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	Iyengar.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, Nspx, OrfXx, N, E, etc…) | 3 required	**RNA Methyltransferase (NSP16-NSP10),** N Protein, Main Protease (monomer), Main Protease (Dimer), NSP1, NSP3 (this report talks about the **RNA Methyltransferase**)

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 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 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 1: RNA Methyltransferase (NSP16/NSP10)

Target structure: 6W4H (10) - 1.80 Angstrom Resolution Crystal Structure of NSP16 - NSP10 Complex from SARS-CoV-2

The target protein was downloaded from the protein data bank (11). Before running POOL on this structure, it was analyzed on YASARA (12) and pKa prediction and energy minimization using YAMBER3 force field were done for this model. This structure was further prepared before docking using the Protein Preparation Wizard on Maestro. The protein preparation wizard allows the user to take the protein in its raw state-which might be missing hydrogen atom and have incorrect bond orders-and convert it into a state which is properly prepared for use by Schrodinger products such as Glide. Protein Preparation step on 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, deleting waters among many others whenever 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
e) ZINC Library library (https://zinc15.docking.org/)
f) Merck Library provided by Organizers of the JEDI challenge

The ligands from all these libraries were prepared using 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

Part A: POOL- Prediction of Binding Sites:

POOL generates a rank-ordered list of all the amino acids in a protein structure, in order of likelihood of biochemical activity. The POOL predicted sites for the NSP16/10 complex protein are as follows.

Chain A: 6809MET 6823CYS 6844LYS 6845TYR 6849CYS 6866ILE 6867HIS 6868PHE 6894VAL 6895ASP 6897ASP 6902VAL 6904ASP 6926ILE 6927SER 6928ASP 6930TYR 6935LYS 6968LYS 6971GLU 6979TYR 7001GLU 7023HIS 7026TYR 7027ILE 7029TRP 7030ARG 7033ASN
Chain B: 4294CYS 4324ALA 4325SER 4326CYS 4327CYS 4330CYS 4331ARG 4335ASP 4336HIS 4343CYS 4344ASP 4346LYS 4383CYS

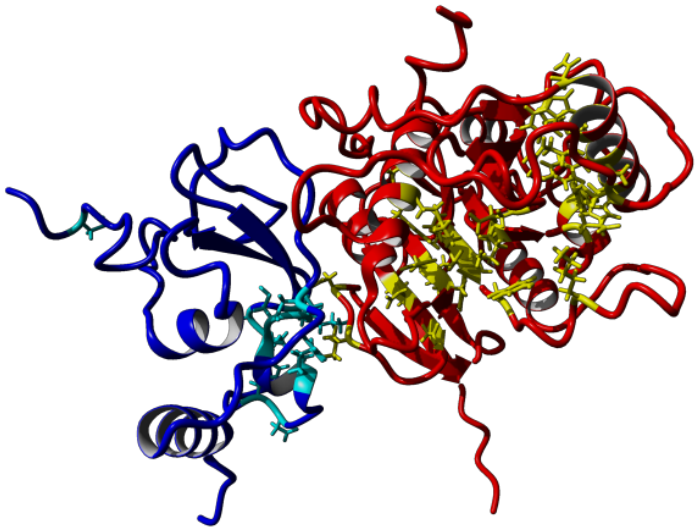

Figure 1: POOL Predicted residues shown in Yellow- Chain A and Cyan-Chain B with the secondary structure in Red and Blue for A and B respectively

Section B: Molecular Docking:

From the libraries using for testing the top hits were obtained from the CAS Antiviral library. 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.

cas.rn	docking score	XP GScore	Interactions **BOLD-POOL** H-bond, PI-PI, salt bridge, halogen bond, Pi-cation
435297-57-7	-14.883	-15.111	Gly6869, Gly6869, Gly6911, Cys6913, **Tyr6930**, **Lys6844**, **Lys6968**
435297-58-8	-14.759	-14.987	Gly6869, Gly6869, Gly6911,Ser6896, Cys6913, **Tyr6930**, **Lys6844**
1312805-81-4	-14.516	-14.516	Asn6996, Ser6999, **Lys6844**, Asn6841, **Lys6968**, Asp6873, Gly6871, **Tyr6930**, Gly6869, Cys6913, Gly6911,Ser6896, Leu6898
1002334-92-0	-14.417	-14.417	Asp6928, **Tyr6930**, Gly6871, Gly6869, Cys6913, Gly6911, Ser6896, Asn6899
192865-92-2	-14.076	-14.076	Asn6996, **Glu7001, Lys6844,** GLy6869, Cys6913, Gly6911, Ser6896, **Tyr6930**

Top Hits from ZINC FDA Library:

ZINC ID	Generic Name	Docking Score	XP Gscore	Interactions (**BOLD POOL** H-bond, PI-PI, salt bridge halogen bond)
ZINC000000518554	Arbutin	-9.89	-9.89	Asp6928, **Tyr6930**, **Phe6947**, Cys6913, Asp6912, Asn6899
ZINC000049783788	valrubicin	-9.699	-9.699	Leu6898, Cys6913, **Tyr6930**, Asp6931, Lys6933

ZINC000085537017	cangrelor	-9.301	-9.315	Asp6928, Lys6968, Lys6968, Asn6841, Ser6999, **Lys6844, Lys6844**, Asn6996
ZINC000000000856	formoterol	-8.713	-8.725	**Tyr6930**, Cys6913, Gly6911, Gly6869, **Asp6897**
ZINC000002599970	formoterol	-8.713	-8.725	**Tyr6930**, Cys6913, Gly6911, Gly6869, **Asp6897**

Top Hits from Enamine FDA database:

Generic Name	docking score	XP GScore	Interactions (**BOLD POOL** H-bond, PI-PI, salt bridge halogen bond)
Esculin	-9.48	-9.491	Leu6898, Ser6896, Gly6869, Asp6912
Pemetrexed disodium	-8.121	-8.121	**Lys6844,** Asn6841, Asp6873, Asp6912, Cys6913
Calcium leucovorin	-7.724	-7.724	**Tyr6930**, Asn6996, Ser6999, Asn6841, **Lys6844** **Lys6844**, Leu6898 Lys6968

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. 6W4H-Not yet published
11. H.M. Berman, K. Henrick, H. Nakamura (2003) Announcing the worldwide Protein Data Bank Nature Structural Biology 10 (12): 980.Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 1997, 25:3389-3402. (www.wwpdb.org)
12. 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.