BindingSite-Augmented DTA - Supplementary information

March 2, 2023

1 Computational method supplemental information content

1.1 AttentionSiteDTI

Our proposed framework is built based on our previously developed model, called AttentionSiteDTI, which was initially developed for the classification task of Drug-Target Interaction (DIT) prediction. AttentionSiteDTI is inspired by models developed for sentence classification in the field of Natural Language Processing (NLP), where the drug-target complex is treated as a natural language sentence with structural and relational meaning between its biochemical entities, a.k.a. protein pockets and drug molecule. In this regard, each protein pocket or drug is analogous to a word, and each drug-target pair is analogous to a sentence. AttentionSiteDTI utilized an end-to-end Graph Convolutional Neural Network (GCNN)-based model to simultaneously learn context-sensitive graph embeddings of protein pockets and small molecules as well as a DTI prediction model capturing contextual and relational information contained in the sentence.

Unlike most of the graph-based models that use amino acid sequence representations for proteins, our model uses the 3D representation of pocket-like regions of the proteins as the input for target proteins. Considering the fact that the intermolecular interactions between protein and many ligands occur at different binding pockets (of the protein's surface) rather than the whole protein, in our model, we represent protein pockets as graphs where the key protein residues correspond to the nodes that are connected based on residue proximity. Furthermore, the features associated with each node are encoded as a vector describing the local amino acid environment.

AttentionSiteDTI is highly generalizable due to the use of protein pockets encoded as graphs to represent the target protein. This allows the model to focus on learning generic topological features from protein pockets, which can be generalized to new proteins that are not similar to the ones in the training data. AttentionSiteDTI is also highly explainable due to its self-attention

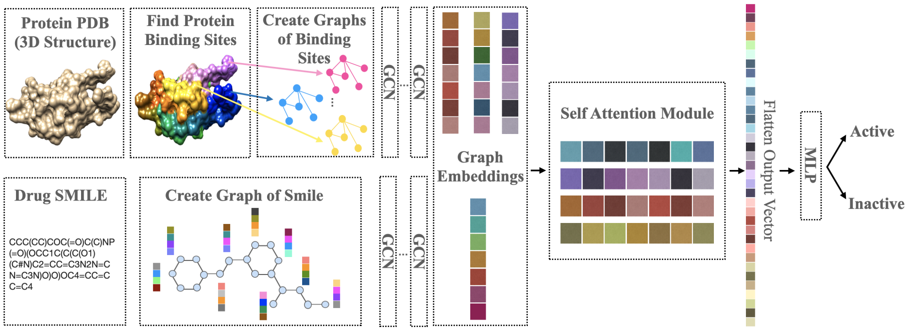
Protein PDB (3D Structure)
Find Protein Binding Sites
Create Graphs of Binding Sites
GCN
GCN
Graph Embeddings
Self Attention Module
Flatten Output Vector
MLP
Active
Inactive
Drug SMILE
CCC(CC)COC(=O)C(C)NP(=O)(OCC1C(C(C(O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O)OC4=CC=CC=C4
Create Graph of Smile
GCN
GCN

Figure 1: Architecture of AttentionSiteDTI: Following the extraction of protein's binding sites, the graphs of protein pockets and ligands are constructed and fed into a graph convolutional neural network to learn corresponding graph embeddings. The concatenated representations are then fed into a binary classifier for predicting drug-target interactions. The self-attention mechanism in the network uses concatenated embeddings to compute the attention output, which enables interpretability by making the model learn the most probable binding sites of the protein in interaction with the ligand in a given drug-target pair.

mechanism, which is used to capture any relationship between binding sites of a given protein and the drug in a sequence (i.e., sentence) and thus provide a better understanding of their binding relationships. To be self-contained, here we provide a brief description of AttentionSiteDTI, and we refer the readers to the original paper for more details (1).

AttentionSiteDTI is an end-to-end graph-based deep learning model which was originally designed to address the problem of drug-target interaction prediction. It consists of four modules: (1) data preparation to find the binding sites of the proteins using the algorithm proposed in (2), (2) graph embedding learning module to learn the embeddings from constructed graphs of protein pockets and ligands as the inputs to the graph convolutional neural network, (3) prediction module to predict drug-target interactions using learned drug-target complex representations, and (4) interpretation module, quipped with a self-attention mechanism, to detect the most probable binding sites of the protein when interacting with the ligand in a given drug-target pair. The output of this module is indeed the main component that we use in this study to make the DTA prediction models focus on the most important parts of the protein when learning interactions between drug-target pairs.

1.2 Experiments Setup and Hyperparameters

The hyper-parameter settings for the eight models were found by grid search and utilized and kept the same as reported in their original studies. For that, five training sets were used in 5-fold cross-validation to train the model, and the final CI scores were reported as the average of these five results. These hyper-parameters are summarized in 1 and 2.

Table 1: Summary of Parameter Setting for String Representation-Based Models

Hyperparameters	DeepDTA	WideDTA	AttentionDTA
Number of filters	32,64,96	32,64	32,64,96
Filter length (compounds)	[4,6,8]	-	[4, 6, 8]
Filter length (proteins)	[4,8,12]	-	[4, 6, 12]
Epoch	100	100	350
Hidden neurons	1024,1024,512	1024,1024,512	-
Batch size	256	256	64(Davis)- 256(KIBA)
Dropout	0.1	0.3	-
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.0001

Table 2: Summary of Parameter Setting for Graph Representation-Based Models

Hyperparameters	GraphDTA-GAT	GraphDTA-GIN	DGraphDTA	DeepGS
FC layers after GNN	1	1	2	1
FC layers after concatination	4	8	2	4
GIN layers	0	5	0	0
GAT layers	2	0	0	0
Epoch	100	100	2000	100
Batch size	512	512	512	1
Dropout	0.2	0.2	0.2	0.2
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.0005	0.0005	0.001	0.0001

Since the model should always be trained under distinct prediction scenarios and regarding the fact that the few labeled data points from lab experiments are inadequate for training the models from scratch, we train AttentionSitDTI and DeepDTA on the BindingDB dataset, which includes a wide variety of drug-like compounds and target proteins. This enables both models to learn generic rules governing the ligand-protein interactions from the BindingDB dataset, containing 2,303,972 binding data for 8,561 protein targets and 995,797 small molecules, as of July 2021. Once the *general* complex interaction patterns are extracted, the models then utilize this pre-learned knowledge and transfer it to the *specific* task of binding affinity prediction between 13 drug-like compounds and the Spike-ACE2 complex.

1.3 Tabular Results on Kiba and Davis Datasets

Tabular results for KIBA and Davis datasets are provided below in Tables 3 and 4. Also, we performed the one-tailed t-test at the significance level of $\alpha = 0.05$ to test whether improved performance of the models wiht AttentionSiteDTI is statistically significantly compared to those of in the plain version of the models.

Table 3: Comparison with state-of-the-arts on Kiba dataset

	Without AttentionSiteDTI				**With AttentionSiteDTI**			
Methods	**CI(std)**	**MSE**	r_m^2 **(std)**	**AUPR (std)**	**CI(std)**	**MSE**	r_m^2 **(std)**	**AUPR (std)**
String Representation-Based Approaches								
DeepDTA	0.863 (0.001)	0.194	0.673 (0.009)	0.788 (0.004)	0.881 (0.003)	0.151	0.778 (0.004)	0.830 (0.002)
WideDTA	0.875 (0.001)	0.179	0.675 (0.009)	0.788 (0.002)	0.880 (0.004)	0.170	0.764 (0.001)	0.821 (0.003)
AttentionDTA	0.882 (0.006)	0.162	0.735 (0.005)	0.829 (0.002)	0.874 (0.009)	0.152	0.777 (0.005)	0.830 (0.002)
Graph Representation-Based Approaches								
GraphDTA-GAT	0.849 (0.003)	0.189	0.726 (0.003)	0.806 (0.002)	0.886 (0.001)	0.157	0.775 (0.004)	0.826 (0.001)
GraphDTA-GIN	0.859 (0.001)	0.168	0.724 (0.007)	0.803 (0.004)	0.887 (0.002)	0.147	0.788 (0.006)	0.831 (0.001)
DGraphDTA	0.904 (0.002)	0.126	0.786 (0.006)	0.827 (0.003)	0.912 (0.001)	0.119	0.799 (0.007)	0.840 (0.002)
DeepGS	0.841 (0.005)	0.367	0.659 (0.002)	0.740 (0.006)	0.889 (0.003)	0.201	0.771 (0.005)	0.792 (0.002)
DeepH-DTA	0.927 (0.003)	0.111	0.799 (0.004)	0.861 (0.002)	-	-	-	-

Table 4: Comparison with state-of-the-arts on Davis dataset

	Without AttentionSiteDTI				**With AttentionSiteDTI**			
Methods	**CI(std)**	**MSE**	r_m^2 **(std)**	**AUPC**	**CI(std)**	**MSE**	r_m^2 **(std)**	**AUPC**
String Representation-Based Approaches								
DeepDTA	0.878 (0.008)	0.261	0.630 (0.002)	0.714 (0.004)	0.887 (0.003)	0.209	0.734 (0.002)	0.820 (0.002)
WideDTA	0.886 (0.008)	0.262	0.633 (0.007)	0.711 (0.003)	0.894 (0.004)	0.241	0.709 (0.004)	0.772 (0.005)
AttentionDTA	0.887 (0.007)	0.245	0.657 (0.007)	0.746 (0.003)	0.902 (0.009)	0.192	0.755 (0.004)	0.830 (0.002)
Graph Representation-Based Approaches								
GraphDTA-GAT	0.892 (0.001)	0.232	0.662 (0.005)	0.728 (0.009)	0.898 (0.003)	0.206	0.733 (0.001)	0.812 (0.007)
GraphDTA-GIN	0.893 (0.002)	0.229	0.649 (0.001)	0.720 (0.004)	0.901 (0.001)	0.198	0.746 (0.004)	0.807 (0.005)
DGraphDTA	0.894 (0.003)	0.216	0.698 (0.002)	0.700 (0.004)	0.894 (0.002)	0.207	0.736 (0.001)	0.816 (0.001)
DeepGS	0.746 (0.009)	0.598	0.240 (0.012)	0.547(0.008)	0.834 (0.002)	0.307	0.678 (0.005)	0.745 (0.007)
DeepH-DTA	0.924 (0.001)	0.195	0.725 (0.009)	0.801 (0.010)	-	-	-	-

As the p-values, reported in Tables 5 and 6 show, except for the one case of AttentionDTA in KIBA dataset, in all other cases the improved results are statistically significant.

Table 5: t-test P-Values on Davis dataset

Methods	**CI**	r_m^2	**AUPC**
String Representation-Based Approaches			
DeepDTA	0.023145 (Yes)	<.00001 (Yes)	<.00001 (Yes)
WideDTA	.040258 (Yes)	<.00001 (Yes)	<.00001 (Yes)
AttentionDTA	0.009331 (Yes)	<.00001 (Yes)	<.00001 (Yes)
Graph Representation-Based Approaches			
GraphDTA-GAT	0.001414 (Yes)	<.00001 (Yes)	<.00001 (Yes)
GraphDTA-GIN	0.000022 (Yes)	<.00001 (Yes)	<.00001 (Yes)
DGraphDTA	0.5 (No)	<.00001 (Yes)	<.00001 (Yes)
DeepGS	<.00001 (Yes)	<.00001 (Yes)	<.00001 (Yes)

Table 6: t-test P-Values on KIBA dataset

Methods	**CI**	r_m^2	**AUPC**
String Representation-Based Approaches			
DeepDTA	<.00001 (Yes)	<.00001 (Yes)	<.00001 (Yes)
WideDTA	.013296 (Yes)	<.00001 (Yes)	<.00001 (Yes)
AttentionDTA	0.068466 (No)	<.00001 (Yes)	0.226015 (No)
Graph Representation-Based Approaches			
GraphDTA-GAT	<.00001 (Yes)	<.00001 (Yes)	<.00001 (Yes)
GraphDTA-GIN	<.00001 (Yes)	<.00001 (Yes)	<.00001 (Yes)
DGraphDTA	0.000022 (Yes)	<.00001 (Yes)	.000021 (Yes)
DeepGS	<.00001 (Yes)	<.00001 (Yes)	<.00001 (Yes)

1.4 DREAM Challenge Data as Additional Validation Data

In addition to the KIBA and Davis datasets, we also tested our framework on the recent IDG-DREAM challenge benchmark dataset (3) for drug-target interaction prediction. We utilized this dataset for further validation of our results, since it contains activity data on understudied human kinomes, so-called dark kinases. These quantitative bioactivities enable performance evaluation of prediction models on both on- and off-target kinase activities for rather challenging compounds and target spaces of multi-targeting kinase inhibitors. As suggested by this challenge, a subset of DrugTargetCommons (DTC) (https://drugtargetcommons.fimm.fi) (4) is used as the training dataset. DTC is a publicly available web-platform tool that provides standardized resources to retrieve crowd-sourced compound-target bioactivity data. The original dataset was filtered to the activity type related to equilibrium dissociation constant (Kd or pKd), where 49791 interactions were obtained. For the test dataset, two rounds were suggested by the IDG Kinase-DRGC program. Round 1 data with 430 Kd interactions consist of 70 compounds and 199 kinases, and round 2 data with 394 Kd interactions consist of 25 compounds and 207 kinases. All of the major kinase families and groups were covered by the round 1 and round 2 kinase targets, of which 111 of them overlapped. Together these 824 Kd values of the compound-kinase pairs were not available publicly and were unpublished at the time of the challenge. Table 7 summarizes the description of the training and test datasets suggested by the IDG Kinase-DRGC program. (3)

Table 7: IDG-DREAM Drug-Kinase Challenge dataset

	Compounds	Proteins	Interactions
DTC (Train set)	7564	827	49791
Round 1 and 2 (Test set)	95	295	824

We selected DeepDTA to predict compound-protein binding affinities on the DREAM challenge dataset. Beside its wide adoption in the literature, our rationale for choosing DeepDTA for this set of experiments is that it is one of the competing teams (Team: Boun - DeepDTA) in the IDG-DREAM Challenge. Also, DeepDTA is the first DL approach developed to predict drug-target binding affinities, and it is among state-of-the-art methods that have shown relatively good performance compared to many DL-based models with higher complexity in architecture and computation time. As reported in Table 8, our proposed framework improves DeepDTA's prediction on round 1 and round 2 (combined) datasets in all metrics.

Table 8: Comparison of DeepDTA with and without AttentionSiteDTI on Dream challenge testing dataset

	Without AttentionSiteDTI				**With AttentionSiteDTI**			
Experimental Setup	**CI**	**MSE**	r_m^2	**AUROC**	**CI**	**MSE**	r_m^2	**AUROC**
Round 1 and 2	0.469	6.278	-3.651	0.548	**0.511**	**3.274**	**-1.426**	**0.556**

Furthermore, we performed another set of experiments on the DTC dataset in order to evaluate the potential of our framework under a more challenging and realistic setting. We show separate results under four different evaluation settings, whether training and test sets share common drugs and targets, only drugs or targets or neither. In this experimental setting, we take into consideration the differences between the following four scenarios under which the model can learn to predict the label of a query drug–target pair (x_d, x_t). Also, for these experiments, we used a simple train-test split for model evaluation.

- S1. Bioactivity Imputation Scenario: Both drug x_d and protein x_t are present in the training set.
- S2. New Drug Scenario: The protein x_t in present in the training set, but the drug x_d is unseen in the training phase.
- S3. New Target Scenario: The drug x_d is encountered in the training phase, whereas the target protein x_t is not.
- S4. New Drug-Target Pair Scenario: Neither the drug x_d nor the target protein x_t is encountered in the training phase.

Table 9: Comparison of DeepDTA with and without AttentionSiteDTI on Dream challenge training dataset

	Without AttentionSiteDTI				**With AttentionSiteDTI**			
Experimental Setup	**CI**	**MSE**	r_m^2	**AUROC**	**CI**	**MSE**	r_m^2	**AUROC**
S1	0.788	0.879	0.477	0.719	**0.874**	**0.431**	**0.743**	**0.850**
S2	**0.696**	2.174	-0.355	**0.692**	0.687	**1.559**	**0.028**	0.679
S3	0.693	1.252	0.241	0.649	**0.725**	**1.170**	**0.290**	**0.682**
S4	0.663	**1.596**	**0.053**	0.662	**0.684**	1.671	0.009	**0.678**

As the results show, the integration of this model with our AttentionSiteDTI leads to improved performance in 12 out of 16 cases, as reported Table 9. These improvements are consistent with our findings on the other two datasets (KIBA and Davis), indicating that our proposed framework is, indeed, effective in boosting the performance of DTA prediction models.

1.5 Separated results for experimentally-measured and AF-predicted structures of proteins on Davis dataset

We used AlphaFold to predict 3D structure for proteins not having experimentally measured structures in PDB. We show the prediction results separately for experimentally measured and AF-predicted structures of proteins on Davis dataset (with83 out of 442 AF-predicted proteins) to see potential differences in the target activity predictions. We provide the results as you can find below in Table 10.

Table 10: Comparison of target activity predictions on experimentally-measured vs AlphaFold-Predicted proteins in Davis dataset

	Experimentally measured				**AF-Predicted**			
Models	**CI**	**MSE**	r_m^2	**AUROC**	**CI**	**MSE**	r_m^2	**AUROC**
DeepDTA	0.889	0.212	0.735	0.816	0.885	0.202	0.731	0.842
WideDTA	0.891	0.247	0.703	0.774	0.901	0.231	0.712	0.772
AttentionDTA	0.914	0.181	0.753	0.821	0.9	0.195	0.759	0.834

2 In-Lab testing supplemental information content

2.1 Materials and Methods

ML has been recently used to enhance the sampling of MD simulation trajectories to capture millisecond scale in-lab processes using nanosecond scale simulations. It has also been used to remove noise in simulation data and make it more accessible for interpretation (5; 6). Free energies calculated from MD simulations have been used to train DL models to accelerate the prediction of free energies based on the structural information of small molecules (7). Joshi et al. (8) have used DL based prediction framework to screen molecules from the Selleck database and perform molecular docking to further optimize the search results for final use in MD simulations to find the molecules having the most potential to bind with the main protease of SARS-CoV-2. In the present study, a comparison has been made between MD simulations of certain selected molecules with the RBD of SARS-CoV-2 from the pool of molecules checked using DL-based frameworks and in-lab experiments. The purpose is to check the similarity of DL predictions with that of MD simulations and use a two-prong approach to guide in-lab experiments. The details are presented as supplementary information.

For the third approach mentioned in the main article, we have initially performed molecular docking using AutoDock Vina and AutoDock tools, and the highest scored ligand-protein complex conformation structures are chosen for the further all-atom Steered Molecular Dynamics (SMD) simulations. Here, only 5 of the previously mentioned 13 molecules have been simulated using SMD simulations to determine the peak unbinding force and center of mass separation of the corresponding ligands with time (Supplementary material Figures 3 and 4). Our molecular docking results show that three ligands bind to one single binding pocket of the protein, which is in line with previously mentioned theories of having specified ligand binding locations in spike protein (Supplementary material section 1.5). Next, our SMD simulation results showed that N-Acetyl neuraminic acid (sialic acid) requires the highest amount of steering force before complete dissociation takes place (Supplementary material Figure 3 and Table 2). Based on this, the five molecules analyzed via SMD simulations are ranked in such a way that rank one is given to the molecule that shows the highest binding strength and force requirement, while the lowest ranked molecule has the least tendency to bind with the protein in a favorable environment. The comparison of this ranking analyzed via different in silico approaches, MD simulations, and in-lab experiments based on binding affinity is given in Supplementary information Table 3. From Table 7, it is clearly seen that both our SMD simulation and DeepDTA-AttentionSite model predict N-acetyl neuraminic acid as the highest ranked molecule in terms of binding affinity with spike protein, and interestingly, DeepDTA, our MD simulation, and the in-lab experiments all predict N-glycolyl neuraminic acid as the second based inhibitor. Cytidine-5' has been ranked 5th both by DeepDTA and our SMD simulation as the molecule least prone to bind, but the N-acetyllactosamine is ranked last in the in-lab experiment. Therefore, even though there are similarities in results based on the DL model and our SMD simulations, the in-lab experiment results seem to vary for the molecules having the least binding affinity. So, the in-lab experiments might be optimized based on the DL and MD results to accelerate the selection of potential inhibitors.

2.1.1 Materials

The ACE-2 SARS-CoV2 inhibition assay used in the presented work was purchased as a kit from BPS Bioscience (ACE-2: SARS-CoV-2 Spike Inhibitor Screening Assay Kit, Catalog no: 79936; Spike-S1 neutralizing antibody, SARS-COV-2, Clone 414-1, Catalog no: 100793). Test molecules (3α,6α-mannopentaose, 2-keto-3-deoxyoctonate ammonium salt, N-glycolylneuraminic acid, Benzanol Brilliant scarlet 3B, methylene blue, cytidine-5-monophospho-N-acetylneuraminic acid sodium salt, evans blue, darunavir, congo red, N-acetyl-neuraminic acid, direct violet 1, N-acetyllactosamine, and chlorazol black) were purchased from Fisher Scientific and Millipore Sigma and used without further purification.

2.1.2 ACE-2/SARS-CoV-2 spike Inhibitor screening ELISA assay

Testing of candidate compounds (noted above) was performed via a commercial assay, with measurements of ACE-2:SARS-CoV2 spike protein complex inhibition activity performed per manufacturer specifications/procedure. The ELISA-based assay was performed for each candidate molecule at concentrations from 0.01 nM to 40 nM. Based on results from the assay, IC50 values were determined by extrapolating compound concentration at which 50% of ACE-2/Spike protein are left unbound over the test concentration range. These values were then used to compare the efficacy of each molecule. Note that we used converted PIC50 values to compare the experimentally measured binding affinities against computational results, which are reported in the paper. Here, we report the assay procedure, in brief.

Initially, spike protein concentration was optimized for luminescent emission (in arbitrary units) following incubation with secondary antibody. 20 nM was determined to be the optimal concentration. ACE-2 was then suspended to 1 μg/mL in 10 mM phosphate buffered saline and added to a nickel-coated 96-wellplate. Incubation was maintained for 1 h at room temperature under constant, low speed shaking. ACE-2 was then washed three times with immunobuffer, incubated in blocking buffer, and washed an additional three times with immunobuffer. 10 μL of an inhibitor solution (1% dimethyl sulfoxide in PBS) was then added to each well, and incubated for an additional hour at room temperature under shaking, to inhibit further binding. SARS-CoV2 spike protein was then added at approximately 20 nM (1 ng/μL) in 1% DMSO in PBS were added to each well, excluding wells designated as blanks. 1% DMSO in PBS was added to positive control and blank wells. From here, the plates were incubated for an additional hour at room temperature under constant, low speed shaking. Then, the plates were washed with immunobuffer and incubated for an additional 10 minutes in blocking buffer. Mouse anti-HRP (horseradish peroxidase) was added and the plates shaken for an hour. Lastly 100 μL of HRP substrate was added to all wells. Chemiluminescence was finally measured using a 96-wellplate reader.

2.1.3 Candidate compound selection

In validating our produced drug-target interaction model, we looked to identify its ability to identify positive hits while predicting weak or no interaction for low affinity compounds. Further, we looked to identify the

precision of the model through testing compounds which vary slightly, in a defined character, from well-performing compounds. In identifying candidates, compounds are often chosen which mimic the chemical structure of native bio-molecules. In pharmaceutical applications, the candidates often provide either agonist or antagonist binding. Therefore, compounds are chosen which mimic, and exacerbate or otherwise optimize, natural binding behavior towards biomolecule substrates. In the present work, the natural binding pair of SARS-CoV2 S-protein (spike protein) and human angiotensin-converting enzyme 2 (ACE2) were selected as a model binding system. It has been determined and presented in literature that this binding is mediated by interaction between S-protein and the N-acetylneuraminic acid species presented, among other species of a dense glycan shield, on the surface of specific cell types in the host. In consideration of this, test molecules were chosen which possessed small (measurable/identifiable) variations to the chemical structure of N-acetylneuraminic acid (e.g., inclusion of functional groups with greater polarity, increase in the length of the carbon backbone, increase in degree of substitution to exaggerate steric interactions, etc.). Darunivir is structurally distinct from N-acetylneuraminic acid, as well as the structurally similar family of biomolecules known as sialic acids, though has been suggested as an antiviral agent towards infection by SARS-CoV2. Similarly, several organic species, used traditionally as dyes, have been suggested. We have chosen to include these molecules in our validation set in order to analyze our models ability to extract and prioritize physicochemical features towards the identification of new drugs hits, candidate molecules towards inhibition of SARS-CoV2:ACE2 binding.

Table 11: Results of Inhibition Assay experiments: For 3α,6α-Mannopentaose no clear interaction was observed in our assay; For this compound we set the PIC50 value to 4.9 nM, which corresponds to the high concentration of 12500 nM.

Rank	Candidate Compound	IC50(nM)	pIC50 (nM)	pIC50 (μM)
1	2-Keto-3-deoxyoctonate ammonium	0.95	9.02	6.02
2	N-glycolylneuraminic acid	1.7	8.77	5.77
3	Benzanol Brilliant Scarlet 3 B	2.25	8.65	5.65
4	Methylene Blue	2.5	8.60	5.60
5	Cytidine-5-monophospho-N-acetylneuraminic acid sodium salt	3.2	8.49	5.49
6	Evans Blue	4.5	8.35	5.35
7	Darunavir	5.5	8.26	5.26
8	Congo Red	14.8	7.83	4.83
9	N-Acetyl-neuraminic acid	19	7.72	4.72
10	Direct Violet 1	19.5	7.71	4.71
11	N-Acetyllactosamine	23.5	7.63	4.63
12	Chlorazol Black	38	7.42	4.42
13	3α,6α-Mannopentaose	12500	4.90	1.90

graphicx

3 Molecular dynamics simulations results in support of the machine learning predictions and the in-lab validations

3.1 Synopsis

Docking simulations are performed to generate ligand-RBD (Receptor Binding Domain) complexes for five of the selected molecules. Then MD simulations are carried out to find out the maximum force obtained in forced dissociation of the ligand from the RBD and this is benchmarked to the ACE2-SP complex. It is observed that top performing DL predictions compare favorably with MD simulation results and predict the sialic acid to be a strong candidate for a potent inhibitor. However, this is not corroborated by in-lab validations. Interestingly the second rank for prediction from MD simulation matches with in-lab results whereas the DL models differ widely. This indicates that a more synergistic prediction framework involving DL and MD may help to explore potential inhibitors with greater fidelity to accelerate in-lab experimentation and ensuing selection.

3.2 Introduction

Molecular docking and molecular dynamics (MD) simulations (9) can complement the machine learning based predictions and provide theoretical confirmation for experimental binding curves as well as substantiate the predictions of DeepPurpose and GCNN. A statistically sound estimation of the correct binding modes of ligands in protein-ligand complexes, as obtained from all atom molecular simulations, is an important step in the drug discovery process and also helps in illuminating potential toxicity mechanisms (10). Typically, selected ligands are docked to target proteins using molecular docking techniques to generate stable complexes, following which the binding energy can be calculated using dynamics simulation methods like umbrella sampling (11). When crystallographic structures are readily available, docking may not be required as those structures can be used as input for the dynamics simulations.

3.3 Methods

Before screening new compounds as possible ligands for blocking the SP, it is imperative to create a benchmark by calculating the free energy of binding between SP and ACE2. To achieve this the crystallographic structure of SP-ACE2 complex (PDB ID : 6M0J) is equilibrated and properly samples in a rectangular box containing water molecules in an ionic concentration of 0.1 M NaCl, using the GROMACS 2020 software and the AMBER99SB force field (12). The stable configuration is then used to prepare trajectories for the umbrella sampling simulations, wherein the ACE2 is held restrained, and the SP is steered away from the binding pocket, so that all bonds are broken, and molecular interactions are entirely disrupted. The resultant center of mass (COM) separations between SP and ACE2 is 5 nm. The trajectory is sliced into 49 windows of 0.1 nm spacing and each window is sampled for 10 ns. The resultant simulations are analyzed using the WHAM algorithm (13) to create the potential of mean force (PMF) profile for the unbinding of SP from ACE2, which gives a binding energy of 16.03 ± 0.91 kcal/mol. The computed value is similar to that reported by Lee in March 2021, using the OPLS force field (14). This is equivalent to a KD 1.9 ± 0.2 pM, indicating a very strong binding of SP with that of ACE2. Based on this benchmark, the purported sialic acid based blocking agents can be docked (in case the crystallographic structures are not available) and simulated to calculate the binding energy with respect to SP.

To perform molecular docking and generate the complexes with RBD of SP, RBD structure is obtained from RCSB PDB ID - 7JVB. All corresponding ligand structures are imported from PUBCHEM database. Molecular docking is performed using AutoDock Vina and AutoDock Tools and highest score docked conformation is chosen. All atom MD simulations similar to that described above are performed for the ligand-RBD complexes, following which SMD simulations are conducted to calculate the peak unbinding force and COM separation curves. A description of the molecules chosen for the docking simulation is provided in 2. Resulting configuration of selected docking simulations are shown in 3.

3.4 Results

In the SMD simulations, the ligand is pulled with a harmonic spring attached to its COM, and over time it breaks away from the binding pocket of the complex with RBD. The resultant reaction force that is experienced by the harmonic spring, is dependent on the strength of binding between the ligand and the

Molecules of interest:

1. N-Acetyl Neuraminic Acid ($C_{11}H_{19}NO_9$)
2. N-Acetyllactosamine ($C_{14}H_{25}NO_{11}$)
3. Cytidine-5′-monophospho N-Acetyl neuraminic acid sodium salt ($C_{20}H_{30}N_4NaO_{16}P$)
4. 2-keto 3-deoxyoctonate Ammonium Salt ($C_8H_{17}NO_8$)
5. N Glycolyl Neuraminic Acid ($C_{11}H_{19}NO_{10}$)

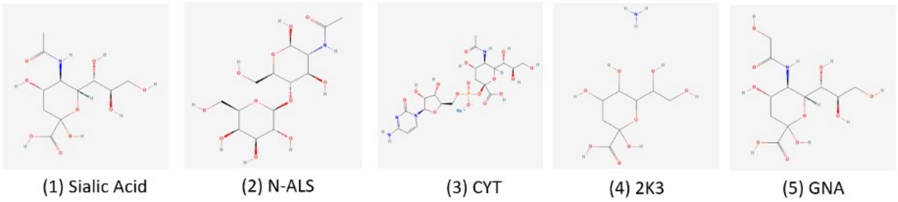

Figure 2: The molecules which were selected as ligands for the docking simulation with RBD are presented. The acronyms for the individual molecules are provided as they are used in following figure legends.

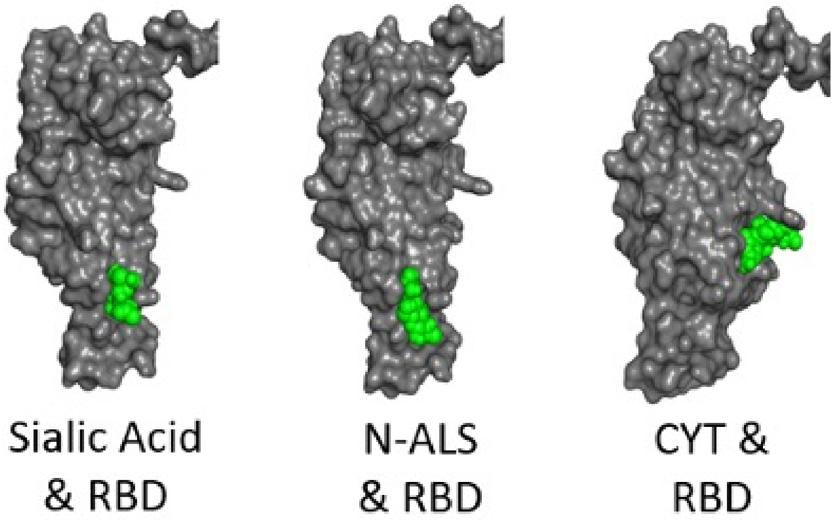

Figure 3: Configurational view of docked complexes of RBD and selected molecules are presented. The ligands are green and RBD is grey.

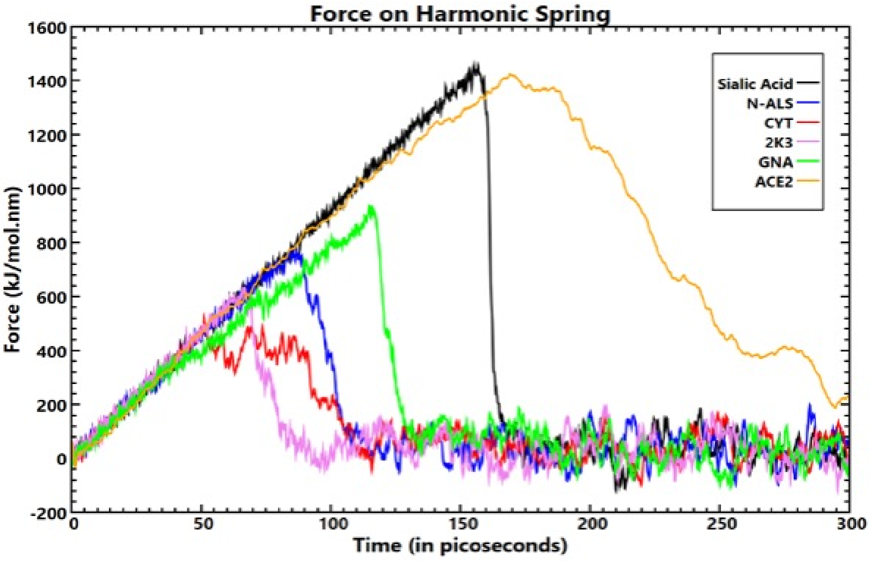

Figure 4: Force curves from SMD simulations for the 5 selected molecules and ACE2

RBD. It is trivial to observe that higher forces indicate stronger binding. The force vs simulation time curves for the 5 selected ligands as well as ACE2 is provided in 4. It is observed that N-acetyl neuraminic acid (Sialic Acid) shows highest binding affinity while Cytidine-5'-monophospho (CYT) ligand shows the lowest among them. Maximum force required to unbind Sialic Acid (1456 kJ.mol-1nm-1) is almost 3 times as high as that required to unbind CYT (518 kJ.mol-1nm-1). In comparison, the force required to unbind ACE2 is about 1425 kJ.mol-1nm-1, suggesting that Sialic Acid probably has similar binding strength. COM separation curves provided in 4 also show that Sialic Acid and ACE2 take almost same simulation time range of 170-190 ps before the ligands start breaking away from the RBD. In comparison, CYT and 2K3 break away within the first 100 ps of simulation time. This is reflected by the much lower maximum force observed in SMD simulations for those molecules.

3.5 Discussion

The primary goal of the deep learning enabled prediction platform revolves around finding out the specified and repeatable binding sites of the viral protein which are most amenable to bind with different ligand molecules and quantify the binding affinity of the corresponding cases. Likewise, from the molecular docking studies it is observed that the best docked poses are around a specified protein binding pocket of the RBD, for the 3 out of 5 simulated cases discussed here. Previous studies (15) show how different sialic acids bind with the SARS-

Table 12: Molecular docking scores in terms of binding affinity with spike protein RBD structure

Compound	Binding affinity with RBD (kcal/mol)
2-keto-3-deoxyoctonate ammonium salt (2K3)	-5.5
N-Glycolylneuraminic acid (GNA)	-5.5
Cytidine-5-monophospho-N-acetylneuraminic acid (CYT)	-6.2
N-acetyl-neuraminic acid (Sialic acid)	-5.8
N-Aceytllactosamine (N-ALS)	-6.3

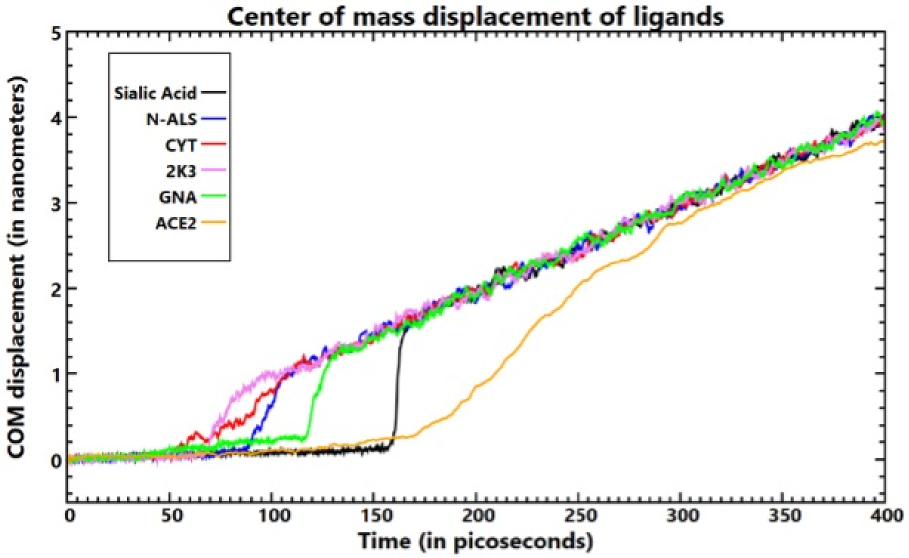

Figure 5: COM separation curves from SMD simulations for the 5 selected molecules and ACE2.

CoV-2 viral protein and how the presence of the glycosylation sites impact their binding on the protein surface. In addition, the N-terminal domain (NTD) of the viral glycoprotein of SARS-CoV-2 shows higher binding propensity with host sialic acid molecules as compared to two of its previously found variants: SARS-CoV and MERS-CoV (16). From our results, the sialic acids, N-ALS and GNA show highest docking scores when they bind with the RBD at a specified binding pocket of the RBD structure, generated by residues ASP457, LYS458, ASP467, SER469, GLU471, ILE472, TYR473 and GLN474 while the highest scored docked poses for the other two molecules (CYT and 2K3) are different. The best docked poses (shown in 12) for each ligand are further considered for the MD simulations.

To prepare 15, the predictions made using the frameworks Morgan-CNN and Morgan-AAC are used as these models have the highest CI value. It is observed that both the models and the MD simulations rank N-acetyl-neuraminic acid i.e.

Table 13: Simulated compounds and their respective maximum SMD unbinding force and corresponding ranks.

Compound	Peak SMD force (kJmol^{-1}nm^{-1})	Rank
2-keto-3-deoxyoctonate ammonium salt	632	4
N-Glycolylneuraminic acid	938	2
Cytidine-5-monophospho-N-acetylneuraminic acid	518	5
N-acetyl-neuraminic acid	1456	1
N-Aceytllactosamine	772	3
ACE2 Protein	1425	Benchmark

Table 14: Ranking of compounds based on binding affinity measured from different techniques. Ranks within '()' refer to absolute ranking based on the 13 molecules analyzed. Ranks without () refer to relative ranking based on only the 5 molecules for which MD simulations have been performed.

Compound	DeepDTA-AttentionSite	DeepDTA	MD simulation	In-Lab Validation
2-keto-3-deoxyoctonate ammonium salt	5(10)	4(12)	4	1(1)
N-Glycolylneuraminic acid	4(8)	2(8)	2	2(2)
Cytidine-5-monophospho-N-acetylneuraminic acid	3(7)	5(13)	5	3(5)
N-acetyl-neuraminic acid	1(2)	3(10)	1	4(9)
N-Aceytllactosamine	2(3)	1(5)	3	5(11)

sialic acid as the ligand with highest binding affinity towards RBD. However, this is not supported by the in-lab validation experiments which were conducted using biochemical reactions. The in-lab results rank 2-keto-3-deoxynononic ammonium salt and N-Glycolylneuraminic acid as number 1 and 2 respectively which is not supported by the DL predictions. Interestingly the MD simulation supports the in-lab results by ranking N-Glycolylneuraminic acid as number 2. The same could be said about Cytidine-5-monophospho-N-acetylneuraminic acid which is ranked 5th in both MD and in-lab validation. Hence, it may be said that combining both DL and MD simulations, the design space of ligand-protein simulations can be more reliably sampled which can optimize the selection of in-lab experiments. It may be worthwhile to note here that sialic acid has been shown to be having strong affinity to spike glycoprotein of SARS-CoV-2 in experimental lateral flow assay studies (17). Therefore, it would be important to re-check the in-lab validation experiments.

Table 15: List of antiviral drugs

Drug Names		
Abacavir	Famciclovir	raltegravir
Abacavir sulfate	Favipiravir (FPV)	Raltegravir potassium (RAL)
acyclovir	Fomivirsen sodium (FMV)	Remdesivir
Acyclovir	Fosamprenavir calcium (FPV)	Ribavirin
Adefovir Dipivoxil	Foscarnet sodium (PFA)	Rilpivirine
Amprenavir (agenerase)	ganciclovir	Rilpivirine hydrochloride (RPV)
Asunaprevir (BMS-650032)	Ganciclovir	Rimantadine (RIM)
Atazanavir	Ganciclovir sodium (GCV)	ritonavir
Atazanavir sulfate (BMS-232632-05)	Imiquimod (IMQ)	Ritonavir
Baloxavir marboxil (BXM)	indinavir	S/GSK1349572
Boceprevir	Indinavir sulfate (IDV)	saquinavir
Brivudine (BVDU)	Interferon alfa-2b (INT2B)	Saquinavir mesylate
Cidofovir	Laninamivir octanoate (LO)	Simeprevir
Daclatasvir (BMS-790052)	Letermovir (LET)	Simeprevir sodium (SIM)
danoprevir	Lopinavir	Sofosbuvir (SOF)
Darunavir	lopinavir	Telaprevir (VX-950)
Darunavir ethanolate (DRV)	MK-5172	Tenofovir
Delavirdine mesylate (DLV)	nelfinavir	tenofovir
Dolutegravir sodium (DTG)	Nelfinavir	Tenofovir alafenamide fumarate (TAF)
Doravirine (DOR)	Nelfinavir Mesylate	Tenofovir Disoproxil Fumarate
Efavirenz	Nevirapine	Tipranavir (TPV)
efavirenz	Oseltamivir	Trifluridine (TFT)
elvitegravir	Oseltamivir acid	valaciclovir
Elvitegravir (GS-9137)	Oseltamivir phosphate	Valaciclovir HCl
Enfuvirtide (T20)	penciclovir	Valganciclovir HCl
entecavir	Penciclovir	VX-222 (VCH-222, Lomibuvir)
Entecavir (ENT)	Peramivir	Zanamivir
Entecavir Hydrate	Peramivir Trihydrate	
Etravirine (TMC125)	Podofilox (PDX)	

References

[1] Yazdani-Jahromi M, Yousefi N, Tayebi A, Kolanthai E, Neal C J, Seal S, and Garibay O O. Attentionsitedti: an interpretable graph-based model for drug-target interaction prediction using nlp sentence-level relation classification. *Briefings in Bioinformatics* , 2022;**23(4)**:bbac272.

[2] Saberi Fathi S M and Tuszynski J A. A simple method for finding a protein's ligand-binding pockets. *BMC Structural Biology* , 2014;**14(1)**:18. URL `http://dx.doi.org/10.1186/1472-6807-14-18`

[3] Cichońska A, Ravikumar B, Allaway R J, Wan F, Park S, Isayev O, Li S, Mason M, Lamb A, Tanoli Z et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nature communications* , 2021; **12(1)**:3307.

[4] Tang J, Ravikumar B, Alam Z, Rebane A, Vähä-Koskela M, Peddinti G, van Adrichem A J, Wakkinen J, Jaiswal A, Karjalainen E et al. Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell chemical biology* , 2018;**25(2)**:224–229.

[5] Wang Y, Ribeiro J M L, and Tiwary P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current opinion in structural biology* , 2020;**61**:139–145.

[6] Ribeiro J M L and Tiwary P. Toward achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through rave. *Journal of Chemical Theory and Computation* , 2019; **15(1)**:708–719.

[7] Bennett W F D, He S, Bilodeau C L, Jones D, Sun D, Kim H, Allen J E, Lightstone F C, and Ingolfsson H I. Predicting small molecular transfer free energies by combining molecular dynamics simulations and deep learning. *J. Chem. Inf. Model.* , 2020;**60(11)**:5375–5381.

[8] Joshi T, Joshi T, Pundir H, Sharma P, Mathpal S, and Chandra S. Predictive modeling by deep learning, virtual screening and molecular dynamics study of natural compounds against sars-cov-2 main protease. *Journal of Biomolecular Structure and Dynamics* , 2020;**39(17)**:6728–6746.

[9] Ling R, Dai Y, Huang B, Huang W, Yu J, Lu X, and Jiang Y. In silico design of antiviral peptides targeting the spike protein of sars-cov-2. *Peptides* , 2020;**130**:170328.

[10] Fratev F, Steinbrecher T, and Jónsdóttir S Ó. Prediction of accurate binding modes using combination of classical and accelerated molecular dynamics and free-energy perturbation calculations: an application to toxicity studies. *ACS omega* , 2018;**3(4)**:4357–4371.

[11] Gosai A, Ma X, Balasubramanian G, and Shrotriya P. Electrical stimulus controlled binding/unbinding of human thrombin-aptamer complex. *Scientific reports* , 2016;**6(1)**:1–12.

[12] Berendsen H, van der Spoel D, and van Drunen GROMACS R. A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm. 91 (1995)* , 1995;:43–56.

[13] Hub J S, De Groot B L, and Van Der Spoel D. g_wham a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of chemical theory and computation* , 2010;**6(12)**:3713–3720.

[14] Lee H. All-atom simulations and free-energy calculations of antibodies bound to the spike protein of sars-cov-2: The binding strength and multivalent hydrogen-bond interactions. *Advanced Theory and Simulations* , 2021;**4(5)**:2100012.

[15] Bingqian L, Lin W, Huan G, Xianglei Z, Penxuan R, Yu G, Wuyan C, Jie L, Wei Z, Wenzhang C, Lili Z, and Fang B. Identification of potential binding sites of sialic acids on the rbd domain of sars-cov-2 spike protein. *Frontiers in chemistry* , 2021;**9**:659764.

[16] Awasthi M, S G, DP S, S T, S K, P R, and SK V. The sialoside-binding pocket of sars-cov-2 spike glycoprotein structurally resembles mers-cov. *Viruses* , 2020;**19(12(9)**:)909.

[17] Baker A N, Richards S J, Guy C S, Congdon T R, Hasan M, Zwetsloot A J, Gallo A, Lewandowski J R, Stansfeld P J, Straube A et al. The sars-cov-2 spike protein binds sialic acids and enables rapid detection in a lateral flow point of care diagnostic device. *ACS central science* , 2020; **6(11)**:2046–2052.