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	**N Protein**, Main Protease (monomer), Main Protease (Dimer), RNA Methyltransferase, NSP1, NSP3 (this report talks about the **N Protein**)

Section 1: Methods:

Part A- Homology modeling of Nucleocapsid protein (N Protein):

The N Protein model structure was built in YASARA (1) using a series of structures from the Protein Data Bank (2). These structures were obtained after a BLAST search of the N Protein sequence. The model was built by manually providing template structures with sequence homology to N Protein. These templates are Crystal Structure of SARS-CoV-2 nucleocapsid protein N-terminal binding domain (PDB ID:6M3M) (3), Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS-CoV-2 (PDB ID: 6VYO) (4) 2.05 Angstrom resolution crystal structure of C-terminal dimerization domain of Nucleocapsid Phosphoprotein from SARS-CoV-2 (PDBID: 6WJI) (5) and the N-Terminal binding domain of the SARS-CoV-2 nucleocapsid phosphoprotein (PDBID: 6YI3)(6). Using these three structures as templates, YASARA built a hybrid model for the N protein. Figure 1 shows the hybrid model generated for N Protein. The final model generated consists of only the RNA binding domain for the N Protein (highlighted in light blue).

```
MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQHG 60
KEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAG 120
LPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQGTTLPKGFYAEGSRGGS 180
QASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAALALLLLDRLNQLESKMSGKGQQ 240
QQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKH 300
WPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAY 360
KTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA  419
```

This model was further validated using the ERRAT and VERIFY 3D servers which are a part of Structural Analysis Verification Server (SAVES) (7) and SWISS model (8).

NOTE: The model structure is also included with the files uploaded as a .zip folder.

Part B- Binding site prediction:

For the binding site prediction, Partial Order Optimum Likelihood (POOL) (9, 10) was used. Partial Order Optimum Likelihood (POOL) 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 (11,12) 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 (9,10) and surface topological metrics (13).

Part C- Molecular Docking:
Molecular Docking was performed using Schrödinger Glide (14). For docking in Schrödinger Glide, the ligands were prepared using LigPrep (15), 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)(16) 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) (17).

Section 2: Targets

Target 1: Nucleocapsid protein
The Nucleocapsid protein regulates the viral genome transcription, replication and packaging, and it is essential for viability. This protein contains two structural domains: the N-terminal domain that acts as a putative RNA binding domain and a C-terminal domain that acts as a dimerization domain. The N protein binds to the RNA genome of the SARS-CoV-2 virus and creates a capsid (shell) around the enclosed nucleic acid. Given the importance of nucleocapsid- mediated RNA packaging to the viral life cycle, small molecules that inhibit nucleocapsid self-assembly may be effective at reducing the severity of infections and the infectivity of patients.

The target protein was the homology model built in YASARA (1) for the N protein (the templates were downloaded from the protein data bank, the details can be found in the methods section). Before running POOL on this structure, it was analyzed in YASARA (1) and pKa prediction and energy minimization using YAMBER3 force field were done for this model. This model 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 (15). 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) ZINC library (https://zinc15.docking.org/)
c) CAS Antiviral set (https://www.cas.org/covid-19-antiviral-compounds-dataset)
d) Enamine FDA library (https://enamine.net/hit-finding/compound-collections/bioreference-compounds/fda-approved-drugs-collection)
e) Merck Library- Provided by the organizers of 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

Section A: Validation of the homology model:

As mentioned in the methods section, the homology model built for the Nucleocapsid protein was further validated using ERRAT and VERIFY 3D servers which are a part of the Structural Analysis and Verification Server (SAVES). VERIFY 3D analyses the residues based on their location and environment in the protein. It determines the compatibility of the model generated with its own amino acid sequence by assigning a structural class based on the location and environment and comparing the results to good structures. Verify 3D assigned a 3D-1D score of >0.2 for at least 99.17% of the amino acids. This implies that the model is compatible with its sequence.

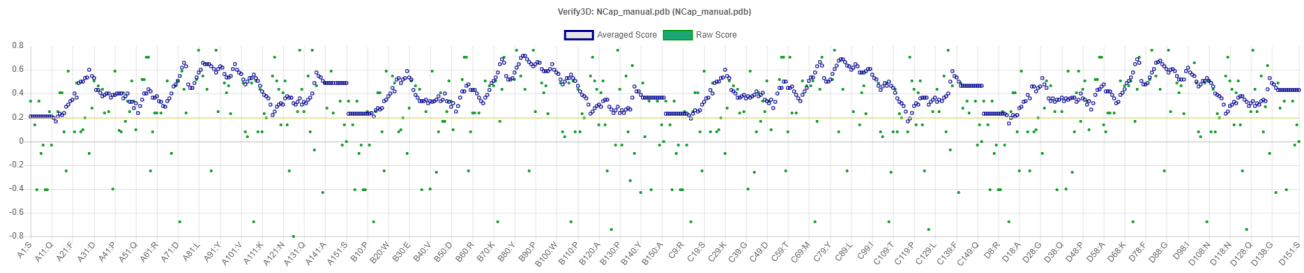

Figure: Verify 3D scores as a function of sequence number for the model

The ERRAT server is another part of the SAVES database. It helps in verifying protein structures. The error values are plotted as a function of the position of a sliding 9-residue window. The function is based on the statistics of non-bonded atom-atom interactions in the structure. The plot for the hybrid model generated by YASARA for N protein is shown below. Regions that can be rejected at 95% confidence level are yellow. 5% of a good protein structure are normally expected to have an error value above this level. Regions that can be rejected at 99% confidence interval are red. It can be seen from the figure below that the model contains significantly low red colored regions. The quality factor for this model is 94.16 for A chain, 94.57 for B chain, 92.42 for C chain and 91.67 for D chain. Therefore, it is a good model according to ERRAT.

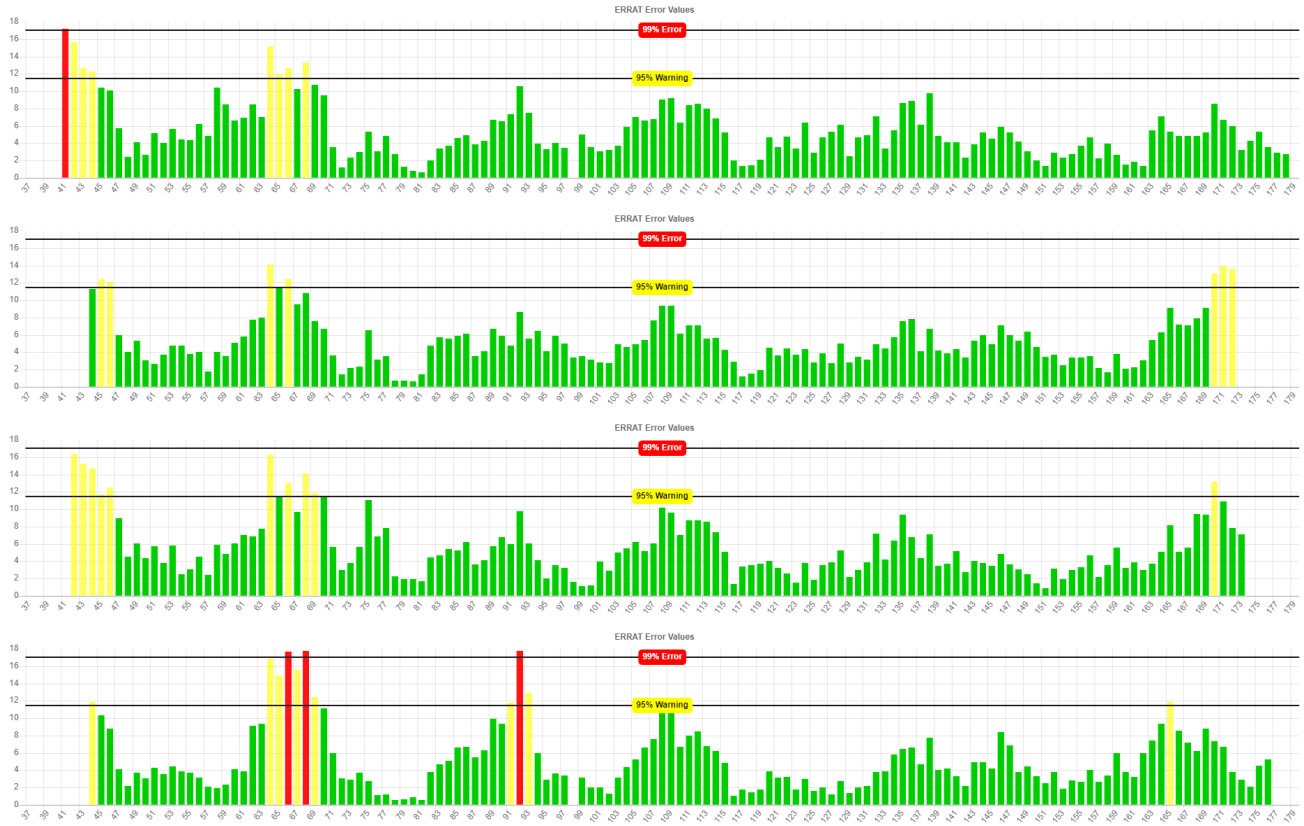

Figure 1: ERRAT scores as a function of sequence number for the model generated by YASARA. The first one is Chain A followed by B, C and D. Green indicates a good score. Yellow indicates regions that can be rejected at 95% confidence. Red indicates regions that can be rejected at 99% confidence.

The second server used to validate the homology model was SWISS model (9). QMEAN is a composite scoring function which derives both global and local absolute quality estimates based on one single model. The global scores are originally in a range [0,1] with one being good. Per default they are transformed into Z-scores to relate them with what we would expect from high resolution X-ray structures. The local scores are a linear combination of the 4 statistical potential terms as well as the agreement terms evaluated on a per residue basis. They are as well in the range [0,1] with one being good. The QMEAN score is 0.29. Below is an image showing the sequence of the protein colored by local quality. The orange areas denote poor quality whereas blue ones are of good quality. It can be seen from the image that most of the residues forming a part of the beta sheets have high confidence that they are predicted accurately whereas the loops connecting them do not have good confidence scores for accuracy.

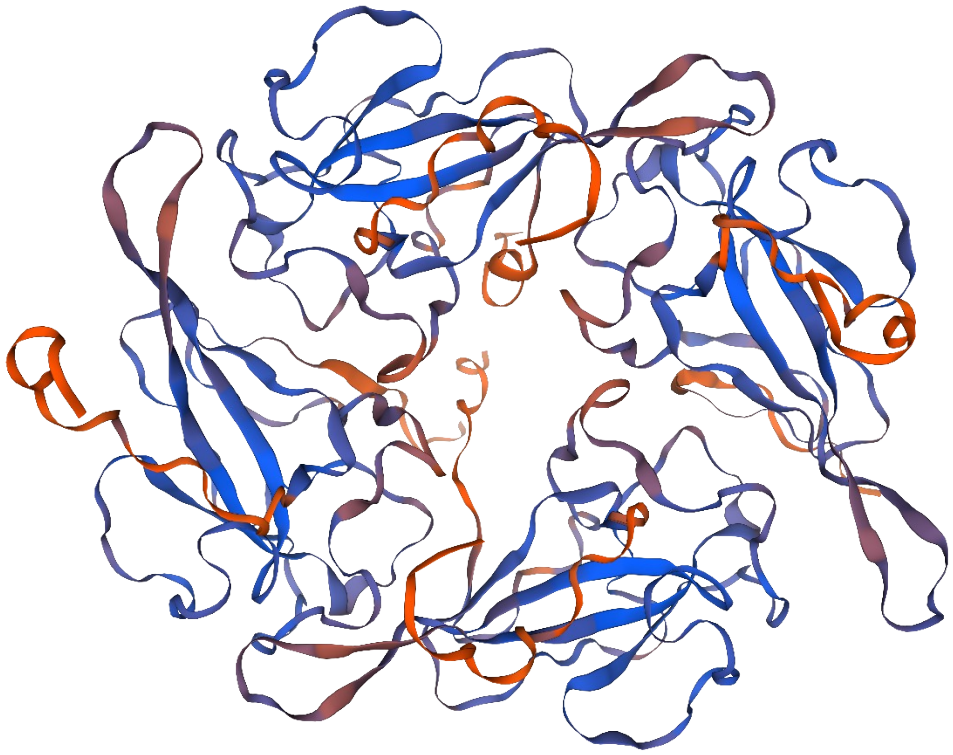

Figure 2: Image generated by QMEAN showing the local quality of the model. Blue indicates better quality regions; orange indicates lower quality regions.

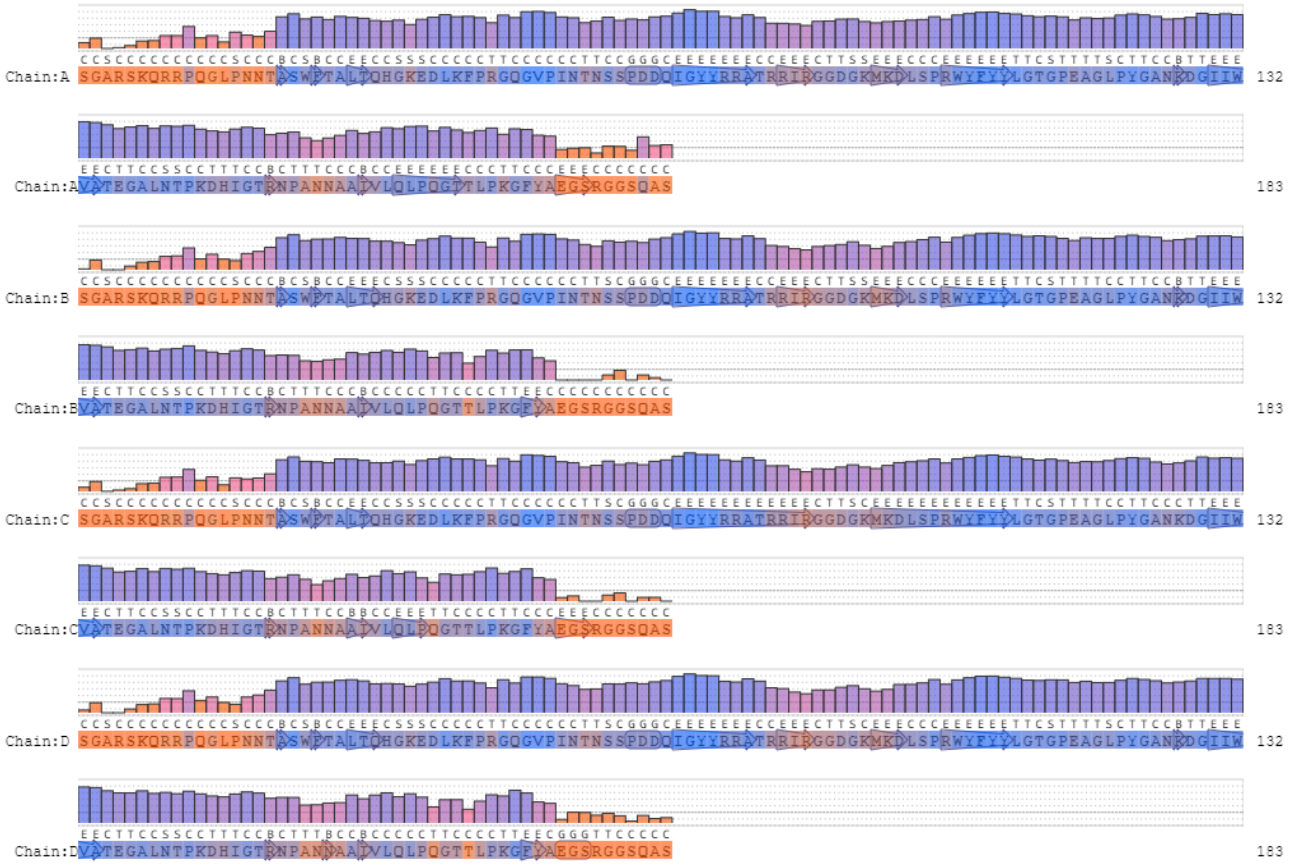

Figure 3: Image showing the local quality of the structure as a function of sequence number for all the chains, generated by QMEAN

The Ramachandran plot for this model can be found in the image below. There are 91.2% residues in the favorable regions and 8.8% residues in additionally allowed regions. There are no residues in disallowed region. This is further evidence of a quality model structure.

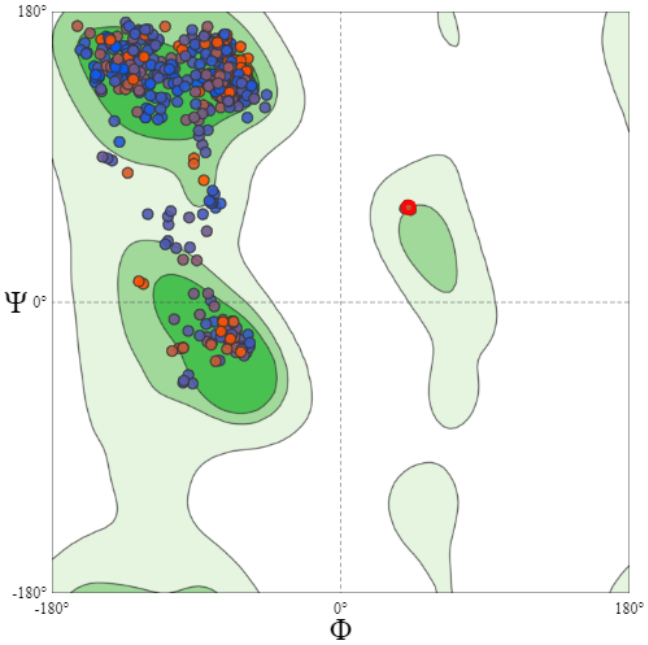

Figure 4: A Ramachandran plot for the hybrid model built by YASARA for N protein. Plot generated with the SWISS Model server.

Section B: 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. The POOL predicted sites for the Nucleocapsid protein are as follows.

88ARG 111TYR 109TYR 107ARG 86TYR 73PRO 87TYR 51SER 149ARG 92ARG 72VAL 49THR 50ALA 110PHE 112TYR

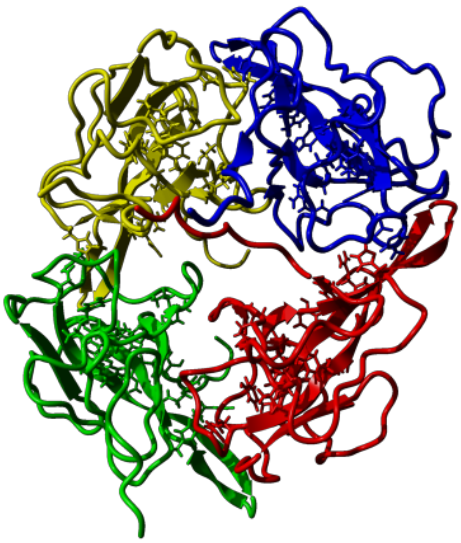

Figure 5: POOL predicted residues shown as stick in Chain a-Green, Chain B-Red, Chain C-Blue and chain D-Yellow.

Section C: 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)
926902-33-2	-12.532	-12.532	**Arg88**, **Arg92**, Thr57, Ser51, **Arg107**, **Arg107**, **Tyr109**
916792-68-2	-12.217	-12.218	**Ala55**, Ser51, **Tyr109**, **Arg107**, **Arg107**, **Arg92**
926902-37-6	-11.396	-11.398	**Tyr111**, **Tyr109**, Ser51, **Arg107**, **Arg107**, Ala55, Thr57, Tyr172, **Arg149**
10276-33-2	-11.134	-11.35	Asn48, **Arg88**, **Tyr109**, **Arg107**, **Arg149**, **Arg107**, Ala55, Thr57, Leu159
2130049-23-7	-11.045	-11.193	Ser51, **Tyr109**, **Arg149**, **Arg149**, **Arg107**, Ala156, **Arg107**

From the docked poses of these top ligands and their interactions it can be seen that the ligand sits inside the pocket of POOL predicted residues. The amino acid residue Arg107 is important for both hydrogen bonded interactions as well as forming a salt bridge. Tyr109 is an important residue for Pi-Pi stacking and hydrogen bonded interaction. Most of the residues interacting with the ligands are predicted by POOL.
More details about other ligands can be found in the Excel file attached.

References:

1. E. Krieger, K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, K. Karplus, Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8, Proteins, 77 Suppl 9 (2009) 114-122.
2. 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)
3. 6M3M- Kang S, Yang M, Hong Z, et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites [published online ahead of print, 2020 Apr 20]. Acta Pharm Sin B. 2020;10.1016/j.apsb.2020.04.009. doi:10.1016/j.apsb.2020.04.009
4. 6VYO- Not yet published
5. 6WJI- Not yet published
6. 6YI3- Not yet published
7. Colovos C, Yeates T., Verification of protein structures: patterns of nonbonded atomic interactions. Protein Science. 1993 Sep;2(9):1511-9.

   D. Eisenberg, R. Lüthy, J. U. Bowie, VERIFY3D: assessment of protein models with three-dimensional profiles. Methods in enzymology, Vol. 277 1997, pp. 396-404

8. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research 2018 46(W1), W296-W303.
9. Tong, W., Wei, Y., Murga, L.F., Ondrechen, M.J., Williams, R.J., Partial Order Optimum Likelihood (POOL): Maximum Likelihood Prediction of Protein Active Site Residues Using 3D Structure and Sequence Properties. PLoS Computational Biology, 2009, 5(1): e1000266.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. Schrödinger Release 2019-3: Glide, Schrödinger, LLC, New York, NY, 2019.
16. 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
17. 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.