Drug Discovery with Deep Learning

Muhammad Waheed Ramzan
PhDEE19004
Information Technology University
Lahore, Pakistan
phdee19004@itu.edu.pk

Aqsa Tariq
MSDS19075
Information Technology University
Lahore, Pakistan
msds19075@itu.edu.pk

Zoya Naseer Hashmi
MSDS19005
Information Technology University
Lahore, Pakistan
msds19005@itu.edu.pk

Fatima Farooq Bhatti
MSDS19081
Information Technology University
Lahore, Pakistan
msds19081@itu.edu.pk

Mohsen Ali
Information Technology University
Lahore, Pakistan
mohsen.ali@itu.edu.pk

Abstract

In the wake of the current pandemic of COVID19 it is ever more important to speed up the reliable drug discovery process. Deep learning has been introduced into the field of Cheminformatics to predict the Binding Affinities from pairs of drugs and target proteins. The drugs and proteins can both be represented as text sequences, graphs and 3D structures. Additionally, meaningful molecular images that hold atomic and bonding information of the molecule are also viable representations. Images of Proteins however don't hold any meaningful information. Image, Graph, 3D and Text based representations for molecular inputs have been considered for the task of DTA, but proteins have mostly been used as character sequences or expensive 3D crystal structure is used, which is not suitable for the prompt and low budget prediction task. The baseline we followed uses the protein sequence information to estimate the contact maps for proteins which can be used to generate protein graphs. In our project, we explore the limitations of their method, and perform a series of experiments with the aim to improve upon their results. Finally, we predict the binding affinities for COVID Protienase 3CL-PRO with the drugs from baseline Davis dataset, using our experimental models. The contact map for the COVID Proteinase was estimated from the MapPred webserver.
link to the github for our code is provided.

1. INTRODUCTION

The recent outbreak of pandemic caused by SARS CoV-2 has urged the scientific community to identify drugs that can act as inhibitors for this virus. In-silico (computational) techniques have been leveraged for their efficiency to narrow the drug search space of potential inhibitors for laboratory experimentation. A huge body of contributions can be found that employ both simulation based [1] and machine/deep learning based[2] approaches to identify the best binding drugs for SARS CoV2.
The efficacy of a drug for inhibition action against a protein, is measured in terms of binding affinity. Drug-Target Affinity Prediction is the branch of Cheminformatics that predicts the binding affinities of drugs for the target proteins using data driven approaches.

1.1. Drug Target Affinity (DTA)

When two chemical entities bind, in our case being the target protein and the drug molecule, old bonds are broken to form newer bonds. This can result in inhibition of certain chemical reactions that the proteins were originally involved in. So, the action of the drug can be measured in terms of the inhibition in the original function of the protein or dissociation (breaking of bonds) of the protein by the introduction of drug molecules.

These chemical reactions (dissociation and protein function) are represented by chemical equations, and equilibrium measures (equilibrium concentrations and equilibrium constants) of these chemical reactions are used to represent strength of the binding affinities of drug-target pairs.

The binding affinities are measured in terms of equilibrium concentrations in units of micro Molars, or equilibrium constants that are ratios of concentrations of reactants and products of a chemical reaction when the reaction is at equilibrium. Dissociation constant (K_d) , Inhibition constant (K_i) and half maximal inhibitory concentration (I_{50}^{2}) are the three common measures of binding affinity.

Data driven approaches can be used to regress for the binding affinity measures, using drug-target pairs as inputs to deep learning models.

1.2. Problem Statement

Accurate input representation, specially accurate protein representation, for the task of affinity prediction is an open problem.

2. LITERATURE REVIEW

Machine learning techniques like SVM (Support Vector Machine) and RF(Random Forecast) were already being used with handcrafted features like molecular fingerprints for Drug Target Affinity prediction.

With the introduction of deep learning, hand crafted features have been replaced with learned complex higher order features. The input representations, play an important role in the type of features that are learned.

Deep learning techniques were introduced to directly process drug SMILES strings with Natural Language Processing(NLP) techniques, through Recurent Neurall Network (RNN), 1D CNN [3],[4] or Transformer[5] based approaches for representation learning[6]. Around the same time graph-based representation learning from molecule graphs[7],[8] was also gaining ground. Only recently however, image based features[9] have been explored and have proven to hold remarkable performance.

Although great effort has been expended on molecular feature representation, very little attention has been given to better representation learning for proteins. The FASTA format proteins just capture the peptide chain structure, where in fact proteins are topologically more complex and exist in the form of globules. One of the papers that we are following[10] uses protein sequence alignment to estimate the contact map for protein residues and represents proteins as graphs. We will further explore the limitations and possible improvements to the literature in our methodology section.

3. METHODOLOGY

For our problem, we need to estimate the binding affinities of known drugs with a protein that is novel. It can be naturally formulated as inference on a DTA prediction model.

3.1. Baselines

We followed two papers on DTA as baselines, that worked with different input representations. [10] proposed DGraphDTA that uses protein and molecular graphs, projected to embedding space with Graph Neural Networks to regress a numeric affinity measure for each pair of drug molecule and target protein. The DgraphDTA model is shown in Figure1.

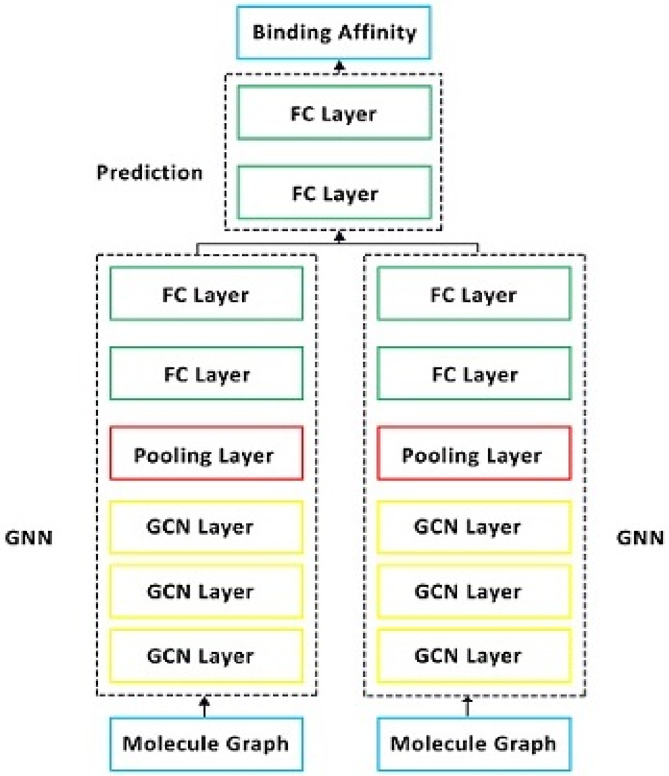

Figure 1. DGraphDTA model

They worked with two benchmark datasets Davis and KIBA, for our experiments we only used Davis dataset. Davis is a dataset 68 unique drug molecules and 442 unique proteins. 68 x 442 pairs of these molecules and proteins are labeled with the measures of binding affinities in terms of Dissociation Constant K_d in units of micro Molars. The molecule and drug representations in this dataset are in the forms of character strings. The molecules are in SMLES format (simplified molecular input line entry system), refer to Figure 2. SMILES are a standard in chemistry used to represent atoms and bonds in a molecule by unique characters. And Proteins are provided in FASTA format, which is a string of characters that represents the chain of residues in the protein, each unique character representing a unique residue. A sample from the dataset is given in Table 1
Other molecular representations can be visualized in Figure 2

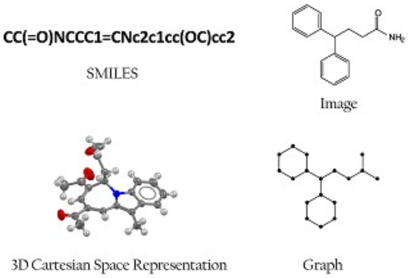
CC(=O)NCCC1=CNc2c1cc(OC)cc2
SMILES
NH2
Image
3D Cartesian Space Representation
Graph

Figure 2. Molecules Represention

The paper uses RDKit a python library to convert the drug molecule SMILES to graph representation. And Protein Sequence Alignment to convert the protein chains into graphs as shown in Figure 3.

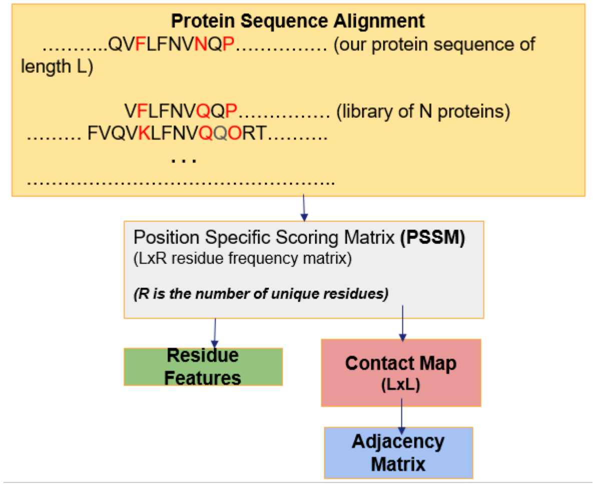
Protein Sequence Alignment
...........QVFLFNVNQP............... (our protein sequence of length L)
VFLFNVQQP............... (library of N proteins)
.... FVQVKLFNVQQORT..........
. . .
....
Position Specific Scoring Matrix (PSSM)
(LxR residue frequency matrix)
(R is the number of unique residues)
Residue Features
Contact Map (LxL)
Adjacency Matrix

Figure 3. Protein Sequence Alignment

Molecule dictionary	{"11314340": "CC1=C2C=C (C=CC2=NN1)C3=CC (=CN=C3)OCC(CC4= CC=CC=C4)N",...}
Protein dictionary	{"AAK1": "MKKFFDSR REQGGSGLGSGSSGG GGSTSGLGSGYIGRV F. . . ",..}

Table 1. A molecule (drug)[2]and protein samples from Davis Dataset

[9] used molecular images as inputs to their DEEPScreen model and solved a simpler problem of binary classification for activity/inactivity against a single target protein, using as input the positive and negative samples of a single protein only.

The model of DEEPScreen Figure 4 is a stack of five convolution layers, with 2 fc layers, before the last binary classification layer. DEEPScreen, since it does not train on pairs of drugs and targets, can't be used for inference on our problem. But the baseline served the important purpose of comparison of input representations and gains in performance for various experiments.

For all experiments the hyperparameter were kept same as provided by the paper, unless mentioned otherwise. The only hyperparameter that we failed to control was the number of epochs, due to the limitation of compute resources and unreliable colab sessions.

4. Experiments and Results

Against these baselines we designed a series of experiments to determine the advantages and disadvantages of each input representation, and performed a few architectural modifications as well. The list of experiments is follows:

- Replacing molecular graphs with molecular images in DGraphDTA
- Replacing molecular images with molecular graphs in DEEPScreen
- Combining molecular images with molecules graphs in DGraphsDTA

[2] In above table, Molecule key is drug ID reference from https://pubchem.ncbi.nlm.nih.gov/compound/. From their website exact name of drugs can be found using the these keys. Protein key is the standard keys used in Davis dataset.

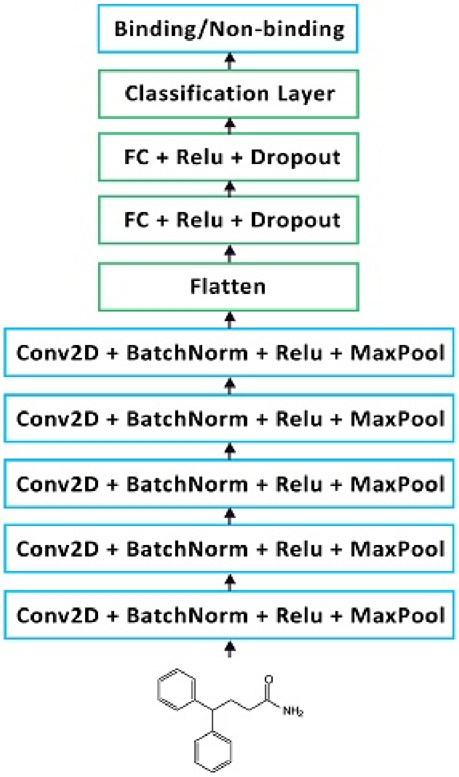
Binding/Non-binding
Classification Layer
FC + Relu + Dropout
FC + Relu + Dropout
Flatten
Conv2D + BatchNorm + Relu + MaxPool
Conv2D + BatchNorm + Relu + MaxPool
Conv2D + BatchNorm + Relu + MaxPool
Conv2D + BatchNorm + Relu + MaxPool
Conv2D + BatchNorm + Relu + MaxPool
O
NH2

Figure 4. DeepScreen Model

- Applying edge dropout to the Graph Neural Network (GNN) branch of proteins in DGraphDTA

For each experiment we modified the original code provided with the paper, for custom preprocessing, data loading and model definition, suitable for the new input representation.

4.1. Experiment 1

For the first experiment the the InMemoryDataset defined in original code by DGRaphDTA was modified to accept images of molecules in place of molecule graphs. Loading images of 200 x 200 resolution for 68 molecules and the protein graphs for 442 proteins in the Davis dataset, in the memory left little RAM space for the training. The original DGraphDTA model was trained with the batch size of 512, and original DEEPScreen model was trained with images of resolution 200 x 200. With high RAM consumption of our modified dataset the maximum batch size we could use was 128 with image resolution of 100 x 100, to avoid out of memory error on CUDA.The modified model can be seen in Figure 5 .The initial training seemed very promising, with the validation loss Figure 6 falling at a better rate than the original DGraphDTA model.

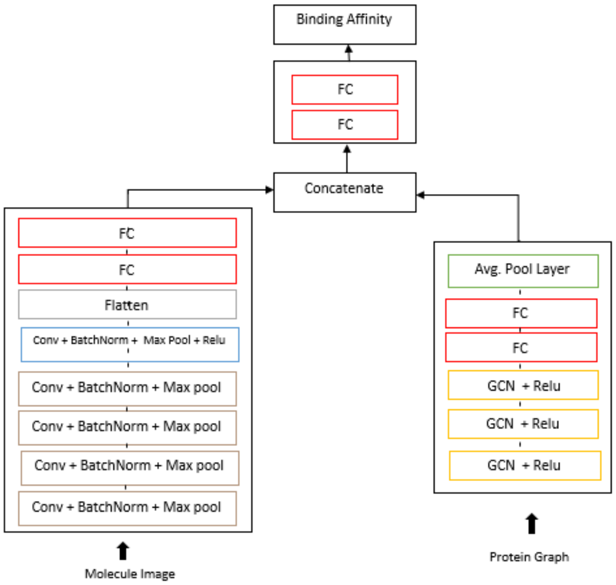
Binding Affinity
FC
FC
Concatenate
FC
FC
Flatten
Conv + BatchNorm + Max Pool + Relu
Conv + BatchNorm + Max pool
Conv + BatchNorm + Max pool
Conv + BatchNorm + Max pool
Conv + BatchNorm + Max pool
Molecule Image
Avg. Pool Layer
FC
FC
GCN + Relu
GCN + Relu
GCN + Relu
Protein Graph

Figure 5. DGraphDTA with CNN

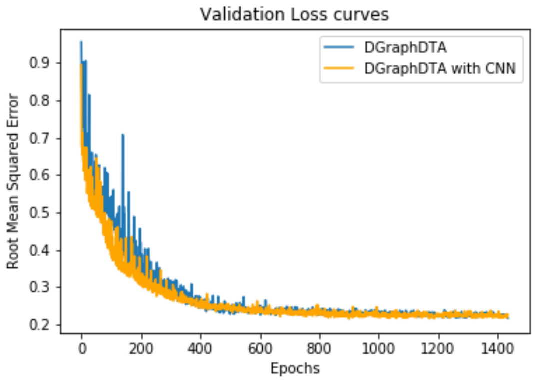

Figure 6. Loss Curve-Replacing Molecule Graphs with Images (100x100) in DGraphDTA

But the test results were seriously degraded, as can be seen in CNN (100 x 100) row of the DGraphDTA experiments Table 2.

This led to a series of more experiments designed to identify the root cause of the problem. The first hunch was that reduced image size could be causing the performance degradation. To verify this hypothesis, the DEEPScreen model was trained with lower image resolution at 100 x 100. And the loss curve on the validation set Figure 7 being very oscillatory, verified our hypothesis that lower image resolution was resulting in the degradation of results

Experiments	CI(std)	MSE(std)	Pearson(std)
reported	0.89	0.21	0.85
Reproduced	0.89	0.23	0.84
CNN(100x100)	0.76	0.58	0.55
CNN(200x200)	0.88	0.25	0.83
no aug.(100x100)	0.88	0.23	0.84
Ensemble no aug.	0.88	0.23	0.83
Dropout	0.88	0.23	0.84

Table 2. Experiment (DGraphDTA) results produced on Davis test dataset

Experiment	Pre.	Recall	F1-sc.	Acc.	MCC
reported	0.89	0.92	0.90	0.88	0.76
Reproduced	0.87	0.92	0.90	0.87	0.74
100 x 100	0.82	0.90	0.86	0.82	0.63
CNN with GNN	0.85	0.95	0.90	0.87	0.73

Table 3. Experiment (DEEPScreen) results produced on Davis test dataset

through lack of robustness.

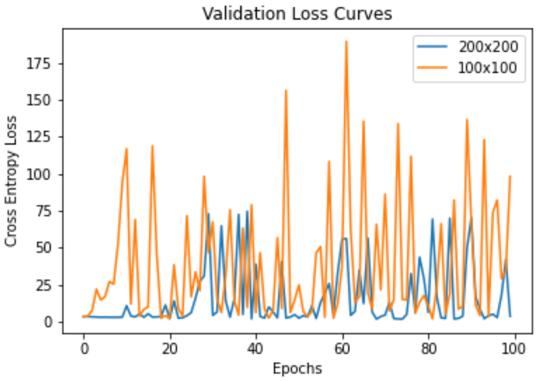

Figure 7. Loss curves for 100x100 molecule images experiment with DEEPScreen

To counter this negative effect of the lower image resolution, the DGraphDTA code was again modified, by saving all the preprocessed data to files and modifying the data loader of pytorch to read from files, we saved a lot of RAM space that was being consumed with the InMemoryDataset and thus the image resolution and batch size could be doubled. This resulted in metrics that were better than the last experiment, as can be verified in Table 2 with row CNN (200 x 200). And the validation loss was still better as seen in the Figure 8.

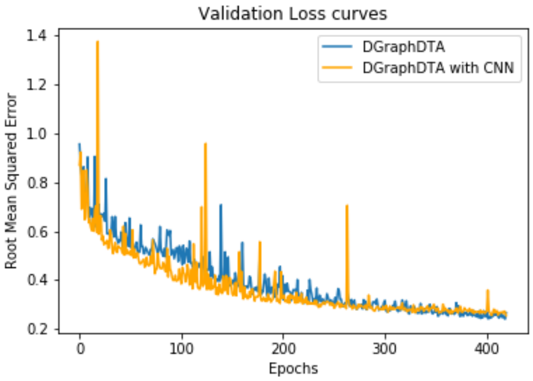

Figure 8. Loss Curve - Replacing Molecule Graphs with Images in DGraphDTA (200x200)

Along the parallel dimension, the image augmentation (random rotations, and skew) was dropped and the results were seen to improve. Refer to the Table 2 for results. The transformations applied at lower resolution could be distorting the useful information passed to the model.

4.2. Experiment 2

Since the validation loss curve with CNN based molecular features in DGraphDTA was observed to give lower loss, this urged us to cross check if the performance of DEEPScreen would also degrade if the CNN was replaced with GNN from DGraphDTA as shown in Figure 9.

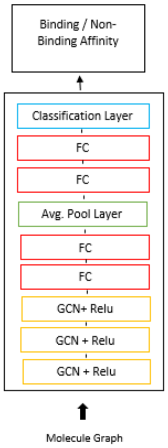

Figure 9. DEEPScreen with GNN

But the experiment gave contradictory results and shown in Figure 10.

But then, since the distribution of data for DEEPScreen was very different from the distribution of data from DGraphDTA, we concluded that the comparison is not justified. Since the DEEPScreen model was trained for

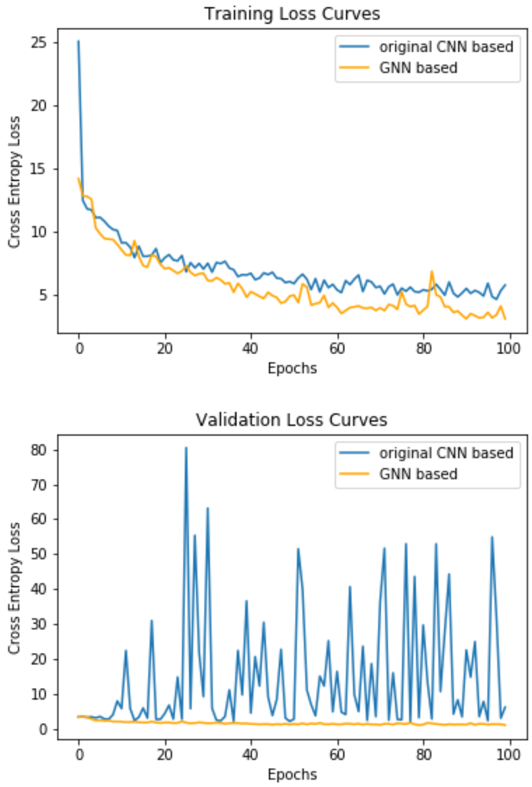

Figure 10. Comparison of loss curves for GNN experiment with DEEPScreen For ChEMBL 286 only

a single protein, the positive and negative samples of the proteins were bound to share much more similarities, then in the case of binding and non-binding drugs across different proteins.
The experimental results on binary classification metrics on test set can be seen in Table 3. GNN based features proved to be more robust in this case and resulted in a remarkable improvement in Recall, while maintaining comparable performance on other metrics.

4.3. Experiment 3

Assuming the ensemble of molecular graph and image based features might work better. We trained an ensemble model, shown in the Figure 11.

The results on the test set can be seen in the Table 2. Ensemble model gave even worse results than the model with image based features. This is a really unlikely event, since ensembles are always seen to perform better than individual models. But we believe that it might be the protein representation that is causing the bottleneck and not the molecular representation.

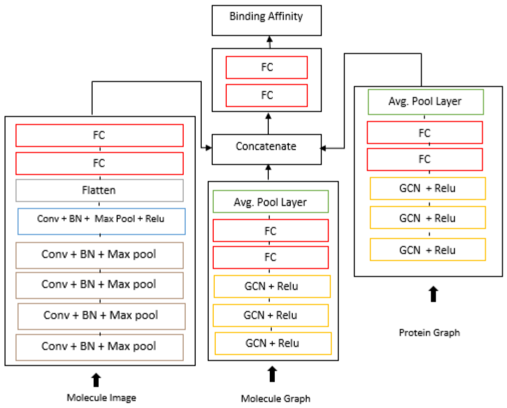

Figure 11. Ensemble on DGraphDTA

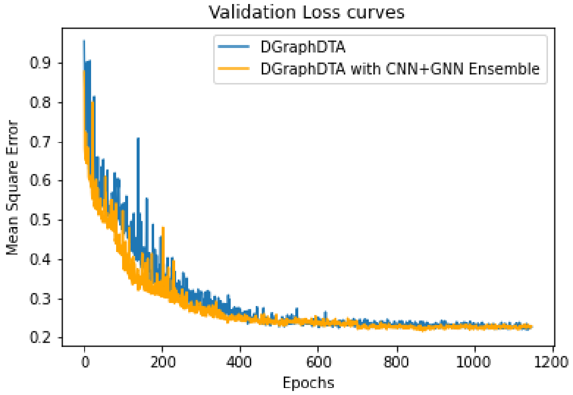

Figure 12. Loss Curve-Ensemble Molecule Graphs with Images(100x100) in DGraphDTA

4.4. Experiment 4

Additionally, one limitation we observed in the DGraphDTA paper was that it used predictions of contact maps for protein residues. The ground truth contact maps could be obtained from the 3D structure of the protein, but the 3D structure is obtained through crystallography, which is an expensive process making 3D protein structures hard to obtain. But the paper needed these contact maps to define protein graphs. In order to make the predictions from these noisy and inaccurate graphs more robust, we suggested using edge dropout[11]. The edge dropout technique randomly holds out certain edges from the graph, at the time of training, to make the predictions more robust to the noisy or absent edges. We experimented with different configurations of dropout, by changing the number of dropout layer and the dropout probability. With dropout probability of 0.5 and edge drop applied before each graph convolution layer, the validation loss curve is visualized in the Figure 4.4.
And the metrics on the test set are shown in the Dropout

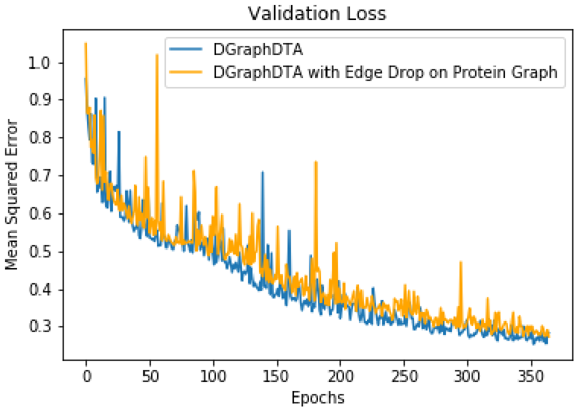

Figure 13. Edge Drop on Protein Graphs in DGraphDTA

row of the Table 2.

5. Affinity Prediction For CoVID Proteinase

Since for our problem, we need to predict affinities for drugs with target protein as COVID main protease. We needed to perform preprocessing on the sequence of proteinase, to generate a contact map for it, which could be used as an adjacency matrix for protein graph generation.

Proteinase 3CL-PRO. Amino acid sequence was taken from here generated through their deep learning algorithms. For preprocessing, 1stly inter residue protein contact map was generated using online webserver MapPred from link . They used their Deep residual neural network model which is trained on available databases and modification of DeepMSA [12] from Zhang lab from where we took covid 19 proteinase structure (link is shown above). Based on their trained model contact map was generated as shown in Figure 14.

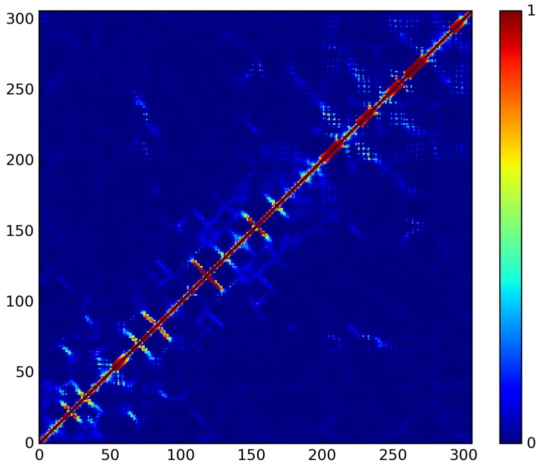

Figure 14. COVID 19 Proteinase Contact Map

Drug ID	Predicted Affinity in pKD
126565	9.907675
44259	9.612669
9809715	8.944876
11984591	8.742044
11427553	8.671042
16038120	8.623212
5329102	8.481388
16722836	8.377007
447077	8.320545

Table 4. Covid19 Proteinase 3CL-PRO protein Binding Affinities[4]with Davis dataset drugs predicted using DGraphDTA's supplied model.

Drug ID	Predicted Affinity in pKD
44259	6.672781
126565	6.1242332
11984591	6.019435
16722836	5.5392294
16038120	5.535989
11427553	5.5091653
5287969	5.4815736
176155	5.403635
11667893	5.3888054

Table 5. Covid19 Proteinase 3CL-PRO protein Binding Affinities with Davis dataset drugs predicted using DGraphDTA's supplied model with edge dropout.

The Mappred generated the .a3m file containing inter residue sequence alignments in pairs and multiple contact maps. From script code of DGraphDTA, first two steps (generating .seq file and using hhblits to generate .a3m file) were already performed so they were skipped. From hhfilter to reformat and pconsc4 were performed to generate contact map in numpy array where pconsc4 generated contact maps based on multiple contact maps from given file. The output of pconsc4 was used to as an input to generate PSSM to generate scoring matrix. The top 10 most affine drug molecules from Davis dataset for COVID Proteinase, predicted using our experimental models are shown in Tables 4,5,6,7. These affinity values can only be validated through laboratory experimentation of computer based docking simulations if 3D structure of the binding pair is available. So validating the results is out of our scope.

[4] Affinity values are in log scale $pK_D(nM) = -\ln(\frac{K_D}{10^9})$. Where K_D is in 10,000 nM range. The 10,000 nM results in $pK_D = 5$nM means no binding. Values from greater than 5 nM are required for drug binding with Covid19 Proteinase.

Drug ID	Predicted Affinity in pKD
44259	9.081171
16038120	7.864847
126565	7.2855363
9809715	7.118546
5328940	6.716422
5329102	6.7150183
25127112	6.5924306
11427553	6.4434547
11984591	6.354123

Table 6. Covid19 Proteinase 3CL-PRO protein Binding Affinities with Davis dataset drugs predicted using the Drugs Images with Protein Graph.

Drug ID	Predicted Affinity in pKD
44259	7.8419065
126565	6.424454
9809715	6.2218676
25127112	5.7972865
11984591	5.781021
51004351	5.7474995
11409972	5.6736794
5329102	5.639513
16038120	5.5416374

Table 7. Covid19 Proteinase 3CL-PRO protein Binding Affinities with Davis dataset drugs Ensemble model.

6. Conclusion and Future Work

From the results that we have, it appears that graph-based features are more robust molecular representations and image-based features result in faster error decay in the early phases of the training. Edge dropout although resulting in comparable performance is not showing any improvement on the protein representation learning. This contradicts our hypothesis that protein graphs are noisy, but can be justified by considering that proteins are huge, with small binding sites where a drug molecule can dock. Identifying that one site in the huge protein might work better with attention-based approaches to narrow down the unnecessary information and improve predictions, which can be explored in the future. Although our ensemble model also did not work quite well that might because of the individual representations were not complimentary but contradictory or might be it was in actual protein branch that was really the bottleneck. From the above discussion, the predicted binding affinities of Covid19 Proteinase 3CL-PRO,protein with drugs produced using the baseline DGraphDTA model is more reliable than other models.

References

[1] Yash Gupta, Dawid Maciorowski, Raman Mathur, Catherine M Pearce, David J Ilc, Hamza Husein, Ajay Bharti, Daniel Becker, Rathi Brijesh, Steven B Bradfute, et al. Revealing sars-cov-2 functional druggability through multi-target cadd screening of repurposable drugs. 2020.

[2] Bo Ram Beck, Bonggun Shin, Yoonjung Choi, Sungsoo Park, and Keunsoo Kang. Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. *Computational and structural biotechnology journal*, 2020.

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

[4] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. Widedta: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[6] 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.

[7] 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.

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[9] Ahmet Sureyya Rifaioglu, Esra Nalbat, Volkan Atalay, Maria Jesus Martin, Rengul Cetin-Atalay, and Tunca Doğan. Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-d structural compound representations. *Chemical Science*, 11(9):2531–2557, 2020.

[10] Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, 10(35):20701–20712, 2020.

[11] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2019.

[12] Chengxin Zhang, Wei Zheng, SM Mortuza, Yang Li, and Yang Zhang. Deepmsa: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, 36(7):2105–2112, 2020.