Team name	Yoda Disciples
Team member(s) (firstname lastname; …)	Maria Rita Gulotta, Nedra Mekni, Maria De Rosa, Giada De Simone, Jessica Lombino, Alessandro Padova, Ugo Perricone
Affiliation	Fondazione Ri.MED
Contact email	uperricone@fondazionerimed.com
Contact phone number (optional)	
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	Spike (S) glycoprotein, 3-Chymotrypsin-like Protease (3CL^{pro}) and Papain-like Protease (PL^{pro})

Section 1: methods & metrics
Describe what methods you have used, how they are independent from one another, what your workflow was, how you performed the cross-correlation between your methods. If applicable, please report estimated performance metrics of your methods, such as accuracy, sensitivity, false-discovery rate, etc., and how those metrics were obtained (e.g. cross-validation). Please provide key references if available.

Methods:

SPIKE GLYCOPROTEIN

This work was based on a supervised consensus approach including several computational techniques, as described below. The first step consisted of I) a computational alanine scanning of the complex between spike (S) receptor-binding domain (RBD) and angiotensin-converting enzyme 2 (ACE2) peptidase domain (PD), using the PDB structures 6M17 and 6M0J, respectively; and II) two molecular dynamics (MD) simulations (each lasting for 200 ns) performed on both PDBs. This analysis allowed identifying the key residues of S RBD, and putative hot-spots were mainly found at the N-terminal of this potential binding pocket. These findings guided the next step, an *in silico* high-throughput screening campaign based on docking and pharmacophore screenings. For this purpose, ZINC12 and SWEETLEAD libraries were used in order to perform supervised docking screening by means of Glide software. Three hydrogen bond constraints were introduced at the N-terminal S RBD region. In detail, two H-bonds involved on the Gly496 carbonyl backbone carbonyl and amine groups; the third H bonding on the oxygen atom of Tyr505 side chain hydroxyl group. Both libraries were first screened using HTVS mode; the first 100,000 molecules were re-screened using Glide (Schrödinger, LLC) standard precision (SP) mode. The resulting molecules were further processed towards a pharmacophore map created by merging two more pharmacophore maps previously built on the SARS-CoV-2 PDB structures 6M17 and 6M0J, respectively. The pharmacophore maps were generated with LigandScout software. The obtained merged pharmacophore model showed four hydrogen bond acceptors for ACE2 and corresponding to Glu37 and Asp38 side chain carbonyl groups, Tyr41 side chain hydroxyl group, and Lys353 backbone carbonyl group, together with two hydrogen bond donors referring to ACE2 Gln42 side chain amine group and Tyr41 side chain hydroxyl group. The screening retrieved about 43,800 consensus molecules, that were filtered according to PAINS (Pan-Assays INterference compoundS) and REOS (Rapid Elimination Of Swill) structures. Finally, the first 10,000 consensus compounds were ranked and retained according to the best docking score values, and they are listed in the csv file attached to this application form.

3-CHYMOTRYPSIN-LIKE PROTEASE (3CL^{pro})
Several SARS-CoV-2 3-chymotrypsin-like protease (3CL^{pro}), also called main protease (M^{pro}), X-ray structures are available. Particularly, Diamond Light Source synchrotron facility has largely

contributed to enrich this collection with numerous M^{pro} complexes co-crystallized with fragments of covalent and non-covalent nature. From the analysis of these PDBs, we focused on the key residues non-covalently bound to the substrate with the help of Maestro GUI for molecular visualization. Table 1 summarizes all the key residues involved in non-covalent interactions detected.

Table 1. Non-covalent interactions observed in 3CL^{pro} PDB structures.

PDB id	π-π	H-bond						
	HIS41	CYS44	ASN142	GLY143	CYS145	HIS163	GLU166	GLN189
5R7Y	x							
5R7Z	x							
5RE4						x	x	
5R82	x							x
5RE9	x							
5REB	x	x						
5R83	x					x	x	
5REH						x		
5R84						x	x	
5REZ	x						x	
5RF1							x	
5RF2	x	x						
5RF3			x			x		
5RF6	x						x	
5RF7	x						x	
5RFE	x							

The key residues involved in ligand-protein complex stabilization were further analyzed by running a 50 ns MD using Desmond by Schrödinger v5.6 and using the PDB 6M2N. The simulation was run using an orthorhombic box and TIP3P water model. The box volume was minimized and OPLS3e force field (FF) was applied. The same FF was used to perform the MD simulation.

X-ray resolution and chemotype variety were considered to select three M^{pro} PDB structures including non-covalent inhibitors (PDB IDs: 6W63, 5RF7 and 5R83) were chosen in order to perform docking calculations.

Guided by the information collected from the simulation and the key residues analysis, the binding site was defined using the co-crystallized ligands coordinates, and adding some important constraints on specific amino acids, such as H-bonds on the carbonyl group side chain of Glu166, on the amine group of Gly143, on the nitrogen atom of His163, on the thiol moiety of Cys145. A constraint on π-π interaction with His41 ring was also introduced.

On the three receptor grids, the Glide virtual screening workflow was run. ZINC12 and SWEETLEAD libraries were initially docked in HTVS mode, and the first 100,000 molecules underwent docking in SP mode. The final outcomes were filtered from PAINS and REOS groups.

Finally, the outputs from each virtual screening were compared and only compounds prioritized by all the three screening campaigns, and based on enthalpy of binding, were extracted and selected by using KNIME analytic platform.

PAPAIN-LIKE PROTEASE (PL^{pro})
In order to identify the key residues for SARS-CoV-2 papain-like protease (PL^{pro}), a similarity analysis between SARS-CoV PL^{pro} and SARS-CoV-2 PL^{pro} was performed and a high identity percentage of 86% was obtained. Furthermore, the alignment of the C-α, performed with the Maestro superimposition tool, highlighted that the catalytic sites are conserved on the same region of the two proteases.

Therefore, through the study of PL^{pro} crystal structures (PDB IDs: 6WUU and 6WX4), we observed the crucial residues for the binding of inhibitors, i.e. Thr264, Asp164, Pro248, Gly171.

Moreover, a MD simulation of 50 ns was performed using an optimized PL^{pro} crystal structure (PDB ID: 6WX4).

On the basis of the information retrieved from PDBs analysis and the MD simulation, the receptor grid was defined including two constraints, on the hydroxyl moiety of Tyr264 side chain and on the carboxyl group of Asp164 side chain.

Then, the Glide virtual screening workflow was run. Firstly, ZINC12 and SWEETLEAD libraries were docked in HTVS mode, then, the first 100,000 results were docked in SP mode. Finally, the resulting best compounds in terms of binding energy were filtered from PAINS and REOS groups and they are listed in the attached csv file.

Section 2: targets
Describe for each protein target: why you chose it, from which source you obtained it (e.g., insidecorona.net / covid.molssi.org / rcsb.org) and why this is the best quality structure, if any pre-processing (e.g., energy minimization, residue correction, alternative folding, …) was performed.

Target 1: S GLYCOPROTEIN
During the viral infection, the first entry step for SARS-CoV-2 life cycle is mediated by the interaction of S glycoprotein to the host ACE2. S promotes the receptor recognition through S1 subunit RBD, and the membrane fusion via S$_2$ subunit.

Therefore, viral S RBDs bind ACE2 proteins to the N-terminal peptidase domain allowing the virus to enter the host cells. Due to its importance in viral infection process, targeting SARS-CoV-2 S$_1$ subunit interface could represent a good strategy to prevent the viral entry, blocking the spread and reducing the viral load at the very beginning of the infection[1–4].

Two PDB structures (PDB IDs: 6M17 and 6M0J) obtained from Protein Data Bank (https://www.rcsb.org/) were processed using Protein Preparation Wizard tool by Schrödinger, in order to optimize protein improprieties as missing hydrogens, loops and to avoid atomic clashes. The protonation state was set in range pH 7.0 ± 2.0. The protein was further optimized using energy minimization with the OPLS3e force field. Structures were then minimized by performing MD simulations and computational alanine scanning, with the aim of I) exploring the interactions between SARS-CoV-2 RBD and ACE2 PD; II) identifying the key residues for S. PDB 6M0J with a better resolution of 2.45 Å compared to the Cryo-EM (6M17, resolution of 2.9 Å) was chosen to perform virtual screening campaign of ZINC12 and SWEETLEAD libraries.

Target 2 and 3: 3-Chymotrypsin-like protease (3CL^{pro}) and papain-like protease (PL^{pro})
SARS-CoV-2 3CL^{pro} and PL^{pro} proteases have a crucial role in the viral replication process, as they are responsible for the cleavage of viral peptides into functional units for virus replication and packaging within the host cells[5–7].

The protein structures were downloaded from PDB database (https://www.rcsb.org/) and were prepared using Protein Preparation Wizard tool by Schrödinger, in order to optimize protein improprieties, such as missing hydrogens, loops, and to avoid atomic clashes. The protonation state was set in the pH range of 7.0 ± 2.0. The proteins were further optimized using energy minimization with the OPLS3e force field. In particular, for the 3CL^{pro} crystal structures, the water molecule interacting with His41 of 3CL^{pro} catalytic site was retained. In fact, experimental evidences[8] have reported that the water molecule occupying the catalytic site might stabilize the protonated His residue during catalysis. 3Cl^{pro} and PL^{pro} PDB structures used for the study showed resolutions ≤ 2.15

Å, which is an indication of a good crystal to perform computational studies and virtual screening calculations, as reported in literature[9,10]. Additionally, the electron density map well covered the proteins side chains and the co-crystalized ligands.

Section 3: libraries
Describe which libraries you have used, how they were combined, if any compounds were removed / added, why additions are relevant, any unique features of your library, etc. Please provide the sources you obtained the libraries from (if publicly available). Describe the procedure of data preparation (removal of duplicates, standardization, etc.). Indicate if different libraries were used for different targets, and why. If possible, provide a download link to your version of the library.

Libraries:
ZINC12 and SWEETLEAD libraries were used for the three above described virtual screening campaigns. Before performing calculations, for both compound libraries, we calculated ligand properties and we filtered them using Ligfilter tool of Schrödinger suite. In particular, on the basis of the calculated physico-chemical properties, the following criteria were set:

$logP_{o/w}$ range $\geq$ -2 and $\leq$ 6.5
MW > 250
PSA range $\geq$ 7 and $\leq$ 200
logS range $\geq$ -6.5 and $\leq$ 0.5
logHERG > -5
Nr of rotamers < 15 (for S RBD) and < 10 (for $3CL^{pro}$ and PL^{pro})
QPPCaco > 25
Nr of allowed Lipinski's rule of five violations $\leq$ 4

After calculations, the molecules were further filtered by using PAINS and REOS structures in order to delete all the compounds that could interfere with biological assays or containing reactive and toxic groups.

Section 4: results
Briefly describe your key findings, any interesting trends in your data, a description of your top 5 compounds for each target. If possible, provide a link to a code and/or data repository. Please do not submit randomly selected compounds!
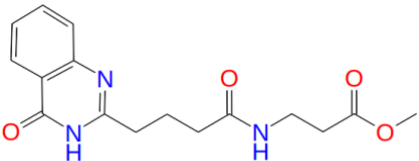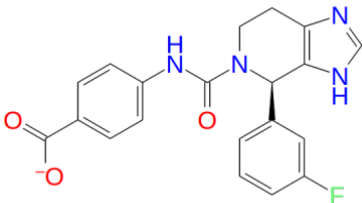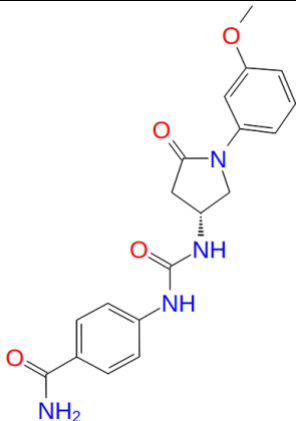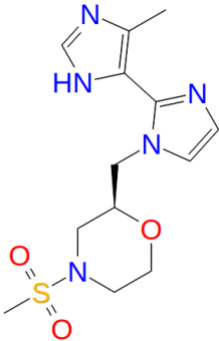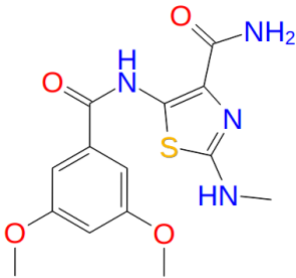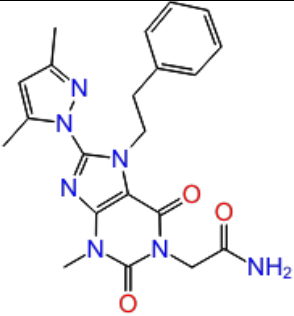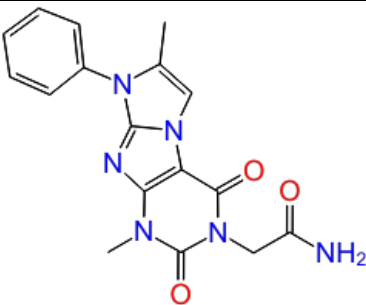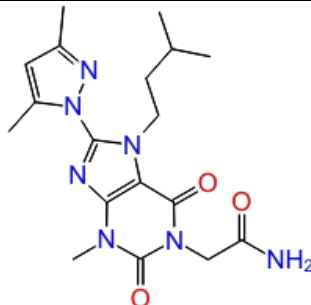

Results:
The attached csv file lists the first 10,000 consensus compounds per each protein target ranked according to the binding energy values expressed as enthalpy of binding.
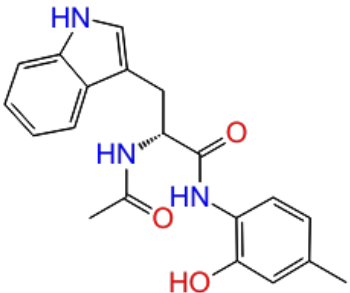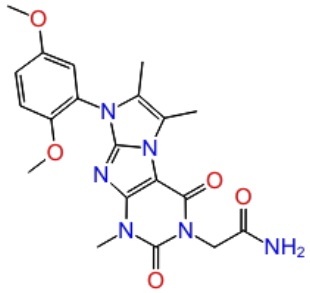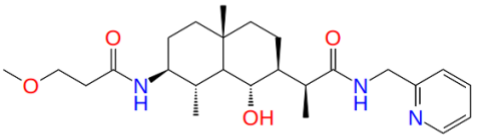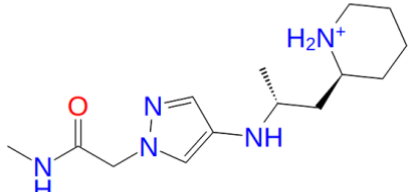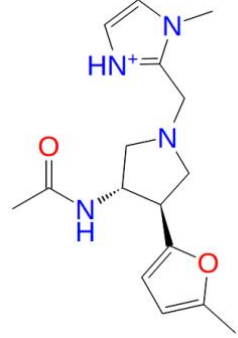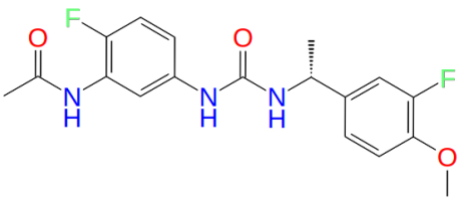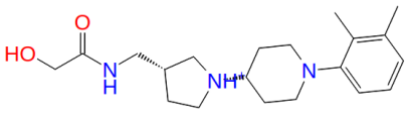
For S protein, the molecules were selected on the basis of docking and pharmacophore screenings results, only if they met at least two of the three defined docking constraints, and only when they met at least four of the six pharmacophore features, respectively.

For $3CL^{pro}$ the consensus molecules were identified retaining all those compounds shared among the three docking outputs. According to the docking grid constraints, each molecule met at least 2 H-bonds for PDB 6W63, 1-bond for PDB 5RF7 and 2 H-bonds for PBD 5R83.

For PL^{pro}, every resulting molecule met two constraints defined for the docking grid, such as a hydrogen bond on Asp164 side chain and a positional constraint on Tyr264 side chain.

Table 2 shows the structure of the top-ranked five molecules for each protein target.

S PROTEIN		
ZINC23172592 clogP*: 1.380 MW: 317.344	ZINC40284669 clogP*: 2.719 MW: 380.378	ZINC05190094 clogP*: 0.772 MW: 368.391
ZINC77322703 clogP*: 1.257 MW: 325.385	ZINC20220352 clogP*: 1.305 MW: 336.365	
3-CL^pro		
ZINC12231832 clogP*: 2.151 MW: 421.458	ZINC04303874 clogP*: 1.47 MW: 352.352	ZINC12231664 clogP*: 1.5 MW: 387.441

ZINC89618901 clogP*: 2.63 MW: 351.401	ZINC09713137 clogP*: 1.619 MW: 426.431	
PL^pro		
ZINC30884046 clogP*: 2.398 MW: 445.601	ZINC93486313 clogP*: 0.467 MW: 279.384	ZINC91625332 clogP*: 1.820 MW: 302.375
ZINC44015908 clogP*: 2.995 MW: 363.363	ZINC67754489 clogP*: 1.960 MW: 345.484	

* Calculated values

References

1. Yan, R. *et al.* Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* **2762**, 1–10 (2020).
2. Prabakaran, P., Xiao, X. & Dimitrov, D. S. A model of the ACE2 structure and function as a SARS-CoV receptor. *Biochem. Biophys. Res. Commun.* **314**, 235–241 (2004).
3. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structural biology: Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science (80-.).* **309**, 1864–1868 (2005).
4. Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu. Rev. Virol.* **3**, 237–261 (2016).
5. Xue, X. *et al.* Production of Authentic SARS-CoV Mpro with Enhanced Activity: Application as

a Novel Tag-cleavage Endopeptidase for Protein Overproduction. *J. Mol. Biol.* **366**, 965–975 (2007).

6. Ghosh, A. K., Brindisi, M., Shahabi, D., Chapman, M. E. & Mesecar, A. D. Drug Development and Medicinal Chemistry Efforts Toward SARS-Coronavirus and Covid-19 Therapeutics. *ChemMedChem* (2020). doi:10.1002/cmdc.202000223

7. Jin, Z. *et al.* Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature* (2020). doi:10.1038/s41586-020-2223-y

8. *Molecular Biology of the SARS-Coronavirus*. (Springer Berlin Heidelberg, 2010). doi:10.1007/978-3-642-03683-5

9. Sulea, T., Lindner, H. A., Purisima, E. O. & Ménard, R. Deubiquitination, a New Function of the Severe Acute Respiratory Syndrome Coronavirus Papain-Like Protease? *J. Virol.* **79**, 4550–4551 (2005).

10. Rut, W. *et al.* Activity profiling of SARS-CoV-2-PLpro protease provides structural framework for anti-COVID-19 drug design. *bioRxiv* 2020.04.29.068890 (2020). doi:10.1101/2020.04.29.068890