Pacific Northwest
NATIONAL LABORATORY

PNNL- 33349

Machine Learning-driven Molecular Design for Therapeutic Discovery

September 2022

1 Rohith Varikoti
2 Katherine Schultz
3 Mowei Zhou
4 Chathuri Kombala
5 Kris Brandvold
6. Agustin Kruel
7. Neeraj Kumar

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL- 33349

Machine Learning-driven Molecular Design for Therapeutic Discovery

September 2022

1 Rohith Varikoti
2 Katherine Schultz
3 Mowei Zhou
4 Chathuri Kombala
5 Kris Brandvold
6. Agustin Kruel
7. Neeraj Kumar

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

The ongoing novel coronavirus pandemic (COVID-19) has highlighted the need for new therapeutics to counter the threat of emerging viral pathogens. The main proteases are a promising target for developing antiviral inhibitors. In this work, we utilized a novel combination of artificial intelligence-driven iterative design of covalent inhibitor candidates, physics-based computational modeling of protein-inhibitor interactions, and “All in One” Native MS biophysical assay screening and characterization of designed candidates. With our existing expertise in hit generation using a particular scaffold as a starting point, we first generated tens of thousands of compounds that preserve the key scaffold. In order to optimize the candidates, we calculated about 136 descriptors consisting of 2D and 3D features for molecules targeting the SARS-CoV-2 Main protease (Mpro). These compounds were initially filtered according to properties and further sorted by predicted binding affinity using our automated docking modeling and machine learning methods. We tested a handful of candidates and identified two as inhibitors of Mpro with micromolar affinities.

Summary

We develop a computational strategy that will transition from hit-finding based on explainable AI and computational methods to a deeper analysis and iterative design-make-test cycles to include a set of chemical modifications around a common core with clear structure-activity relationships (SAR) of various properties. These candidates were validated using PNNL's screening and native MS to define molecular mechanisms for rapid iteration of AI design. The tight integration between data scientists, modelers, and experimentalists provided a closed loop machine intelligent model that learns from protein specific data and builds an ML algorithm to identify novel candidates and perform lead optimization with broad spectrum antiviral properties, which can possibly revolutionize the drug discovery for fast response to future pandemics.

Acknowledgments

This research was supported by the I3T Investment, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830. The computational work was performed using PNNL Computing at Pacific Northwest National Laboratory. Part of the research was performed using the Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by the DOE's Office of Biological and Environmental Research and located at PNNL. PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DEAC05-76RL0-1830.

Contents

Figures

Tables

Introduction

The process of identifying and discovering novel small molecules with desired properties using conventional methods is time-consuming and expensive (Hughes *et al.*, 2011). Furthermore, the risk of such drug therapeutic candidates ultimately failing in subsequent pre-clinical trials is very high (Sun, D. *et al.*, 2022). Over the past 2 years, the COVID-19 pandemic has posed challenges to develop potent drug candidates while overcoming time and cost constraints. The rapid spread of novel SARS-CoV-2 viral variants underscores the need for an efficient platform for developing small molecule antiviral drug leads that show high efficacy targeting the viruses while being nontoxic. It is a monumental task to obtain a potent drug from scratch using traditional approaches like target identification and validation, hit discovery, high throughput screening assays, and toxicity assays. However, the availability of compound libraries containing structural, functional, and therapeutic information for previously approved drugs and millions of chemical compounds from databases like Enamine (Shivanyuk *et al.* 2007), Mcule (Kiss *et al.*, 2012), and ChEMBL (Gaulton *et al.,* 2017) allows for the leveraging of computational resources and expertise at PNNL to aid in understanding and searching the vast chemical space of the compounds as a starting point. With the availability of several open source in silico tools—including those developed at PNNL— for high throughput virtual screening (HTVS), hit-to-lead identification, and optimization, recent advances in the field of machine learning (ML) and artificial intelligence (AI) have helped to accelerate drug discovery and development, thus expanding the scope of treatments for a wide variety of targets.

One such PNNL-developed tool, 3D-Scaffold (Joshi *et al.,* 2021) that we previously developed, utilizes deep learning with a fragment-based or functional group method, where a fragment or scaffold is used as an initial structure and molecules are generated using the scaffold as their core structure. A benefit of this approach is that scaffolds can be chosen from experimentally validated active compound libraries such that they retain key features with respect to a given target protein. The generated compound library can be further screened based on calculated physicochemical properties to identify whether the compounds are desirable candidates for proceeding to experiments. With this work, we developed a drug discovery and lead optimization (LO) workflow (**Figure 1**) for generating potential therapeutic candidates targeting the Mpro (Kneller *et al.*, 2020). We confirmed two candidates as inhibitors of Mpro using PNNL experimental resources.

Results and Discussions

To identify potential hits, we utilized our 3D-Scaffold model, high throughput virtual screening (HTVS) techniques, and cutting-edge hit identification and optimization methods as shown in our computational workflow (**Figure 1**). The key scaffolds (or chemical fragments) important for Mpro activity were identified and extracted from experimentally validated potential candidates. These scaffolds were then provided as input for our 3D-scaffold model to generate a library of compounds covering extensive chemical space. Predicted novel compounds were then screened and sorted based on cheminformatics, physiochemical properties, and similarity patterns with their parent compounds. The compounds were ranked based on the interpreted results and using molecular docking simulations to predict binding affinity. The compounds' conformations were visually inspected to elucidate the orientation of the compound in the binding pocket and observe key interactions of the compounds with the target protein. However, with the above-mentioned criteria one can reduce the number of molecules but not improve their potency. LO is then utilized to achieve optimal potency and interactions. With the increasing number of 2D and 3D structural properties, and their relevant importance to ADMET properties and activity, LO has become a challenging task to achieve. Several structural and ADMET properties are to be converged to a point where one can identify the key properties that are to be altered to obtain desired potent compounds. Here, this was achieved by 3D-QSAR and MPO analysis which provided us with 5 important descriptors to consider for lead optimization.

The screened hits were optimized further before testing them using experimental validation. Once we finalized the hits (**Table 1**), we ordered them and tested and characterized the final set of compounds with experimental methods using Native MS and FRET based functional assays (Clyde *et al.*, 2021). The capabilities and insights developed with this project will be ultimately applicable to a wide range of protein targets and biological systems of interest.

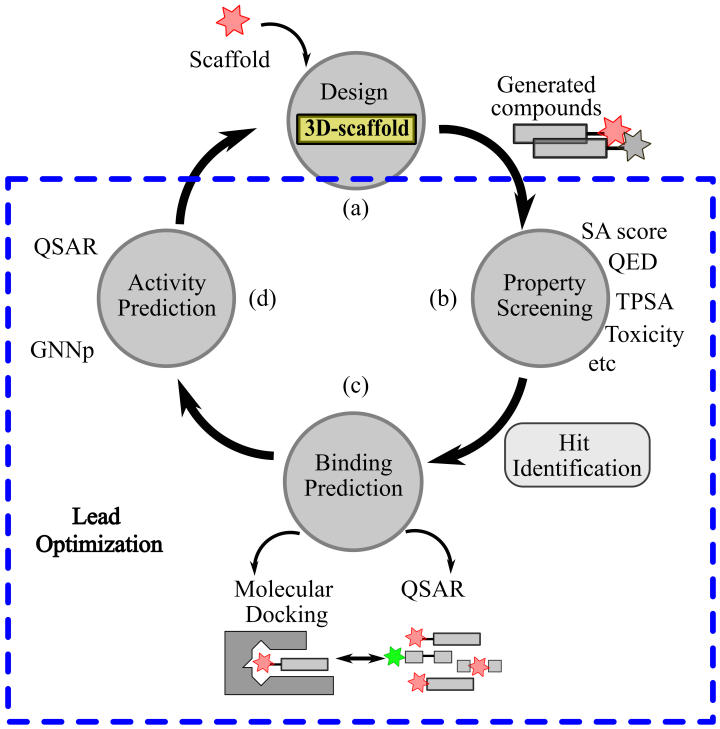

Figure 1. Representation of therapeutic candidate identification and lead optimization procedure followed in our research. (a) The process is initiated by providing a scaffold (*) as an input to our 3D-scaffold model that generates several ligands. (b) The generated compounds are screened based on different physiochemical properties to identify hits. (c) HTVS using molecular docking and QSAR to identify lead compounds; (d) ML/DL based activity prediction of lead compounds.

Table 1. Top 8 high throughput virtually screened compounds with their respective molecular properties that we finalized based on extensive computational studies and used for experimental characterization.

#	Mcule ID	MW	LP	TPSA	HA	HD	Docking Score	Synthetic Accessibility
1	MCULE-8568381615-0	457.17	2.149	104.4	6	1	-7.3	3.83
2	MCULE-8054614126-0	455.19	1.555	95.17	7	1	-7.2	3.96
3	MCULE-7471308738	401.53	0.919	131.00	7	4	-7.4	3.55
4	MCULE-5167696303	427.61	2.375	105.14	6	3	-6.5	3.89
5	MCULE-7052658287	410.54	2.4345	115.04	6	3	-6.7	3.82
6	MCULE-4947886566	334.37	0.779	81.89	6	1	-6.1	2.64

7	MCULE-4926166920	371.43	2.005	91.5	7	2	-6.6	2.93
8	MCULE-2238978486	398.50	1.613	85.25	7	2	-6.6	3.85

MW = Molecular weight; LP = partition coefficient (LogP); TPSA = topological polar surface area; HA and HD = number of hydrogen bond acceptors and donors; Docking score in kcal/mol; Synthetic Accessibility score between 1 (easy to synthesize) and 10 (very difficult to synthesize)

Mpro Library Generation

Recently, we developed our model 3D-Scaffold, a deep learning approach which generates the 3D coordinates of molecules built around a desired molecular scaffold provided as an input and training data sets. The identification of scaffolds is a critical step in the process, as it defines candidate generation. To identify core scaffolds to use as input, we curated a library of potent drug candidates with their IC50 and/or EC50 values (measurements of binding affinity) from various sources such as Protein Data Bank (RCSB PDB) (Burley *et al.,* 2021), PostEra, and published literature (Qin *et al.*, 2022, Ghahremanpour *et al.*, 2020, Narayanan et al., 2022). In particular, we included scaffolds which have shown promising antiviral or inhibitory activity against Mpro experimentally. Ultimately, we generated a broad compound library consisting of both covalent and non-covalent inhibitors. For each scaffold, our 3D-scaffold model generated between 500-4000 molecules not only sharing fingerprint similarity with the training set but also constraining the properties with respect to the input scaffolds. The generated molecules were then checked for validity, uniqueness, and novelty as described in Joshi et al.

Ligand-based Compound Screening

Ligand-based screening techniques were applied to the 3D-Scaffold generated compounds such that the screened compounds contain a high probability of druglike characteristics. The initial screening of the compounds was done by computing basic properties of interest like similarity of the compounds with respect to the parent compound, synthetic accessibility (SA) score, and quantitative estimation of druglikeness (QED). Next, various physicochemical properties were considered including: (i) logP, the partition coefficient, which indicates the lipophilicity of the compound (lipophilic if the value is positive or hydrophilic if the value is negative) and measures its permeability; (ii) topological polar surface area (TPSA), which estimates polarity and is one of the important parameter to measure absorption and blood-brain barrier permeability of the compounds; (iii) molecular weight (MW), selecting a range between 150-500 Da; and (iv) toxicity prediction. In total, 58 such properties were used for screening. This reduced the number of compounds under consideration to fewer than 500 to proceed to the next stage of the pipeline: molecular docking simulations.

Structure-based Compound Screening

We utilized molecular docking simulations for structure-based compound screening. Docking helps not only in understanding the key interactions between the target protein and the screened lead compounds, but also how these compounds bind in the binding pocket of the target protein. We used a homo-dimeric Mpro 3D structure as the target protein, with a binding site near the surface of each monomer. The binding site is further categorized into subsites (S1 to S4) containing a catalytic dyad composed of a cysteine and histidine pair (Cys145 and His41). We used our Automated Modeling Engine for Covalent Docking (AME-CoV) using AutoDockFR for covalent docking and our Automated Modeling Engine for non-covalent docking (AME-Non_CoV) using AutoDock Qvina02. A combination of docking scores and visual inspection of docked poses were used to further filter the set of candidates to proceed to lead optimization.

Lead Optimization

A significant contribution of this work is the code developed in our lead optimization efforts. While ligand- and structure-based screening such as that outlined in the previous sections has established utility, both suffer from higher failure rates than desired. As such, the development of computational lead optimization techniques with the ability to mitigate error have the potential for high impact. A key contribution of this work is the three LO methodologies we employed, and the models thereby obtained. The methods developed include: i) multi-parameter optimization, whereby weights are applied to linear combinations of chemical properties to achieve a model with reliable rankings of compounds; ii) 3D-quantitative structure-activity relationship (3D-QSAR) analysis; and iii) parallel graph neural network for binding affinity prediction. The curated Mpro library obtained as an early step was further utilized to train models using each of these methods. For 3D-QSAR model development, five ML-based models were trained and assessed for performance using multiple metrics. Decoy molecules were also obtained to serve as negative controls during training and assessment. Following training, filtered candidate molecules were submitted to each model and assessed for their predicted ability to target Mpro. Toxicity models were also utilized as a critical part of our LO process. The LO results helped inform which candidates to submit for experimental validation.

Experimental Validation

Native MS Experiments: We experimentally tested top 8 screen hits (Table 1) using native mass spectrometry to examine if the designed compounds form stable complexes with the target protein. After mixing 4 μM M^{pro} with each of the compounds at 20 μM, each sample was then subjected to electrospray MS detection under native conditions. From the compounds tested, compounds MCULE-4926166920 and MCULE-7471308738 showed best binding as we can see clear mass shifts to the dimer of the protein. Dimers of M^{pro} showed binding to one molecule in both MCULE-4926166920 and MCULE-7471308738. Based on the peak areas of the apo and holo species, we estimated the Kd to be in the hundreds of μM range.

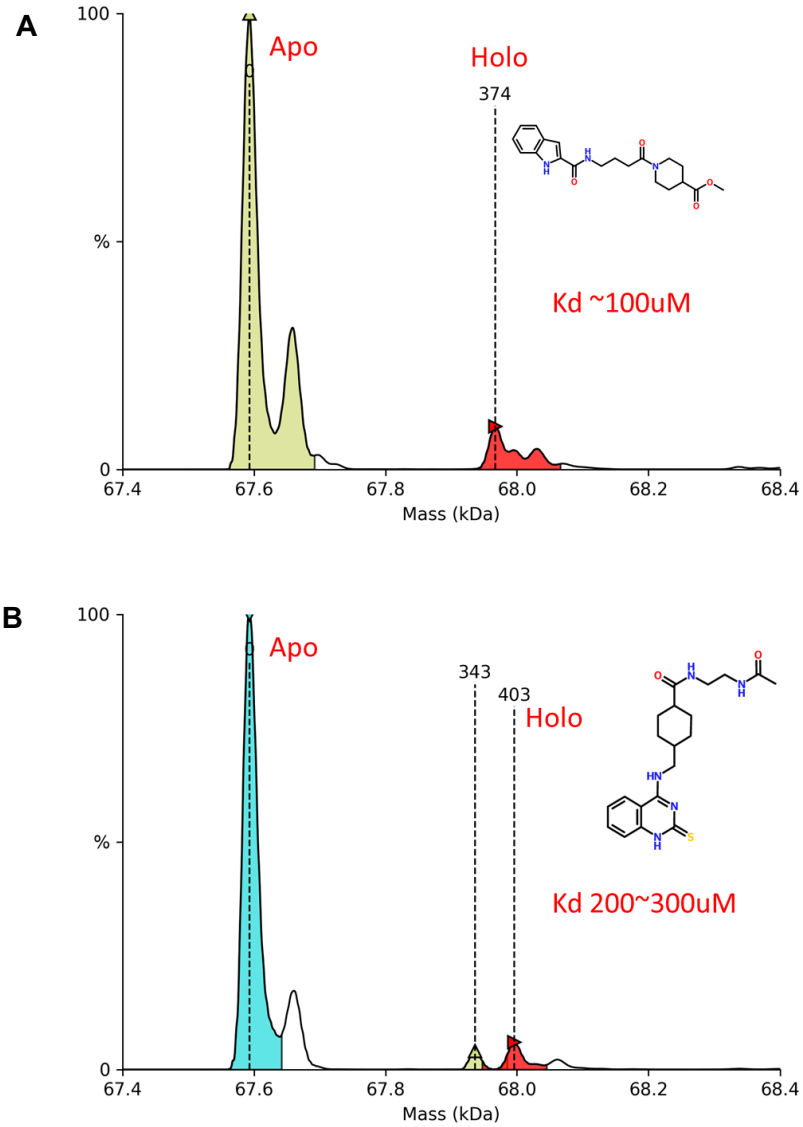

Figure 2. Native MS of Mpro mixed with compound A) MCULE-4926166920 and B) MCULE-7471308738 showed binding of one molecule to the dimer of the protein.

Functional Assay Experiments: We experimentally verified that our screen hits were capable of inhibiting M^{pro} activity using a plate reader-based biochemical assay with purified M^{pro} enzyme and a commercially available fluorogenic FRET peptide substrate. From the 8 compounds tested (Table 1), MCULE-7471308738 showed significant inhibition at higher concentration (62.5 µM) after incubating for 60 minutes. Further experiments will be conducted at higher concentrations to determine the actual IC_{50} of the compounds.

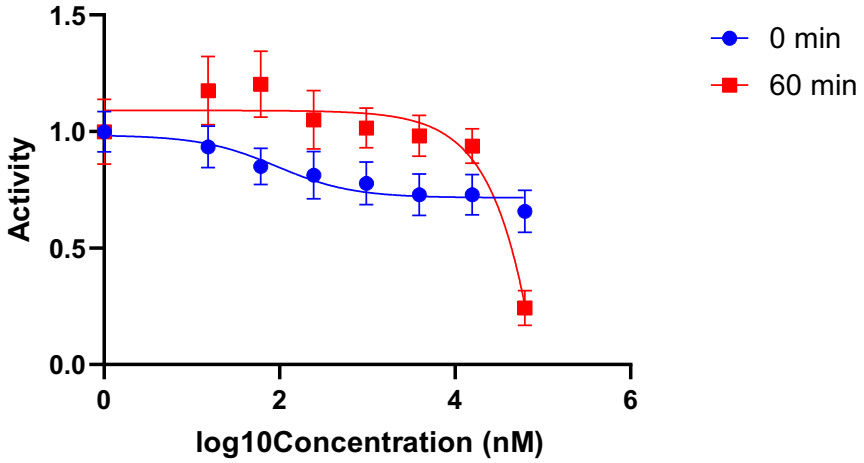

Figure 3: IC50 curves for compound MCULE-7471308738 after 0 minute and 60 minutes of co-incubation of enzyme and inhibitor.

References

1. Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British journal of pharmacology, 162(6), 1239-1249. DOI: 10.1111/j.1476-5381.2010.01127.x

2. Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it?. Acta Pharmaceutica Sinica B. DOI: https://doi.org/10.1016/j.apsb.2022.02.002

3. Shivanyuk, A. N., Ryabukhin, S. V., Tolmachev, A., Bogolyubsky, A. V., Mykytenko, D. M., Chupryna, A. A., ... & Kostyuk, A. N. (2007). Enamine real database: Making chemical diversity real. Chemistry today, 25(6), 58-59.

4. Kiss, R., Sandor, M., & Szalai, F. A. (2012). http://Mcule. com: a public web service for drug discovery. Journal of cheminformatics, 4(1), 1-1. DOI: https://doi.org/10.1186/1758-2946-4-S1-P17

5. Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., ... & Leach, A. R. (2017). The ChEMBL database in 2017. Nucleic acids research, 45(D1), D945-D954. DOI: https://doi.org/10.1093/nar/gkw1074

6. Joshi, R. P.; Gebauer, N. W. A.; Bontha, M.; Khazaieli, M.; James, R. M.; Brown, J. B.; Kumar, N (2021). "3D-Scaffold: A Deep Learning Framework to Generate 3D Coordinates of Drug-like Molecules with Desired Scaffolds." Journal of Physical Chemistry B **125**: 12166– 12176. DOI: 10.1021/acs.jpcb.1c06437

7. Kneller, D. W., Phillips, G., O'Neill, H. M., Jedrzejczak, R., Stols, L., Langan, P., ... & Kovalevsky, A. (2020). Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. Nature communications, 11(1), 1-6. DOI: https://doi.org/10.1038/s41467-020-16954-7

8. Clyde, A., Galanie, S., Kneller, D. W., Ma, H., Babuji, Y., Blaiszik, B., ... & Stevens, R. (2021). High-throughput virtual screening and validation of a sars-cov-2 main protease noncovalent inhibitor. Journal of chemical information and modeling, 62(1), 116-128. DOI: https://doi.org/10.1021/acs.jcim.1c00851

9. Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., ... & Zhuravleva, M. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic acids research, 49(D1), D437-D451. DOI: https://doi.org/10.1093/nar/gkaa1038

10. Qin, B., Craven, G. B., Hou, P., Chesti, J., Lu, X., Child, E. S., ... & Cui, S. (2022). Acrylamide fragment inhibitors that induce unprecedented conformational distortions in enterovirus 71 3C and SARS-CoV-2 main protease. Acta Pharmaceutica Sinica B. DOI: https://doi.org/10.1016/j.apsb.2022.06.002

11. Ghahremanpour, M. M., Tirado-Rives, J., Deshmukh, M., Ippolito, J. A., Zhang, C. H., Cabeza de Vaca, I., ... & Jorgensen, W. L. (2020). Identification of 14 known drugs as inhibitors of the main protease of SARS-CoV-2. ACS medicinal chemistry letters, 11(12), 2526-2533. DOI: https://doi.org/10.1021/acsmedchemlett.0c00521

12. Narayanan, A., Narwal, M., Majowicz, S. A., Varricchio, C., Toner, S. A., Ballatore, C., ... & Jose, J. (2022). Identification of SARS-CoV-2 inhibitors targeting Mpro and PLpro using in-cell-protease assay. Communications biology, 5(1), 1-17. DOI: https://doi.org/10.1038/s42003-022-03090-9

Machine Learning-driven Molecular Design for Therapeutic Discovery

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov