

Team name	Northeastern University Warriors of the Anti-Viral Enterprise (NUWAVE)
Team member(s) (firstname lastname; ...)	Mary Jo Ondrechen (PI), Suhasini Iyengar , Kelton Barnsley, Yen Vu, Penny Beuning, Ian Bongalonta, Alyssa Herrod, Jasmine Scott
Affiliation	Northeastern University, USA
Contact email	lyengar.s@northeastern.edu ; mjo@neu.edu
Contact phone number (optional)	+1-857-415-9653 (Iyengar) +1-508-740-9513 (Ondrechen)
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nsp x , Orf Xx , N, E, etc...) 3 required	N Protein, Main Protease (monomer) , Main Protease (Dimer) , RNA Methyltransferase, NSP1, NSP3 (this report talks about the Main protease)

Section 1: Methods:

Part A- Binding site prediction:

For the binding site prediction, Partial Order Optimum Likelihood (POOL) (1,2) was used. Partial Order Optimum Likelihood (POOL) (1,2) is a machine learning method that predicts biochemically active sites using the three-dimensional structure of the query protein as input. POOL predicts multiple types of binding sites in proteins which include catalytic sites, allosteric sites and other sites, some of which may not be detected by other predictive methods. POOL generates a rank ordered list of all the amino acids in the protein structure in the order of likelihood of biochemical activity. POOL predicts some sites that might be overlooked by other methods because POOL is based primarily on computed electrostatic and chemical properties (3,4) of the query protein, rather than a purely informatics-based approach. POOL points to the residues involved in reversible binding, including catalytic sites, non-catalytic binding sites such as allosteric sites, ligand transport sites, and some protein-protein interaction sites. The other input features for POOL consist of properties of the local environment (1,2) and surface topological metrics (5).

Part B- Molecular Docking:

Molecular Docking was performed using Schrödinger Glide (6). For docking in Schrödinger Glide, the ligands were prepared using LigPrep (7), the protein was minimized and optimized using the Protein Preparation Wizard and the grid for docking was prepared using Receptor Grid Generation using the top 10 % of the POOL predicted residues as the centroid for ligand placement in Schrödinger 2019-3. Molecular Docking was performed on the Discovery Cluster at the Massachusetts Green High-Performance Computing Center using Glide. Glide Standard Precision (SP) (8) was used as a filter to remove false positive results and top predicted ligands with docking score of ≤ -7 kcal/mol were used for Glide Extra Precision (XP) (9).

Section 2: Targets

Target 2: Main Protease (Monomer and Dimer)

The target protein was downloaded from the protein data bank. Following structures of the main protease were used: 6LU7- The crystal structure of COVID-19 main protease in complex with an inhibitor N3 (Resolution 2.16 Å) (10), 5R82- PanDDA analysis group deposition -- Crystal Structure of COVID-19 main protease in complex with Z219104216

(resolution-1.31 Å)(11), 5R7Z- PanDDA analysis group deposition -- Crystal Structure of SARS-CoV-2 main protease in complex with Z1220452176 (12) and 6W63-Structure of COVID-19 main protease bound to potent broad-spectrum non-covalent inhibitor X77 (13). Before running POOL on these structures, they were analyzed in YASARA (14) and pKa prediction and energy minimization using the YAMBER3 force field were done for these structures. They were further prepared before docking using the Protein Preparation Wizard in Maestro. The protein preparation wizard allows the user to take the protein structure in its raw state, which might be missing hydrogen atoms and have incorrect bond orders, and convert it into a state which is properly prepared for use by Schrödinger products such as Glide (6). Protein Preparation step in Maestro contains three basic steps: first is preprocessing the protein structure. This step performs the basic calculations for assigning bond orders, adding hydrogens, creating disulfide bonds, filling missing side chain or missing loops, and deleting water molecules as needed. The second step is protein refinement. This step consists of optimization of the hydrogen bond network by reorienting the hydroxyl and thiol groups, water molecules, amide groups of asparagine (Asn) and glutamine (Gln), and the imidazole ring in histidine (His); and predicting protonation states of histidine, aspartic acid (Asp) and glutamic acid (Glu) and tautomeric states of histidine. The last step is Restrained minimization which provides controls for optimizing the corrected structure, to relieve any strain and fine-tune the placement of various groups.

Section 3: Libraries

The ligands were obtained from the following databases:

- a) ZINC FDA library (<https://zinc15.docking.org/substances/subsets/fda/>)
- b) CAS Antiviral set (<https://www.cas.org/covid-19-antiviral-compounds-dataset>)
- c) Enamine FDA library (<https://enamine.net/hit-finding/compound-collections/bioreference-compounds/fda-approved-drugs-collection>)
- d) Antiviral library consisting of compounds from- Selleck Chemicals Antiviral Library Enamine Antiviral Library and Asinex Antiviral Library

The ligands from all these libraries were prepared using the LigPrep tool in Maestro. Ligprep is a tool designed to prepare high quality all-atom 3D structures for large numbers of drug-like molecules. The LigPrep process consists of a series of steps that perform conversions, apply corrections to the structures, generate variations in the structure, eliminate unwanted structures and optimize all the structures.

Section 4: Results

Section A: Prediction of binding sites by POOL:

POOL generates a rank-ordered list of all the amino acids in a protein structure, in order of likelihood of biochemical activity.

- a) MONOMER- PDBID: 6LU7- The crystal structure of COVID-19 main protease in complex with an inhibitor N3

The POOL predicted sites for the main protease are as follows.

27LEU 38CYS 39PRO 40ARG **41HIS** 42VAL 54TYR 85CYS 105ARG 110GLN 128CYS 140PHE 141LEU 142ASN 143GLY 144SER **145CYS** 160CYS 161TYR 163HIS 164HIS 165MET 166GLU 172HIS 173ALA 181PHE 182TYR 187ASP 188ARG 189GLN

Residues in Red : Previously known catalytic residues

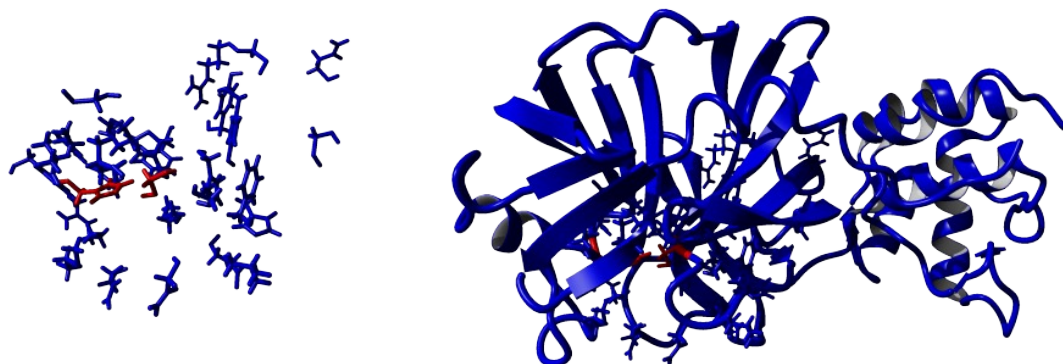


Figure 1: Image showing the POOL predicted residues in blue and catalytic residues in red for 6LU7

b) Monomer- PDBID-5R82: Crystal Structure of COVID-19 main protease in complex with Z219104216

POOL predicted residues are as follows :

25THR 27LEU 38CYS 39PRO 40ARG **41HIS** 42VAL 44 CYS 54TYR 85CYS 110GLN
111THR 128CYS 141LEU 142ASN 143GLY 144SER **145CYS** 161TYR 163HIS 164HIS
165MET 166GLU 167LEU 172HIS 173ALA 186VAL 187ASP 188ARG 189GLN 295ASP
Residues in Red : Catalytic residues

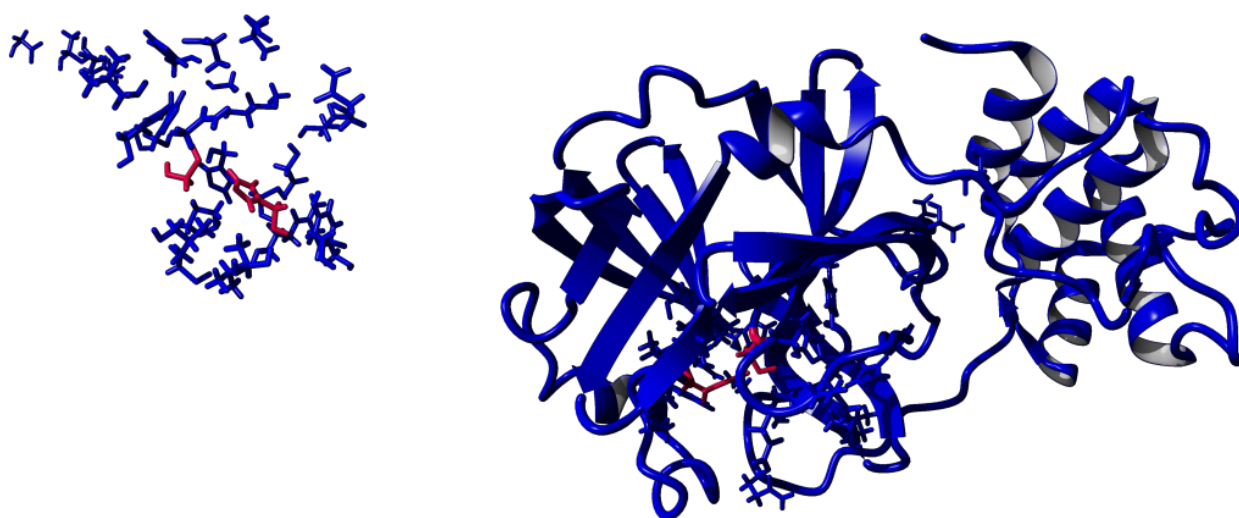


Figure 2: Image showing the POOL predicted residues in blue and active site residues in red for 5R82

c) 6W63 Monomer- Structure of COVID-19 main protease bound to potent broad-spectrum non-covalent inhibitor X77

POOL-Predicted Residues (top 10%):

CYS145 HIS164 HIS41 HIS163 ASP187 TYR161 MET165 HIS172 TYR54 GLU 166
ASN142 CYS38 CYS44 GLY143 TYR182 CYS160 CYS128 GLN110 LEU141 VAL186
ARG40 ARG105 PHE181 SER144 PRO39 ARG188 TYR126 CYS85 ALA173 THR111
GLY174

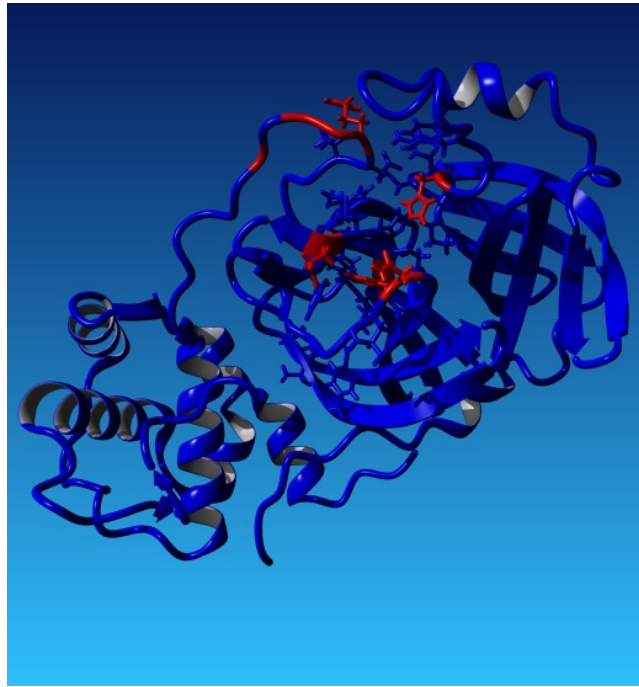


Figure 3: Image showing the POOL predicted residues in blue and active site residues in red for 6W63

d) Dimer: PDBID- 6LU7- The crystal structure of COVID-19 main protease in complex with an inhibitor N3

POOL Predicted Residues FOR BOTH THE CHAINS: (top 10%) (NOTE: Since this structure was a dimer only THEMATICS was used as input for POOL)

1SER 4ARG 5LYS 6MET 7ALA 111THR 112PHE 113SER 114VAL 117 CYS
 118TYR 124 GLY 125VAL 126TYR 127GLN 128CYS 129ALA 137LYS 138GLY
145CYS 160CYS 161TYR 163HIS 164HIS **166GLU** 127HIS 288GLU 289ASP
290GLU [yellow highlighted residues are important for dimerization]

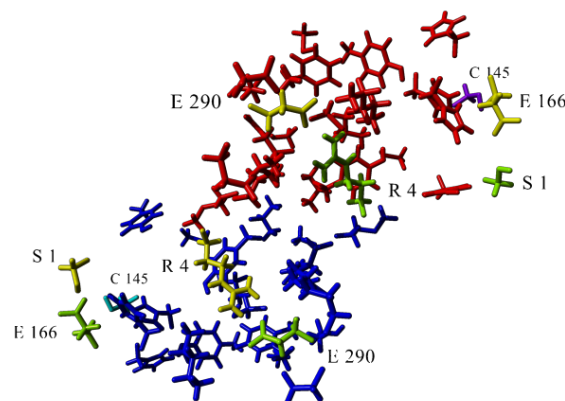


Figure 4: Image showing the POOL predicted residues for 6LU7 dimer, A chain in blue and B chain in red. The residues important for dimerization are colored in lemon green for chain A and yellow for chain B. The active site residue is colored in cyan for Chain A and magenta for B.

Section B: Molecular Docking:

Glide SP docking was performed on the entire library and the top hits from Glide SP were given as input to Glide XP. The results tabulated below are from Glide XP.

a) Main protease monomer- 6LU7:

Top hits from ZINC:

ZINC ID	Generic Name	Docking Score	XP Gscore	Interactions (BOLD-POOL H-bond, PI-PI , salt bridge)
ZINC000008214418	Ioxilan	-9.878	-9.878	His164, Leu141, Thr190, Glu166
ZINC000085540215	Ioxilan	-9.878	-9.878	His164, Leu141, Thr190, Glu166
ZINC000085540219	Ioxilan	-9.878	-9.878	His164, Leu141, Thr190, Glu166
ZINC000085540223	Ioxilan	-9.878	-9.878	His164, Leu141, Thr190, Glu166
ZINC000002036848	Riboflavin	-8.840	-8.841	Glu166, Thr190, Gln192, Gly143, Leu141

b) Main Protease monomer- 5R82:

Top hits from ZINC:

ZINC ID	Generic Name	Docking Score	XP Gscore	Interactions (BOLD H-bond, PI-PI , salt bridge)
ZINC000028232750	Valrubicin	-9.754	-9.754	Glu166, Gln189, Gly143, Asn142
ZINC000049783788	Valrubicin	-8.408	-8.409	Glu166, Ser144, Cys145, Asn142

Hits from ENAMINE Database:

Enamine ID	Generic Name	Docking score	XP Gscore	Interactions (BOLD H-bond, PI-PI , salt bridge)
Z1563146136	Acarbose	-10.653	-11	His164, Gln189, Asn142, Cys44, Glu166, Glu166

Hits from CAS Database:

CAS #	Docking score	XP Gscore	Interactions (BOLD H-bond, PI-PI , salt bridge)
885109-28-4	-9.823	-9.995	Gln189, Asn142, Gly143, Glu166, Thr26
1002334-85-1	-9.492	-9.492	Glu166, Arg188, Gln189, Asn142
65128-60-1	-9.4	-9.4	Thr26, His164, Gly143, Arg188, Thr190, Asn142
2133374-44-2	-9.357	-9.357	Glu166, Gly143, Thr190
956031-61-1	-9.338	-9.529	Glu166, Ser144, Phe140

c) Main Protease Monomer-6W63:

ZINC ID	Generic Name	Docking Score	XP GScore	Interactions (BOLD-POOL , H-bond, PI-PI , salt bridge)
ZINC000085540215	Ioxilan	-10.966	-10.966	HIS163, GLU166, THR190, GLN189
ZINC000043207238	Canagliflozin	-10.134	-10.134	GLU166, GLN192, THR190
ZINC000003830947	Isovue-M	-9.695	-9.695	ASN142, GLN192, THR190, GLN189, GLU166
ZINC000003830957	Iopromide	-9.514	-9.514	HIS163, GLU166, ARG188

ZINC000003830958	Iopromide	-9.430	-9.430	HIS41, ASN142, HIS163, ARG188, GLU166
ZINC000011615926	Lipitor	-9.273	-9.276	GLY143, THR26
ZINC000002036848	Riboflavin	-9.058	-9.058	ASN142, LEU141, GLU166, GLN192

d) Main Protease dimer- 6LU7:

ZINC XP Results:

ZINC ID	Generic Name	Docking Score	XP Gscore	Interactions (BOLD-POOL H-bond, PI-PI, salt bridge, halogen bond, Pi-cation)
ZINC000085540223	loxilan	-9.875	-9.875	B chain- Gln127, Lys5, Glu288 , Phe3, Trp207 B chain- Arg4 , A chain- Lys5 Both A and B- Lys5
ZINC000008214418	loxilan	-9.875	-9.875	B chain- Gln127, Lys5, Glu288 , Phe3, Trp207 B chain- Arg4 , A chain- Lys5 Both A and B- Lys5
ZINC000085540215	loxilan	-9.875	-9.875	B chain- Gln127, Lys5, Glu288 , Phe3, Trp207 B chain- Arg4 , A chain- Lys5 Both A and B- Lys5
ZINC000085540219	loxilan	-9.875	-9.875	B chain- Gln127, Lys5, Glu288 , Phe3, Trp207 B chain- Arg4 , A chain- Lys5 Both A and B- Lys5
ZINC000028957444	Ticagrelor	-7.824	-8.355	B chain- Gln127 , Both A and B- Lys5
ZINC000002005305	5-methyltetrahydrofolic acid	-8.269	-8.285	A chain-Asn216, Asn214, Leu282, Lys5, Glu288 B chain- Leu288 , Trp207

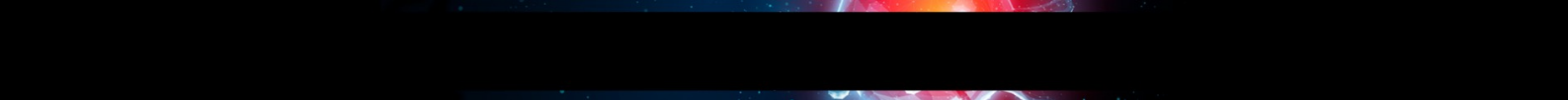
CAS XP results:

CAS #	Docking score	XP Gscore	Interactions (BOLD-POOL H-bond, PI-PI, salt bridge, halogen bond, Pi-cation)
766549-10-4	-11.811	-11.811	A chain- Leu282, Glu290 B chain- Phe3, Lys5 A- Lys5
479677-72-0	-11.357	-11.357	A chain: Lys5, Glu288 B chain: Phe3, Lys 5 , Trp207, Gly283, Gly138 , Gly170 B chain- Lys137
1383925-90-3	-10.881	-10.972	A chain- Phe3, Lys5 B chain- Glu288 A chain- Lys5

			B chain- Lys137 A and B chain- Lys 5
115561-47-2	-10.617	-10.617	A chain- Phe3, Leu282, Glu288 B chain- Phe3, Lys5 , Thr169 B chain- Lys5
926902-31-0	-10.561	-10.565	A chain- Lys5 , Thr169 B chain- Lys5 , Lys137 A and B chain- Lys5 A chain- Arg4

References:

1. Tong, W., Y. Wei, L.F. Murga, M.J. Ondrechen, and R.J. Williams, Partial Order Optimum Likelihood (POOL): Maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. PLoS Comp Biol, 2009. 5(1): e1000266.
2. Somarowthu, S., H. Yang, D.G. Hildebrand, and M.J. Ondrechen, High-performance prediction of functional residues in proteins with machine learning and computed input features. Biopolymers, 2011. 95(6): 390-400.
3. Ondrechen, M.J., J.G. Clifton, and D. Ringe, THEMATICS: A simple computational predictor of enzyme function from structure. Proc. Natl. Acad. Sci. (USA), 2001. 98: 12473-12478.
4. Ko, J., L.F. Murga, P. André, H. Yang, M.J. Ondrechen, R.J. Williams, A. Agunwamba, and D.E. Budil, Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. Proteins, 2005. 59(2): 183-195.
5. Capra, J.A., R.A. Laskowski, J.M. Thornton, M. Singh, and T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol, 2009. 5(12): e1000585.
6. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M.; Perry, J. K.; Francis, P.; Shenkin, P. S., "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," J. Med. Chem., 2004, 47, 1739–1749.
7. Schrödinger Release 2019-3: Glide, Schrödinger, LLC, New York, NY, 2019.
8. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening," J. Med. Chem., 2004, 47, 1750–1759
9. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes," J. Med. Chem., 2006, 49, 6177–6196.
10. 6LU7-Jin, Z., Du, X., Xu, Y. et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature 582, 289–293 (2020). <https://doi.org/10.1038/s41586-020-2223-y>
11. 5R82- Not yet published
12. 5R7Z- Not yet published
13. 6W63: Not yet published
14. Krieger, E., K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus, Improving physical realism, stereochemistry, and side-chain accuracy in



homology modeling: Four approaches that performed well in CASP8. Proteins 2009.
77 Suppl 9: 114-122.