Pacific Northwest
NATIONAL LABORATORY

PNNL-36621

Automated AI-driven Molecular Design for Therapeutic Discovery

September 2024

1. Rohith Anand Varikoti
2. Chathuri Kombala
3. Stephanie M Thibert
4. Deseree J Tennyson
5. Zhou Mowei
5. Agustin Kruel
6. 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
fox: (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-36621

Automated AI-driven Molecular Design for Therapeutic Discovery

September 2024

1. Rohith Anand Varikoti
2. Chathuri Kombala
3. Stephanie M Thibert
4. Deseree J Tennyson
5. Zhou Mowei
5. Agustin Kruel
6. Neeraj Kumar

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

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

In recent years, artificial intelligence and machine learning (AI/ML) approaches have revolutionized the process of designing new therapeutics, enabling scientists to rapidly respond to emerging threats from various pathogens. A prime example is the SARS-CoV-2 main protease, a key target for the development of antiviral inhibitors. In this study, we employed a novel, integrated approach that combines AI-driven iterative design of inhibitor candidates, screening based on physio-chemical properties and toxicity, physics-based computational modeling of protein-inhibitor interactions, and AI-assisted analysis of Native MS biophysical assay and characterization of designed candidates. Our deep learning 3D-scaffold model, which uses an input scaffold as a starting point, generated tens of thousands of compounds while preserving the key scaffold. To optimize these candidates, we calculated a comprehensive set of 136 descriptors, including both 2D and 3D molecular features, for compounds targeting the SARS-CoV-2 Main protease (Mpro) and a neurodegenerative disease-associated protein, cyclophilin (Cyp). The generated compounds were initially filtered based on their properties and then ranked according to their predicted binding affinity using our automated modeling and ML methods. Experimental validation of the Mpro candidates showing inhibitory activity demonstrates that our workflow can expedite the therapeutic discovery.

Summary

We developed 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 advance the therapeutic discovery.

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

Introduction

Artificial intelligence/machine learning (AI/ML) based drug discovery and development approaches have gained significant progress over the past few years. This technological progress can aid in reducing the cost and time for discovering novel small molecules with desired properties compared to conventional methods (Hughes et al., 2011). However, nearly 90% of these new therapeutic molecules fail in the later stages of drug discovery (Sun, D. et al., 2022). With the vast amount of structural, functional, and therapeutic data of the 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), this 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. Utilizing the AI-based models incorporated with several open-source in silico tools—including those developed at PNNL targeting various areas of drug design such as compound generation, high-throughput virtual screening, quantitative structure-activity relationship (QSAR) analysis, toxicity prediction, etc. (Duch, W. et al., 2007). We designed and experimentally validated several potential hits against different protein targets. We also utilized the above approaches to incorporate new applications like lead optimization and drug repurposing, where we modified the fragments of existing drugs to make them more potent or utilized available FDA-approved drugs for different target proteins.

We developed a closed-loop drug discovery and lead optimization (LO) workflow (Figure 1) utilizing various tools, one such PNNL-developed tool, 3D-Scaffold (Joshi et al., 2021) that utilizes deep learning with a fragment-based or functional group method, which generates molecules based on the input scaffold, retaining the key scaffold. A benefit of this approach is that scaffolds can be chosen from experimentally validated active compounds.

Results and Discussions

Computational modeling and AI/ML methods: We utilized our 3D-Scaffold model, high throughput virtual screening (HTVS) techniques, and advanced hit identification and optimization methods in our computational workflow (Figure 1). We did the extensive literature search and key fragments were extracted from experimentally validated potential ligands found in the cyclophilin(s) PDB crystal structures (Kajitani, K., et al. 2007, Mikol, V., et al. 1994). For Mpro, we used the scaffolds from the top 4 compounds from our initial iteration of compounds (Varikoti, R. A., et al. 2023), which were then inputted into our 3D-scaffold model to generate a library of compounds covering extensive chemical space. The generated novel compounds were screened and sorted based on similarity patterns with their parent compounds, as well as on cheminformatics, physiochemical properties, and toxicity. The compounds were ranked based on the interpreted results and using molecular docking simulations to predict binding affinity. Further screening was done by visually inspecting the binding orientations and observing key interactions of the compounds with the target proteins. Additionally, we performed LO using 3D-QSAR and MPO analysis to obtain potent compounds with target-specific properties. The screened hits were further optimized before testing them using experimental validation (Table 1). We searched for the compounds from various vendors like Mcule, Enamine, etc., 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, Joshi, R. P. et al., 2023, Varikoti, R. A., et al., 2023). The capabilities and insights developed with this project will be ultimately applicable to a wide range of protein targets and biological systems of interest.

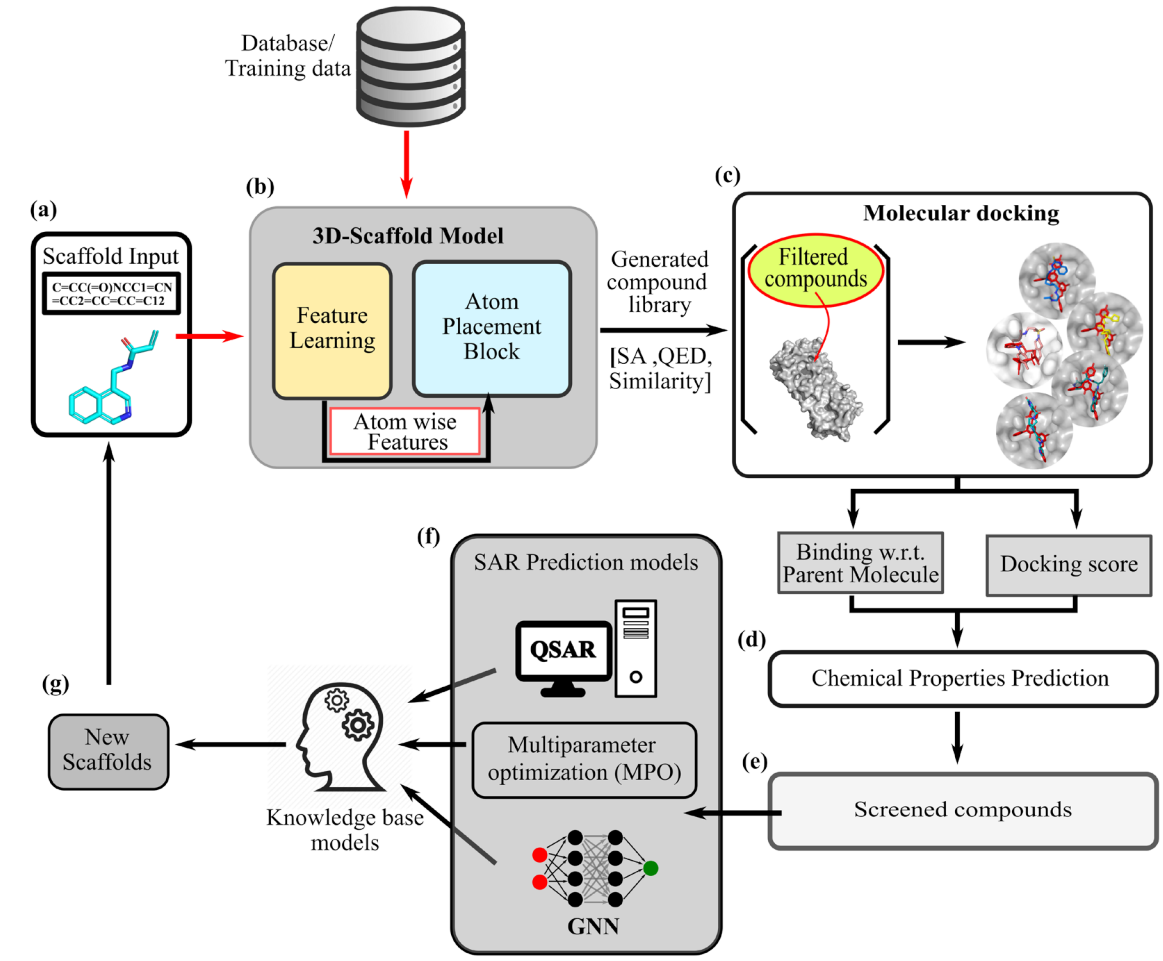


Figure 1. Our research follows a procedure for identifying therapeutic candidates and optimizing leads. (a) The process begins by inputting a scaffold (*) into (b) our 3D-scaffold model, which generates several ligands. (c) High Throughput Virtual Screening (HTVS) uses molecular docking and QSAR to identify lead compounds. (d) The generated compounds are screened based on various physiochemical properties to (e) identify hits. (f) Machine Learning/Deep Learning is used for activity prediction of lead compounds. (g) Fragments/scaffolds from the final compounds are then input into the 3D-scaffold model for lead optimization

Table 1. Top high throughput virtually screened compounds targeting proteins with their respective molecular properties that we finalized based on extensive computational studies.

Target Protein	#	Compound Name	MW	LP	TPSA	HA	HD	Docking Score	Synthetic Accessibility
SARS-CoV-2 Mpro	1	Z4887119528	501.551	4.8689	100.21	8	2	17.0599	3.87
	2	Z4509080715	471.445	5.2496	62.3	5	1	13.29	3.53
	3	Z4605133164	465.944	5.1882	62.3	5	1	13.668	3.64
	4	Z4509080683	505.889	5.903	62.3	5	1	13.7844	3.62
	5	Z4912275806	459.554	5.5283	62.3	5	1	15.5122	3.85
Cyclophilins (D and A)	6	B54_1257	663.576	-2.67694	288.67	12	7	-9.7 (-8.9)	5.55
	7	B54_1842	651.609	-3.20673	269.43	11	9	-9.3 (-9.5)	5.33
	8	B54_1126	664.584	-2.77533	286.17	12	8	-9.1 (-9)	5.26
	9	B54_1929	650.601	-2.96824	269.43	11	9	-9.1 (-8.7)	5.43
	10	B54_939	665.592	-2.93471	291.4	13	8	-9.1 (-8.7)	5.65

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); Cyp D (A): Cyclophilin D (Cyclophilin A)

Compound Library Generation

Utilizing 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. The input scaffolds were selected from a curated library of experimentally validated potent drug candidates with 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) targeting protein of interest. For generating compounds targeting cyclophilin, we used core fragments from a well-studied drug cyclosporin (CsA) and two experimentally validated compounds. For Mpro we used the fragments from our previous iteration of compounds which were experimentally tested to be active. Finally, we generated a broad compound library consisting of non-covalent inhibitors targeting Cyp(s) and both covalent and non-covalent inhibitors targeting Mpro. For each scaffold, our 3D-scaffold model generated between 500-4000 molecules not only sharing fingerprint similarity with the parent compounds set but also constraining the properties. 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 used to screen the 3D-scaffold generated compounds for druglike characteristics. The initial screening involved computing properties such as similarity to the parent compound, synthetic accessibility (SA) score, and quantitative estimation of druglikeness (QED). As a next step, 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. A total of 58 properties were used for screening, resulting in fewer than 500 compounds being considered for the next stage: molecular docking simulations.

Experimental Validation

Automated Native MS Experiments and Analysis: A key contribution of this work is the development of a novel AI/ML frameworks to automated mass-spectra analysis which allowed for high throughput binary classification to identify target compounds that bound to the active site of Mpro taking in m/z intensity pairings for peak identification. Two models were designed based off the datatype chosen for use. If similar runs with Ammonium Acetate are used, the model takes a vector of relative intensities from the runs, between 4000 & 4400 m/z, however an alternate structure was also designed for if there are frameshifts in m/z exceeding 200 m/z, as seen with the EDTA/TCEP samples. This model was trained on 40 test sets including 20 (+) bound samples of MPRO + the Pfizer compound as well as 20 (-) samples consisting of only MPRO. From the final Mpro compounds (**Table 1**) Z4887119528 showed no affinity, Z4605133164 showed low affinity, and the remaining compounds each displayed significant populations of both singly and doubly bound ligand with moderate to high affinity (**Figure 2**). The lack of observed binding for compound Z4887119528 during Native MS screening may be the result of potential non-covalent interactions that do not survive in the gas phase under MS conditions.

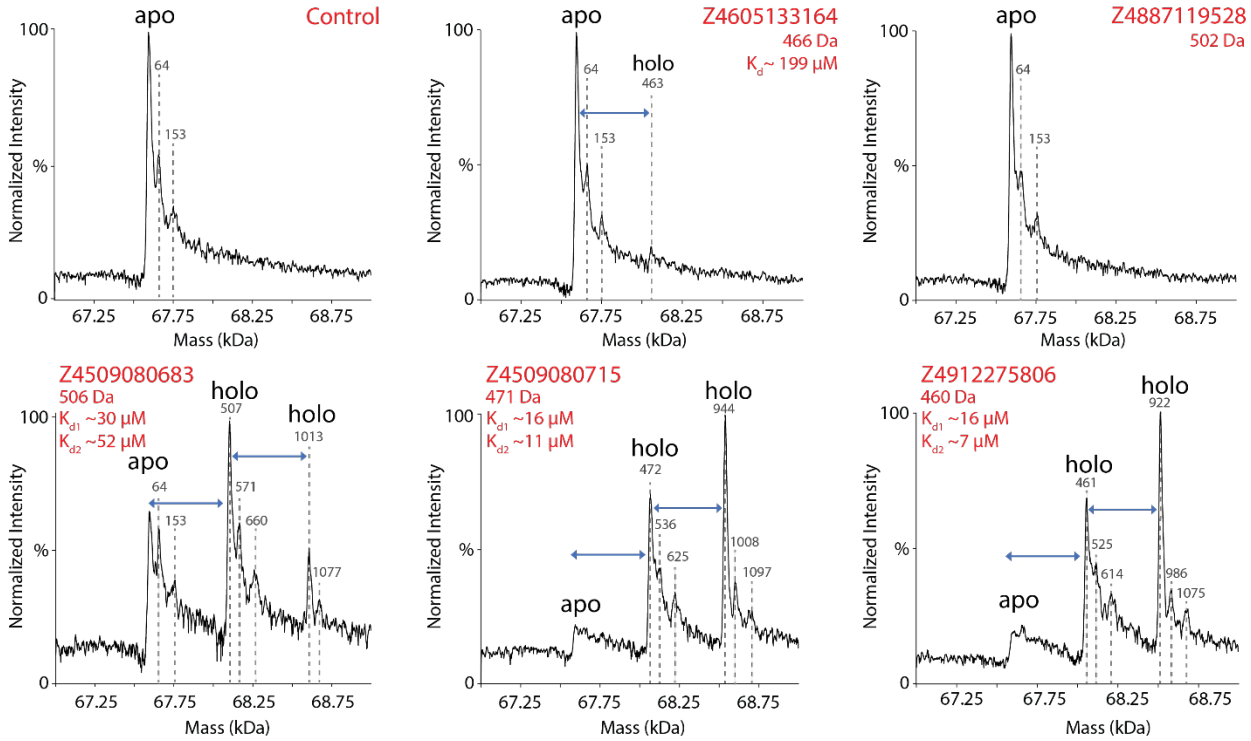
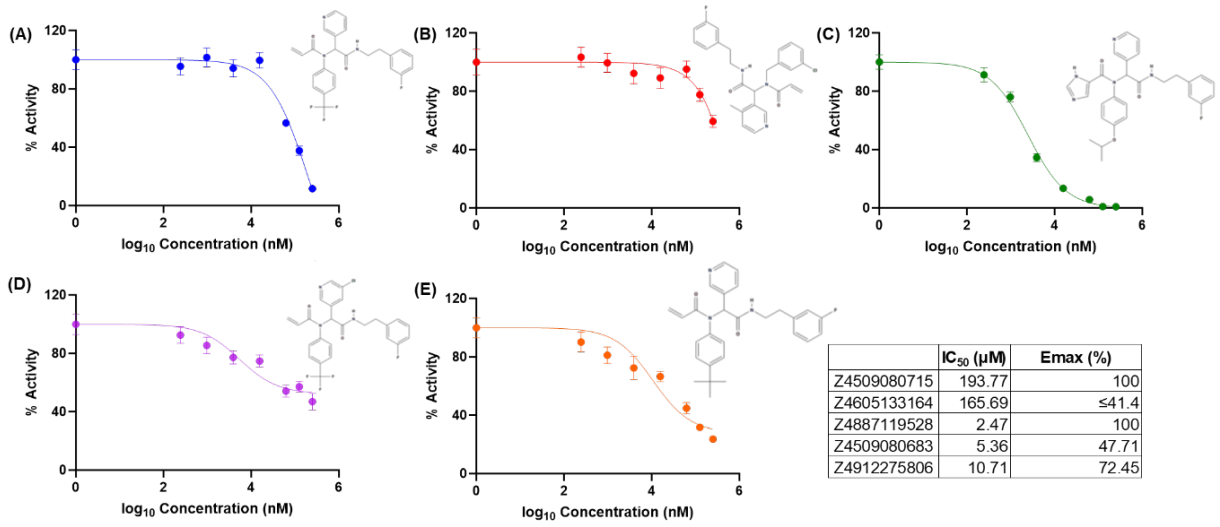


Figure 2. Deconvolved mass spectra of M^{pro} control and M^{pro} with each of the five compounds from the activity assay. The numbers above each dashed line are the mass shift in Da from the apo M^{pro} dimer mass. Note that some adducts are seen, even in the control sample. The blue arrows indicate a mass shift corresponding to the mass of the compound. The K_d values were based on the relative abundance of the apo and holo species. Spectra for M^{pro} with compounds Z4509080683, Z4509080715, and Z4912275806 each display two holo peaks, indicating singly and doubly bound species

A key contribution of this work is the development of a novel AI/ML frameworks to automated mass-spectra analysis which allowed for high throughput binary classification to identify target compounds that bound to the active site of Mpro taking in m/z intensity pairings for peak identification. Two models were designed based off the datatype chosen for use. If similar runs with Ammonium Acetate are used, the model takes a vector of relative intensities from the runs, between 4000 & 4400 m/z, however an alternate structure was also designed for if there are frameshifts in m/z exceeding 200 m/z, as seen with the EDTA/TCEP samples. This model was trained on 40 test sets including 20 (+) bound samples of MPRO + the Pfizer compound as well as 20 (-) samples consisting of only MPRO. From the final Mpro compounds (**Table 1**) Z4887119528 showed no affinity, Z4605133164 showed low affinity, and the remaining compounds each displayed significant populations of both singly and doubly bound ligand with moderate to high affinity (**Figure 2**). The lack of observed binding for compound Z4887119528 during Native MS screening may be the result of potential non-covalent interactions that do not survive in the gas phase under MS conditions.

Functional potency of inhibitor compounds:

We experimentally assessed the inhibition of Mpro enzyme activity by the computational leads using a well-established FRET based biochemical assay. Of the five compounds tested compound Z4887119528 showed highest inhibitory activity (IC_{50} =2.47 μM) compared with other candidates. Compounds Z4509080715, Z4605133164, Z4509080683, and Z4912275806 have covalent acrylamide warhead in common whereas compound Z4887119528 lacks a strong covalent warhead. Thus, compound Z4887119528 could act as non-covalent inhibitor. Further, native MS results showed no binding of compound Z4887119528 (**Figure 3**) with Mpro enzyme. It is possible that some non-covalent interactions such as hydrophobic interactions become weaker in gas phase under MS conditions (Bich, C., *et al.* 2010) and impossible to detect certain ligand-protein assemblies with native MS (Boeri Erba, E., & Petosa, C., 2015).



	IC_{50} (μM)	Emax (%)
Z4509080715	193.77	100
Z4605133164	165.69	≤41.4
Z4887119528	2.47	100
Z4509080683	5.36	47.71
Z4912275806	10.71	72.45

Figure 3 Inhibition screening of computational leads using FRET assay. IC50 curves with representative structures of commercially available screened hit compounds A) Z4509080715 B) Z4605133164 C) Z4887119528 D) Z4509080683 and E) Z4912275806. Enzyme reactions were carried out incubating purified Mpro enzyme with increasing concentrations (0 to 250 μM) of screen hit compounds. Enzyme activity was determined by measuring fluorescence after adding Dabcyl-KTSAVLQSGFRKME-EDANS peptide substrate. Initial reaction rates were used to determine the IC50 values. IC50 and Emax (Maximum possible inhibition) values are reported in the table.

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. Duch, W., Swaminathan, K., & Meller, J. (2007). Artificial intelligence approaches for rational drug design and discovery. *Current pharmaceutical design*, *13*(14), 1497-1508.

7. 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.

8. 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.

9. Stauffer, W. T., Goodman, A. Z., & Gallay, P. A. (2024). Cyclophilin inhibition as a strategy for the treatment of human disease. Frontiers in Pharmacology, 15, 1417945.

10. Kajitani, K., Fujihashi, M., Kobayashi, Y., Shimizu, S., Tsujimoto, Y., & Miki, K. (2008). Crystal structure of human cyclophilin D in complex with its inhibitor, cyclosporin A at 0.96-Å resolution. *Proteins: Structure, Function, and Bioinformatics*, *70*(4), 1635-1639.

11. Mikol, V., Kallen, J., & Walkinshaw, M. D. (1994). X-ray structure of a cyclophilin B/cyclosporin complex: comparison with cyclophilin A and delineation of its calcineurin-binding domain. *Proceedings of the National Academy of Sciences*, *91*(11), 5183-5186.

12. Varikoti, R. A., Schultz, K. J., Kombala, C. J., Kruel, A., Brandvold, K. R., Zhou, M., & Kumar, N. (2023). Integrated data-driven and experimental approaches to accelerate lead optimization targeting SARS-CoV-2 main protease. *Journal of Computer-Aided Molecular Design*, *37*(8), 339-355.

13. 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.

14. Joshi, R. P., Schultz, K. J., Wilson, J. W., Kruel, A., Varikoti, R. A., Kombala, C. J., ... & Kumar, N. (2023). Ai-accelerated design of targeted covalent inhibitors for SARS-CoV-2. *Journal of Chemical Information and Modeling*, *63*(5), 1438-1453.

15. 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.

16. 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.

17. 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.

18. 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.

19. Bich, C., Baer, S., Jecklin, M. C., & Zenobi, R. (2010). Probing the hydrophobic effect of noncovalent complexes by mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *21*(2), 286-289.

20. Boeri Erba, E., & Petosa, C. (2015). The emerging role of native mass spectrometry in characterizing the structure and dynamics of macromolecular complexes. *Protein Science*, *24*(8), 1176-119.

Pacific Northwest National Laboratory

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

1-888-375-PNNL (7665)

www.pnnl.gov