SARS-CoV-2 Drug Discovery based on Intrinsically Disordered Regions

Anish Mudide
Phillips Exeter Academy
20 Main Street, Exeter, NH 03833, USA
Email: amudide@gmail.com

Gil Alterovitz
Biomedical Cybernetics Laboratory, Brigham and Women's Hospital and Harvard Medical School
Department of Veterans Affairs, National Artificial Intelligence Institute
25 Shattuck Street, Boston, MA 02115, USA
Email: ga@alum.mit.edu

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a close relative of SARS-CoV-1, causes coronavirus disease 2019 (COVID-19), which, at the time of writing, has spread to over 19.9 million people worldwide. In this work, we aim to discover drugs capable of inhibiting SARS-CoV-2 through interaction modeling and statistical methods. Currently, many drug discovery approaches follow the typical protein structure-function paradigm, designing drugs to bind to fixed three-dimensional structures. However, in recent years such approaches have failed to address drug resistance and limit the set of possible drug targets and candidates. For these reasons we instead focus on targeting protein regions that lack a stable structure, known as intrinsically disordered regions (IDRs). Such regions are essential to numerous biological pathways that contribute to the virulence of various viruses. In this work, we discover eleven new SARS-CoV-2 drug candidates targeting IDRs and provide further evidence for the involvement of IDRs in viral processes such as enzymatic peptide cleavage while demonstrating the efficacy of our unique docking approach.

1. Introduction

IDRs lack a fixed three-dimensional structure, and instead fold dynamically into a set of continuous conformations based on surrounding conditions [1]. This allows IDRs to have a wide range of binding partners, and as a result, serve significant roles in critical biological processes such as cell signaling and transcription [2-3]. Moreover, certain short IDRs known as molecular recognition features (MoRFs) are essential for initiating protein-protein interactions (PPIs) [4]. For over a decade now, it has been clear that IDRs are functionally important to and incredibly abundant in proteins implicated across the disease spectrum [5].

While IDRs are not incredibly common in the SARS-CoV-2 proteome, the IDRs that do exist contribute greatly to the functioning and overall virulence of the pathogen [6-7]. In fact, nearly all SARS-CoV-2 proteins are predicted to have MoRFs, highly suggestive of the importance of IDRs in PPI networks [7]. SARS-CoV-2 IDRs therefore serve as promising drug targets for antiviral drug discovery.

Of the 27 mature viral proteins within the SARS-CoV-2 proteome, the majority of current drug discovery research is largely focused on three main targets: the RNA polymerase, the Papain-like

protease, and the 3C-like protease (3CLpro) [8-9]. The 3CLpro's main role is to cleave the polyproteins into functional parts [10]. While all three targets are disordered [7], in this study we focus on the CoV-2 3CLpro since it is highly similar (96% sequence identity) to its CoV-1 relative, for which an abundance of bioassay data is available [10]. In particular, we concentrate our efforts on the N-terminally short IDR (residues 1-6; see footnote 'b') predicted to be a MoRF in both CoV-1 and CoV-2 [7]. A drug capable of binding to this IDR could thereby inhibit PPIs within the virus.

Our approach to drug discovery consists of two major steps. First, we compute binding affinities between the CoV-2 3CLpro IDR and over 1400 ligands from the NCI Diversity Set III through a unique docking procedure. While older docking procedures focus on targeting structured protein pockets [11], in this study we account for the wide range of IDR conformations through the allowance of residue side chain rotation as well as through ensemble docking. High binding affinities are a key first indicator of drug potential since they imply a great attractive force toward the receptor and demonstrate that the binding energy can be used to alter the receptor structure. We discovered over 60 ligands with binding affinities of -8.0 kcal/mol or better. However, drug discovery approaches relying solely on docking often fail to produce seriously meaningful results, and expert opinion suggests the cross-verification of results using distinct techniques [12-13]. Thus, in the second step of our approach, we validate and filter our results using a statistical model. The results of bioassay AID 1706, which screens over 290,000 compounds for inhibition of CoV-1 3CLpro-mediated peptide cleavage [14], are used to train a message passing neural network (MPNN) to distinguish between positive (3CLpro inhibiting) and negative (non-inhibiting) compounds. Due to the high similarity between the two CoV 3CLpros, such a model is likely to make meaningful predictions relevant to CoV-2 3CLpro inhibition [10, 15]. This model is then used to predict activity scores for each of the previously docked ligands. We show a correlation between activity scores and binding affinity, suggesting the efficacy of our docking approach. Moreover, we combine the results of our steps to determine 11 new CoV-2 drug candidates, many of which show antibiotic or antiviral properties. Figure 1 summarizes the process.

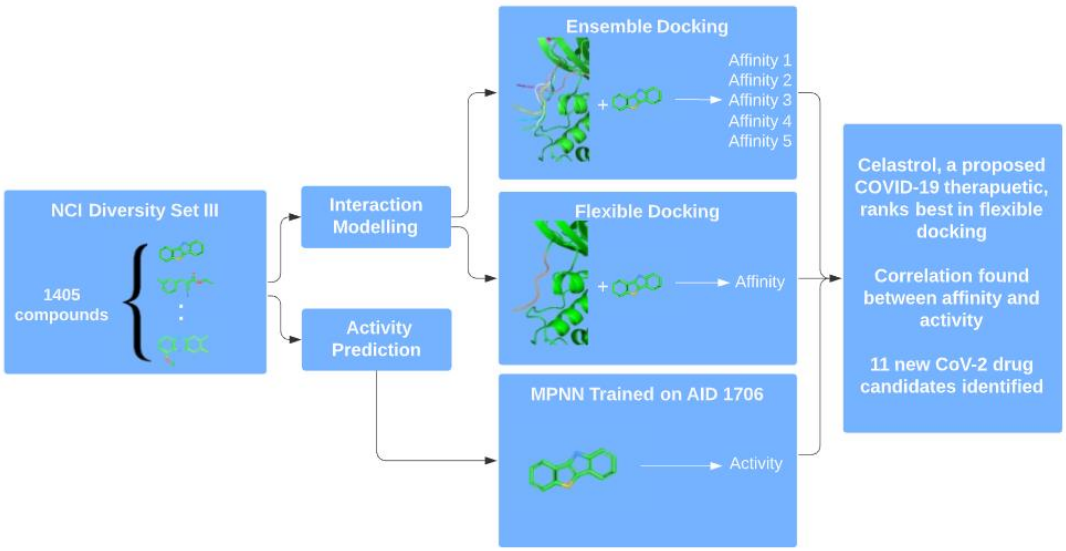

Fig. 1. Drug discovery flowchart.

2. Methods

2.1. *Molecular docking*[a]

2.1.1. *Data collection*

The three-dimensional structures of all 27 mature viral proteins were predicted by the Oak Ridge National Laboratory (ORNL) with a workflow consisting of X-ray crystallography results, homology modeling, and disorder prediction among other techniques. In particular, the structure of the monomeric form of 3CLpro was obtained directly from ORNL's COVID-19 site.[b]

In this study, we use the National Cancer Institute (NCI) Diversity Set III as our ligand dataset. Diversity sets are constrained such that no two ligands can be overly similar to one another, resulting in heterogeneity. A single SDF file was retrieved from NCI's website[c] describing the structures of each of the ligands in the set.

2.1.2. *Data preprocessing*

AutoDockTools was used to prepare and preprocess the PDB file for the 3CLpro before docking. Water molecules were removed, polar hydrogen atoms were added, and Kollman charges were added to the entire structure. The structure was then saved as a PDBQT file.

The ligands were extracted from the SDF file into individual PDB files. Then, the prepare_ligand function from the AutoDock Flexible Receptor (ADFR) suite[d] was used to preprocess each of these ligand files, generating PDBQT files ready for docking.

2.1.3. *Target file generation*

The protein-ligand docking software used in this study is AutoDock Flexible Receptor (ADFR). ADFR requires at least two parameters to be passed: the protein receptor, specified by a target file, and the ligand, specified by a PDBQT file. Target files specify the docking box size and position, calculated binding pockets, residue side chains to be made flexible, affinity maps, as well as other meta-data. AutoGridFR was used to generate such a target file for the 3CLpro. In particular, the docking box was specified to enclose residues 1-9, and residues within the IDR (1-6) were specified as having flexible side chains. Additionally, AutoSite 1.0 was used to generate ligand binding pockets through a clustering algorithm that groups high affinity points into disjoint "fills." Fills with high scores in close proximity to the disordered region were chosen to be targeted during docking. Figure 2 graphically summarizes the parameters chosen for target file generation.

[a] Our code, data and results are available at https://github.com/Biomedical-Cybernetics-Lab2/IDR-SARS-CoV-2.
[b] https://compsysbio.ornl.gov/covid-19/covid-19-structome/.
[c] https://wiki.nci.nih.gov/display/NCIDTPdata/Compound+Sets.
[d] https://ccsb.scripps.edu/adfr/.

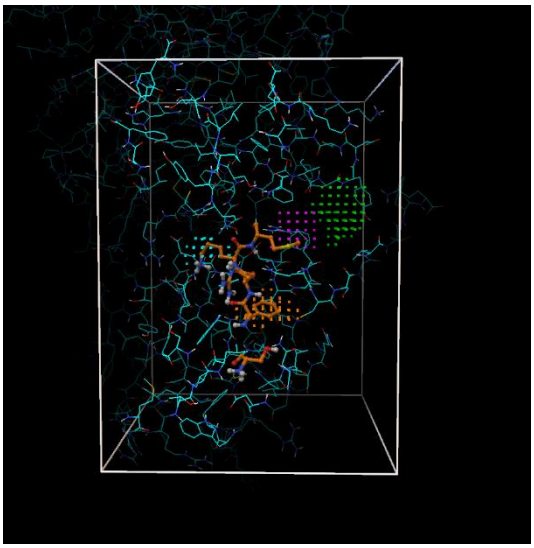

Fig. 2. Orange residues are part of the IDR and specified as flexible. The docking box is shown in white. Fills chosen are shown in purple, light blue, green and orange.

2.1.4. *Flexible docking*

ADFR is unique from other protein-ligand docking software in that it can handle both ligand and receptor flexibility. As a result, ADFR is of incredible use when performing IDR-related docking. ADFR employs a genetic algorithm (GA) to find the best docked position of a given ligand. For each protein-ligand pair, the GA is run several times in case the GA converges to local rather than global optima. Moreover, the user can specify both how many runs are executed as well as an upper bound for the number of times the scoring function is called per run. This allows us to drastically cut down on compute time by potentially terminating searches before they converge. The default values for the number of GA runs and the maximum number of score evaluations are 50 and 2.5 million respectively; in this study, at least initially, we modify these parameters to 7 runs with at most 28,000 evaluations each. Docking is performed with these parameters for 1405 distinct ligands from the NCI Diversity Set III, and results are compiled.

2.1.5. *Ensemble docking*

In our pursuit of simulating the conformational flexibility of the IDR for accurate drug discovery, we also utilize ensemble docking. In this approach, we generate many possible conformations of the IDR, and dock each ligand onto each possible conformation. In this study, we generate conformations by treating the IDR as a loop of the protein. Loop modelling implemented by MODELLER[e] is then used to generate five likely IDR conformations. We then repeat the processes outlined in the above sections: we preprocess each newly generated PDB file, generate a target file for each, and perform docking on each conformation-ligand pair using ADFR. Figure 3 illustrates how the five different conformations of the IDR compare to each other. After docking is complete, results are compiled.

[e] https://salilab.org/modeller/.

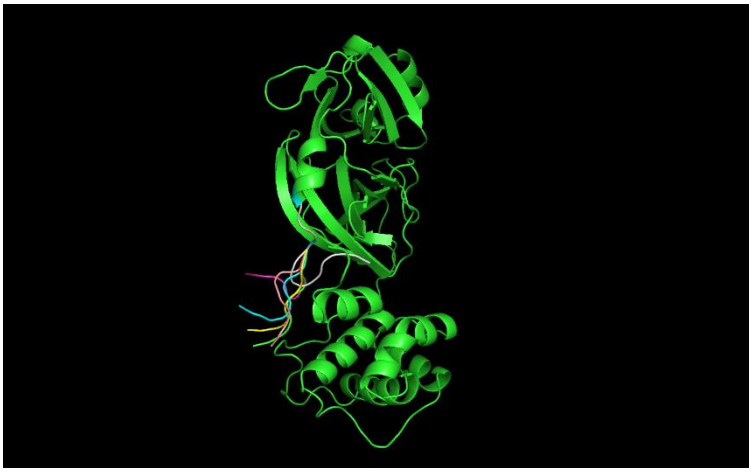

Fig. 3. Five conformations of the 3CLpro IDR superimposed onto the original structure.

2.2. *Statistical model*

2.2.1. *Chemprop*

Chemprop[f] is a freely available implementation of a message passing neural network. Such models are designed to predict properties of graph-based inputs. In the case of molecular property prediction, molecules are transformed to graphs by treating atoms as nodes and the bonds between atoms as edges. Using this representation, a feature vector is generated through a learned algorithm that aggregates chemical features within the graph. This vector is then passed to a typical feed-forward neural network [16]. For our purposes, this neural network outputs a real value between 0 and 1 representing the model's confidence that a certain molecule has a desired binary property.

2.2.2. *Data and training*

Our aim was to train a MPNN model to predict whether a molecule can inhibit the CoV-2 3CLpro *in vitro* to further validate and filter our results from molecular docking. We make use of the results from bioassay AID 1706, which screens over 290,000 compounds for inhibition of CoV-1 3CLpro peptide cleavage, to train such a model. Concretely, the bioassay screens for cleavage inhibition by attaching a fluorescent compound and a quencher to opposite sides of a 3CLpro substrate. A compound can then be classified as active or inactive since fluorescence increases if and only if cleavage occurs [14]. Due to the high similarity between the two CoV 3CLpros, a model trained on CoV-1 data is likely to make meaningful predictions relevant to CoV-2 3CLpro inhibition. Each training example in the dataset[g] consists of one feature (the SMILES string of the compound) and one label (a binary output; 1 for inhibition, 0 for no inhibition). Just 405 of the compounds are classified as positive (label = 1), whereas the other 290,321 compounds are negative (label = 0). To

[f] https://github.com/chemprop/chemprop.
[g] Retrieved from https://github.com/yangkevin2/coronavirus_data.

account for this imbalance between positive and negative data points in the training set, an equal number of positives and negatives are used in each batch during training. Furthermore, additional features generated by RDKit are appended to the feature vector generated before being passed into the neural network during training and predicting. Once trained, the model achieves a test ROC AUC of .739. We then apply the model to predict activity scores for each of the previously docked ligands.

3. Results

3.1. *Interaction modelling*

The binding affinities of over 1400 ligands with the proposed IDR target were analyzed *in silico* using molecular docking. We first simulated IDR conformational flexibility by allowing IDR residue side chains to rotate while searching for the optimal ligand pose. With this docking procedure, 57 ligands were found to have binding affinities of -8.0 kcal/mol or better. Considering that we terminated the docking searches before convergence by bounding the maximum number of score evaluations, their true binding affinities are likely to exceed -8.0 kcal/mol. Therefore, we deemed all 57 ligands as ideal drug candidates. Table 1 summarizes these results of this first docking procedure.

Table 1. Binding affinity results from flexible docking (abridged)

Molecule (NSC)	Binding Affinity (kcal/mol)
70931	-9.8
177862	-9.7
16437	-9.3
96541	-9.1
117987	-8.8
45527	-8.8
…	…

With a binding affinity of -9.8 kcal/mol, the top molecule found is NSC-70931, also known as the triterpenoid named celastrol. Celastrol displays antiviral properties against influenza A virus as well as dengue virus in mice [17-18]. In fact, celastrol has already been suggested as an anti-inflammatory therapeutic for the lethal pneumonia stage of COVID-19 [19]. These results indicate the potential of our first docking method.

When we reran the docking of celastrol onto the 3CLpro IDR with the default parameters mentioned above, the search converged and found a pose with an improved docking score of -11.4 kcal/mol (shown in Figure 4). This further solidifies our claim that the binding affinities presented in this study are likely sub-optimal.

Fig. 4. Docked pose of celastrol (-11.4 kcal/mol) after search converged.

We then simulated IDR conformational flexibility using a different approach known as ensemble docking. Concretely, each ligand was docked onto five distinct conformations of the IDR generated by loop modelling techniques. These five binding affinities were retrieved, but only the highest of the five was used to compare ligands with each other. With this docking procedure, 49 ligands were found to have highest binding affinities of -8.0 kcal/mol or better. Table 2 summarizes the results of this second docking procedure.

Table 2. Binding affinity results from ensemble docking (abridged)

Molecule (NSC)	Best Binding Affinity (kcal/mol)
166259	-9.3
37641	-9.1
121868	-9.1
727038	-9.1
117987	-8.7
70931	-8.6
…	…

The top molecule found is NSC-166259, a close relative of succinic acid found to have a highest binding affinity of -9.3 kcal/mol with conformation 2 of the IDR. NSC-166259 displays anticancer properties, showing activity in human tumor cell bioassays. Upon closer inspection of NSC-166259's docked pose, it becomes apparent that NSC-166259 interacts with the receptor at two sites: residue 126 as well as residue 3, which is within the IDR (see Figure 5). This confirms the notion that our docking approach can find ligand poses that interact directly with the IDR.

Finally, given the current need for efficient drug discovery through repurposing, a set of well-known compounds such as danazol, genistein and estramustine found to perform well in both docking procedures are listed along with their modern uses in Table 3.

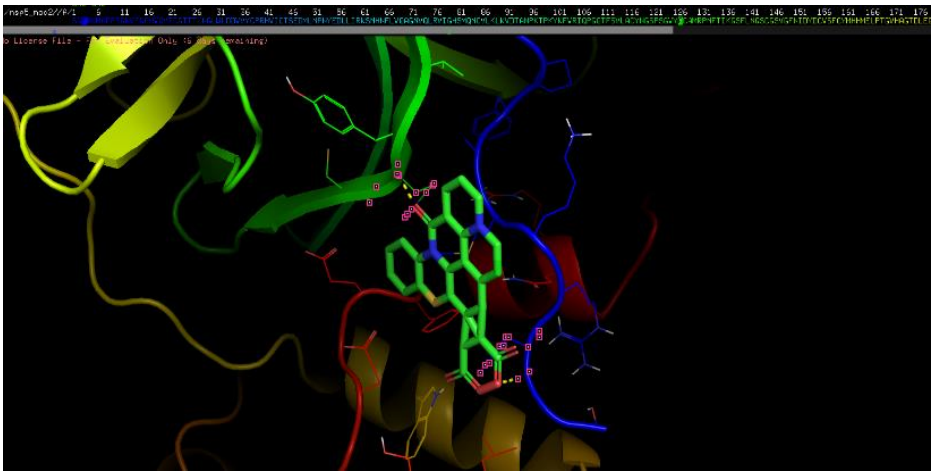

Fig. 5. NSC-166259 interacting with IDR.

Table 3. Drug repurposing candidates with their uses and additional information.

Drug	Pharmacological Use	Additional Information
Celastrol	Inflammation, cancer (lung, prostate)	Suppresses NF-kB signaling
Danazol	Fibrocystic breast disease, endometriosis	Targets estrogen receptor alpha
Estramustine	Cancer (prostate)	Targets estrogen receptor alpha/beta
Camptothecin	Cancer	Targets topoisomerase
Genistein	Cardiovascular risk, cancer	Targets estrogen receptor alpha/beta
Benzbromarone	Heart failure, chronic kidney disease	Targets cytochrome P450 2C9

3.2. *Activity prediction*

The goal of this work is to find CoV-2 3CLpro inhibitors by concentrating our efforts on the IDR/MoRF present at the N-terminus. The first step in this effort using molecular docking yielded promising results; however, in general, docking approaches need to be cross verified by a different method. Thus, our next goal was to create a statistical model capable of predicting *in vitro* inhibition of our protein target to filter and provide further evidence for our docking results. Such a model would be capable of making predictions many orders of magnitude faster than standard bioassays. Here, we train a model to predict whether a compound can inhibit 3CLpro-mediated peptide cleavage.

Due to the scarcity of CoV-2 data, we train our model using CoV-1 3CLpro peptide cleavage inhibition data from bioassay AID 1706. The model structure chosen is a MPNN implemented by Chemprop. Our model achieves a test ROC AUC of .739 (80% train, 10% validation, 10% test).

We then apply the trained model to predict activity scores for each of the previously docked ligands. A total of 11 ligands (see Table 4) are identified as having both high affinity (absolute affinity $\geq$ 7.9 kcal/mol) as well as high activity ($\geq$ 0.8). These ligands have high probabilities of binding to the IDR, having enough binding energy to deform the 3CLpro, and inhibiting peptide cleavage. Therefore, we deem these 11 ligands promising drug candidates. Furthermore, known use

cases for these 11 ligands include orthopoxviruses, foot-and-mouth disease virus, human tumors, and malaria. We are currently investigating a potential collaboration to validate the efficacy of these 11 new drug candidates *in vitro*.

Table 4. Top 11 drug candidates in terms of affinity and activity.

Molecule (NSC)	Activity	Affinity	Active Bioassays
16437	.859	-9.3	Foot-and-mouth disease (FMD) virus
117987	.872	-8.8	
601359	.855	-8.4	Melanoma cell line, Malaria
13294	.825	-8.4	
127133	.908	-8.3	
61610	.823	-8.2	Malaria
107582	.877	-8.1	
128606	.920	-8.0	
211490	.808	-8.0	Hepatitis C virus, Human cytomegalovirus
679525	.894	-8.0	Orthopoxviruses, FMD virus
204232	.800	-7.9	DNA Polymerase Beta

We also investigate the possible link between 3CLpro cleavage inhibition and IDR binding affinity. A scatter plot of the binding affinities and activity scores for each of the 1405 docked ligands is shown in Figure 6. The correlation coefficient *r* measuring the strength and direction of the linear relationship between the two variables is computed to be 0.38, suggesting a weak to slightly moderate correlation. This means that higher binding affinities to the IDR of the CoV-2 3CLpro weakly/moderately correlate with higher rates of cleavage inhibition. This suggests that the IDR/MoRF of the CoV-2 3CLpro is involved in the peptide cleavage process. As a matter of fact, it is well known that the dimerization of 3CLpro that develops its active site involves our targeted IDR [7]. Therefore, since our method realizes this relationship, it suggests that targeting the IDR in the monomeric form is an effective way of finding 3CLpro peptide cleavage inhibitors. This also could suggest that our approach of cross verifying docking results with statistical models could be used to hypothesize other biological relationships key to drug discovery in the future. In Figure 7, we show the distribution of the IDR binding affinities of known CoV-1 3CLpro inhibitors from bioassay AID 1706, and in Figure 8 we show the same distribution for the NCI Diversity Set III. We find the average binding affinity of CoV-1 3CLpro inhibitors to be -6.74 kcal/mol, which is above the typical threshold for choosing possible drug candidates, whereas the average for the NCI Diversity Set III, which we assume to be a representation of the drug-like ligand space, is just -5.93 kcal/mol. Consequently, the distributions indicate that the average 3CLpro inhibitor falls within the top 23.5% of all ligands in terms of binding affinity to the IDR of 3CLpro, further supporting our previous claims. Furthermore, it is possible that the correlation between the IDR and cleavage inhibition is higher than mentioned above but is dampened in our data since the MPNN was trained on *in vitro* results, but high binding affinities do not always correspond to *in vitro* binding.

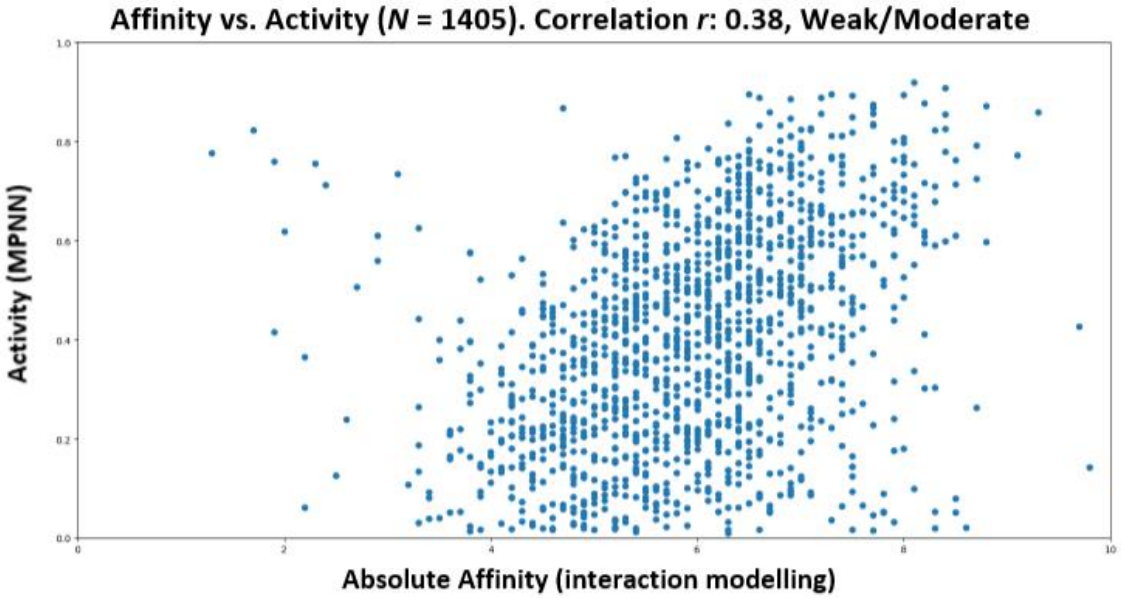

Fig. 6. Affinity versus activity

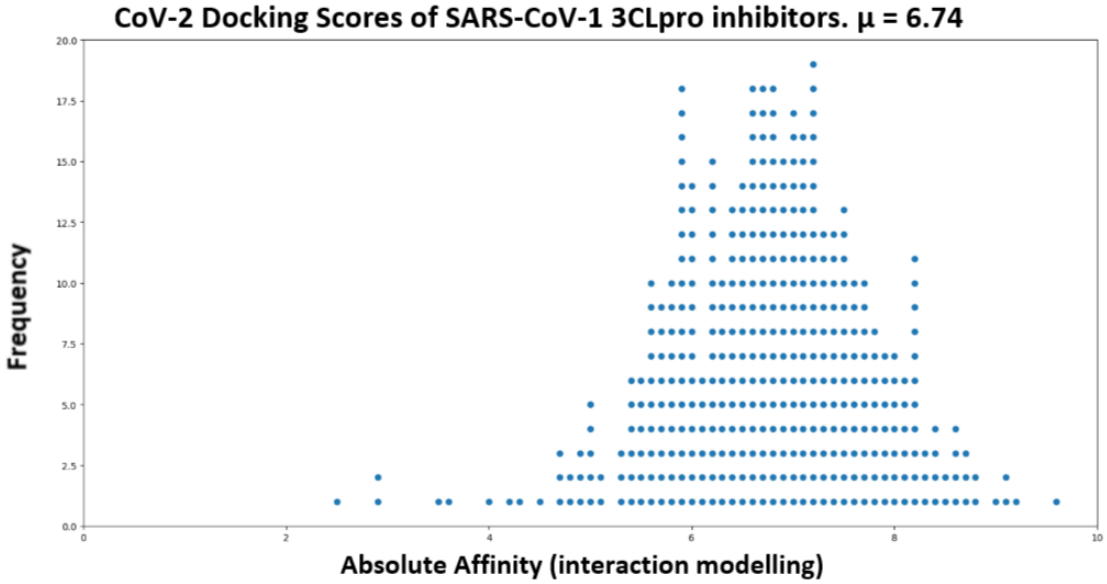

Fig. 7. Distribution of 3CLpro inhibitor binding affinities.

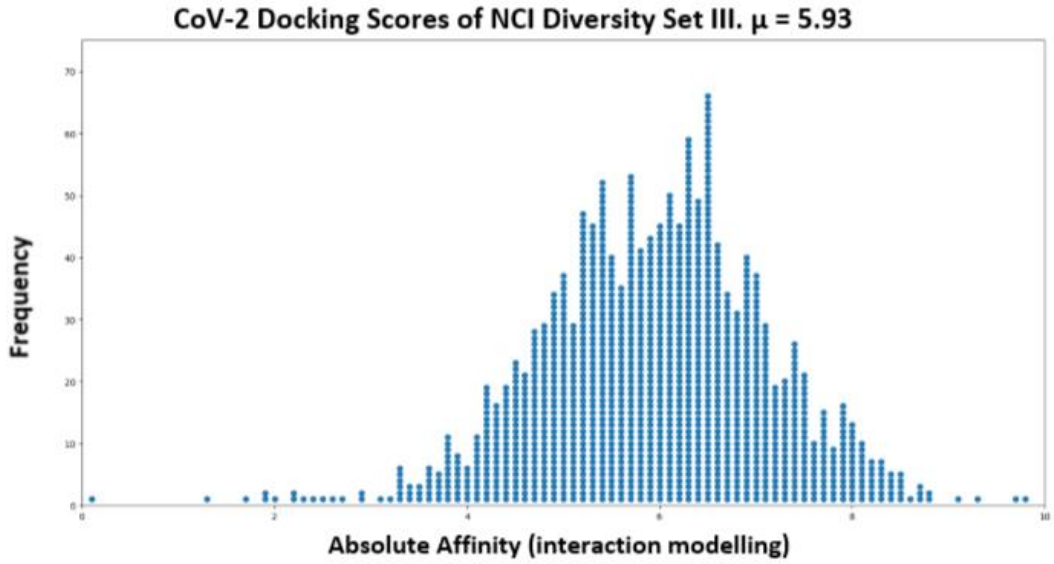

Fig. 8. Distribution of NCI Diversity Set III binding affinities.

4. Conclusion

Currently, there are no widely approved CoV-2 antivirals or vaccines available. Given the infectious and fatal nature of COVID-19, there exists a dire need for immediate drug discovery research. In this work, we make advancements by specifically focusing on targeting disordered protein regions. We demonstrate how these IDRs can be targeted through molecular docking and illustrate how results can be verified in a multi-faceted approach. Ultimately, we identify 11 new drug candidates with high binding and activity scores, along with known antiviral properties. In the future we would like to validate our results *in vitro* as well as further explore the IDR interactions within the SARS-CoV-2 proteome through MoRF mimicry.

5. Acknowledgements

This research was undertaken as part of the MIT-PRIMES program. Ning Xie and Ling Teng of the Biomedical Cybernetics Lab provided frequent support, feedback, and organization.

References

1. Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews. Molecular cell biology*, *16*(1), 18–29.
2. Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, *293*(2), 321-331.
3. Rhoades, E. (2018). *Intrinsically Disordered Proteins*. Academic Press.
4. Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., & Uversky, V. N. (2006). Analysis of molecular recognition features (MoRFs). *Journal of molecular biology*, *362*(5), 1043-1059.

5. Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.*, *37*, 215-246.
6. Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D., & Huang, T. H. (2014). The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral research*, *103*, 39–50.
7. Giri, R., Bhardwaj, T., Shegane, M., Gehi, B. R., Kumar, P., Gadhave, K., ... & Uversky, V. N. (2020). When Darkness Becomes a Ray of Light in the Dark Times: Understanding the COVID-19 via the Comparative Analysis of the Dark Proteomes of SARS-CoV-2, Human SARS and Bat SARS-Like Coronaviruses. *bioRxiv*.
8. Wang, R., Hozumi, Y., Yin, C., & Wei, G. W. (2020). Decoding SARS-CoV-2 Transmission and Evolution and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine. *Journal of chemical information and modeling*, acs.jcim.0c00501. Advance online publication.
9. Joshi, S., Joshi, M., & Degani, M. S. (2020). Tackling SARS-CoV-2: proposed targets and repurposed drugs. *Future medicinal chemistry*, 10.4155/fmc-2020-0147. Advance online publication.
10. Chen, Y. W., Yiu, C. B., & Wong, K. Y. (2020). Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL ^{pro}) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*, *9*, 129.
11. Antunes, D. A., Devaurs, D., & Kavraki, L. E. (2015). Understanding the challenges of protein flexibility in drug design. *Expert opinion on drug discovery*, *10*(12), 1301-1313.
12. Thafar, M., Raies, A. B., Albaradei, S., Essack, M., & Bajic, V. B. (2019). Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. *Frontiers in Chemistry*, *7*.
13. Kairys, V., Baranauskiene, L., Kazlauskiene, M., Matulis, D., & Kazlauskas, E. (2019). Binding affinity in drug design: experimental and computational techniques. *Expert opinion on drug discovery*, *14*(8), 755–768.
14. National Center for Biotechnology Information (2020). PubChem Bioassay Record for AID 1706, Source: The Scripps Research Institute Molecular Screening Center.
15. Suárez, D., & Díaz, N. (2020). SARS-CoV-2 Main Protease: A Molecular Dynamics Study. Journal of chemical information and modeling.
16. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., ... & Palmer, A. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, *59*(8), 3370-3388.
17. Khalili, N., Karimi, A., Moradi, M. T., & Shirzad, H. (2018). In vitro immunomodulatory activity of celastrol against influenza A virus infection. *Immunopharmacology and Immunotoxicology*, *40*(3), 250-255.
18. Yu, J. S., Tseng, C. K., Lin, C. K., Hsu, Y. C., Wu, Y. H., Hsieh, C. L., & Lee, J. C. (2017). Celastrol inhibits dengue virus replication via up-regulating type I interferon and downstream interferon-stimulated responses. *Antiviral research*, *137*, 49-57.
19. Habtemariam, S., Nabavi, S. F., Berindan-Neagoe, I., Cismaru, C. A., Izadi, M., Sureda, A., & Nabavi, S. M. (2020). Should we try the antiinflammatory natural product, celastrol, for COVID-19?. *Phytotherapy Research*.