006)	0.837 (0.003)
	DeepDTA (Öztürk et al., 2018)	0.194 (n/a)	0.863 (0.002)	0.673 (0.009)	0.788 (0.004)
	SimBoost (He et al., 2017)	0.222 (n/a)	0.836 (0.001)	0.629 (0.007)	0.760 (0.003)
	KronRLS (Pahikkala et al., 2015)	0.411 (n/a)	0.782 (0.001)	0.342 (0.001)	0.635 (0.004)

were downloaded using their individual PubChem CIDs to query the Pubchem compound database (Bolton et al., 2008). For the protein sequences, the accession number of each protein was used to locate and extract the protein sequence from the UniProt protein database (Apweiler et al., 2004). The Davis dataset includes interactions between a subset of selectivity assays from a kinase protein family and a set of inhibitors, which were measured using the dissociation constant (K_d) across 442 unique proteins and 68 unique drugs. As in previous work (He et al., 2017; Öztürk et al., 2018), we use the log-transform of the K_d values. As the majority of the Davis dataset is inactive, it leads to a highly unbalanced distribution as a majority of the interactions either have such a low binding affinity value (i.e. $K_d > 10,000$ nM) or was not observed in the primary screen (Pahikkala et al., 2015).

The KIBA dataset was filtered during the Simboost study to yield a total of 229 unique proteins and 2,111 unique drugs (He et al., 2017). This dataset was designed to cover the bioactivity of specific kinase inhibitors from various studies, which were combined to include interactions based on K_i, K_d and IC_{50} values (Tang et al., 2014). For more information on how the KIBA scores were generated, please see citations (He et al., 2017).

The value of using a GCN approach to model the drug-target interactions as apposed to fine-tuning both BERT models simultaneously is realised when we consider the typical sequence length of either drug or protein. In the Davis dataset, the maximum drug length is 103 (average: 64), and a maximum protein length of 2,549 (average: 788). While the KIBA dataset, the maximum drug length is 590 (average: 58), and a maximum protein length of 4,128 (average: 728). To fine-tune BERT models of this size would become very computationally expensive as a considerable amount of padding would be required during this fine-tuning stage, which would increase the time required to optimise both models to the interaction task.

4 Results and discussion

To properly evaluate the predictive performance of our model, we calculated the mean square error (MSE), Concordance Index (CI) (Gönen, 2012), r_m^2 index, and Area Under Precision-Recall (AUPR) (i.e. utilised for binary predictions) scores for all predictions, as shown in Table 2. To calculate the AUPR scores for either dataset, the binding activity values were binarised by selecting a particular threshold value. Following previous evaluation using these datasets for the SimBoost and DeepDTA models (He et al., 2017; Öztürk et al., 2018), a threshold value of 7 was used for binarising pK_d values in the Davis dataset and a value of 12.1 was used for the KIBA dataset.

For a fair comparison to the previous models, we also used a five-fold cross-validation procedure to validate the performance of our approach and averaged the test scores across all five folds. To

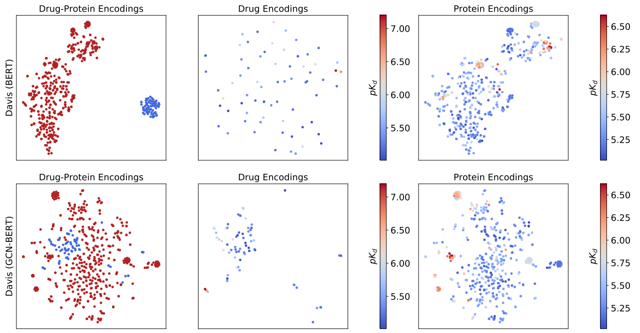


Figure 5: Davis t-SNE plots (see text for details).

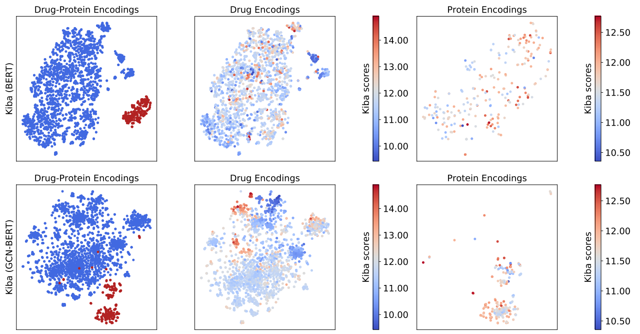


Figure 6: KIBA t-SNE plots (see text for details).

provide an unbiased final measure of performance in each dataset, we evaluated our approach on the same independent test set that was used in previous studies. We then tested our BERT-GCN approach against just using a vanilla BERT approach to modelling the interaction (MLP-BERT). In addition to the DeepDTA and SimBoost models, we also compared our model to the KronRLS algorithm, which like SimBoost is based on employing similarity matrices for both drugs and proteins as input features to the model (Pahikkala et al., 2015). We also examined the MT-DTI approach that learned only drug sequence representation with a BERT block and retained a similar CNN to encode each protein, much like the DeepDTA model (Shin et al., 2019). Table 2 reports the average score for each of these metrics across cross-validation folds for both datasets.

When the vanilla BERT encodings were used as inputs to an MLP, there appeared to be no performance benefits from using these large pre-trained networks. This result is unsurprising as neither the protein BERT model nor the drug RoBERTa model was fine-tuned simultaneously to model drug-protein interactions. As mentioned, given that protein and drug sequences can be considerably long, it would have been far too computationally expensive to run both models to perform fine-tuning. However, to capitalise on the pre-training that was conducted for both BERT networks. We were motivated to find a solution that used both pre-trained models but was also computationally feasible to the end-user. To improve upon the previous pre-trained results, we tested the combination of using these pre-trained models in coordination with a set of GCN layers, which would then model the interactions between the proteins and drugs. This method outperformed all baseline methods with the lowest average MSE scores for both the Davis and Kiba datasets and likewise achieving the highest CI, r_m^2 and AUPR scores on both datasets. In Figures 5-6, we visualise the effect of using a GCN model to enhance the original pre-trained encodings. Unsurprisingly, we see that the original pre-trained encodings do not cluster well concerning their average binding affinity. While in for the GCN counterparts, we observe both drug and protein encodings clusters towards one another as the most active and least active drug-proteins begin to form sub-clusters within each plot.

5 Conclusions

In this paper, we proposed a deep learning model that was capable of accurately predicting drug-target binding affinity values. By adopting a pair of pre-trained BERT models along with and a graph convolutional neural network, and without using any prior knowledge on the biochemistry of these interactions, this model was able to encode the sequence representations of both drugs and targets to produce state-of-the-art results. We evaluated our approach on two benchmark datasets and compared our model to previous state-of-the-art machine learning and deep learning baselines. Our results indicated that the predefined features produced by the BERT models alone could not sufficiently be applied to represent a drug-target interaction. However, when additional GCN layers were used to learn each interaction as a component of a more extensive network, the performance increased significantly compared to baseline methodologies for both datasets. Without the need to directly fine-tuning both BERT models to the DTI task, we were able to improve performance by using a graph neural network to overcome this computationally expensive process.

By analysing only the string representations of each drug and a protein respectfully, this study provides a method that utilises state-of-the-art pre-trained models to produce the most accurate interaction network for binding affinity prediction. Our approach not only saves time and computational resources with regards to training, but it also provides the best overall performance when compared to past state-of-the-art approaches that required additional feature engineering. In future work, we aim to utilise better pre-trained models that apply subword encoding algorithms during pre-training, along with building an interpretable graph neural network system that operates on these pre-trained encodings to provide improved predictions for novel interactions.

References

Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl_1):D115–D119, 2004.

Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.

Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 2009.

Evan E Bolton, Yanli Wang, Paul A Thiessen, and Stephen H Bryant. Pubchem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*, volume 4, pp. 217–241. Elsevier, 2008.

Dong-Sheng Cao, Liu-Xia Zhang, Gui-Shan Tan, Zheng Xiang, Wen-Bin Zeng, Qing-Song Xu, and Alex F Chen. Computational prediction of drug target interactions using chemical, biological, and network features. *Molecular informatics*, 33(10):669–681, 2014.

Tianqi Chen and Carlos Guestrin. Xgboost: a scalable tree boosting system in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.. acm, 2016.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10(2):370–377, 2019.

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):1046–1051, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.

Qingyuan Feng, Evgenia Dueva, Artem Cherkasov, and Martin Ester. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv preprint arXiv:1807.09741*, 2018.

Joffrey Gabel, Jérémy Desaphy, and Didier Rognan. Beware of machine learning-based scoring functions on the danger of developing black boxes. *Journal of chemical information and modeling*, 54(10):2807–2815, 2014.

Mehmet Gönen. Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics*, 28(18):2304–2310, 2012.

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):1–14, 2017.

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

Hongjian Li, Kwong-Sak Leung, Man-Hon Wong, and Pedro J Ballester. Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*, 20(6): 10947–10962, 2015.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016, 2016.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. A comparative study of smiles-based compound similarity functions for drug-target interaction prediction. *BMC bioinformatics*, 17(1):128, 2016.

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

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2014.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2015.

Yifan Peng and Zhiyong Lu. Deep learning for extracting protein-protein interactions from biomedical literature. *arXiv preprint arXiv:1706.01556*, 2017.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. In *Advances in Neural Information Processing Systems*, pp. 9689–9701, 2019.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 115(18):E4304–E4311, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1803294115. URL `https://www.pnas.org/content/115/18/E4304`.

Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19, 2017.

Piar Ali Shar, Weiyang Tao, Shuo Gao, Chao Huang, Bohui Li, Wenjuan Zhang, Mohamed Shahen, Chunli Zheng, Yaofei Bai, and Yonghua Wang. Pred-binding: large-scale protein–ligand binding affinity prediction. *Journal of enzyme inhibition and medicinal chemistry*, 31(6):1443–1450, 2016.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, Technical Report DOI-TR-161, Department of Informatics, Kyushu University, 1999.

Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. *arXiv preprint arXiv:1908.06760*, 2019.

Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, 18(1):277, May 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1700-2. URL `https://doi.org/10.1186/s12859-017-1700-2`.

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.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Yan-Bin Wang, Zhu-Hong You, Xiao Li, Tong-Hai Jiang, Xing Chen, Xi Zhou, and Lei Wang. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Molecular BioSystems*, 13(7):1336–1344, 2017.

Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, 16(4): 1401–1409, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pp. arXiv–1910, 2019.

Lingwei Xie, Song He, Xinyu Song, Xiaochen Bo, and Zhongnan Zhang. Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC genomics*, 19(7): 667, 2018.