Team name	Laboratory of Chemoinformatics
Team member(s) (firstname lastname; …)	Dragos Horvath, Gilles Marcou, Alexandre Varnek
Affiliation	UMR7147, University of Strasbourg & CNRS
Contact email	dhorvath@unistra.fr
Contact phone number (optional)	
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	3CLProtease (6W63), Papain-like Protease (6W9C), RNA polymerase

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:

Four distinct chemoinformatics strategies were used in our approach:

- Chemical space analysis using Generative Topographic Mapping
- Similarity searches using high-content molecular descriptors
- Targeted docking: refinement of above-selected candidates by docking into selected targets
- SVMdock – a fast Support Vector Machine-guided docking of large libraries, without prefiltering

1. Database standardization: The ~1.3 billion ZINC[1] compounds (zinc.docking.org; highest reactivity: "standard", minimum purchasability "boutique") were downloaded from the web site zinc.docking.org. These included only drug discovery-relevant compounds, classified as and submitted to the standardization protocol of our in-house chemoinformatics web server http://infochim.u-strasbg.fr/webserv/VSEngine.html. After standardized rendering and removal of stereochemical information, 800M distinct compounds remained. The gap to the original 1.3 billion is made of genuine duplicates, compounds reported as different formulations (example – the acid and its sodium salt, redundant from a cheminformatics perspective), compounds reported under different tautomeric forms, stereoisomers now considered as a same "skeleton" (this radical measure was undertaken because chemoinformatics descriptors used at stage 2, *vide infra*, are stereochemistry-agnostic) – and last but not least compounds with unwanted groups (including ones discarded by the web server, plus additional anti-PAINS filters). This was part of a distinct project by Y. Zabolotna, Ph.D. student in the Laboratory of Chemoinformatics (publication submitted).
   In addition, the 5M compounds of the AMS Merck library shared to the JEDI participants were also standardized according to the above-mentioned protocol.
2. Chemoinformatics-driven preselection. Several chemoinformatics techniques were used to preselect specific subsets of potential interest for docking:
   a. 55K ZINC molecules present within the Relevant Antiviral Spaces (RAS) of at minimum three of the seven Universal Maps[2] were selected following the already published procedure[3] which originally reported the >400 compounds residing in 4/7 of these RAS. This 55K collection will be termed "ZINC-RAS"
   b. 58K ZINC molecules were selected by a consensus similarity screening using as references 84 antiviral compounds from the DrugBank (compound names ending in "-

vir"). Considering the seven ISIDA fragment descriptor[4, 5] spaces at the basis of above-mentioned Universal maps, each such reference was encoded by its respective descriptor ("located" in the respective descriptor space), and its similarity to each of the 800M ZINC compounds was calculated using the Tanimoto[6] score. In each of the seven descriptor spaces, the list of top thousand ZINC neighbors surrounding the reference compound was established. Only ZINC compounds which "made it" into the top1000 neighbors in two or more of the seven descriptor spaces (consensus neighbors) were kept (on the average, some 690 distinct ZINC compounds/reference molecule), thus 690×84=58K compounds selected in the final library, further on labeled "ZINC-VIR"

   c. The published audit of coronavirus-related structure-activity data revealed a very small series of compounds confirmed to be active on the SARS-CoV 3CL protease, according to the ChEMBL[7] database. These 25 molecules were at their turn employed as similarity search references and deployed for similarity-based virtual screening of ZINC, as described above. A pool of 24885 compounds, herein labeled "ZINC-3CLpro" resulted.
   d. In absence of any RNA polymerase structure-activity data on coronaviruses, a search for RNA polymerase inhibitors of any viruses was conducted in ChEMBL, and 29 most potent nanomolar inhibitors were selected. They too were used in the similarity search-driven virtual screening and returned a "ZINC-RNApol" focused library of 15245 compounds.

3. Docking of 800M compounds is technically unfeasible. Furthermore, relying on a single criterion – the docking score, an often flawed[8] result, is risky, especially when applied to compounds that are different from the training set serving to calibrate that score. Therefore, application of docking to the subsets defined by chemoinformatics is beneficial in terms of both computer effort economy and consensus effect achieved by these orthogonal approaches. Cheminformatics-driven preselection removes many candidates that are manifestly outside the relevant (as to-date understood) chemical space of coronavirus protease inhibitors, but would achieve fake favorable docking scores nevertheless. Compliance to chemoinformatics filters plus favorable docking scores is better than favorable docking score alone. The pipeline of chemoinformatics filters followed by docking will be termed "Targeted" docking.
4. It might be argued that "targeted" docking is likely safer and bound to achieve higher hit rates because of the "conservative" candidate preselection by chemoinformatics filters tending to pick those compounds containing already known features of so-far known antivirals. As such, completely novel "paradigm breaking" species that potentially dock without resembling known antivirals might be discarded. Such discoveries are low-likelihood events but "open-ended" docking might *in principle* recognize such compounds. Therefore, an additional strategy of machine-learned enhanced docking (termed "SVMdock", because it relies on Support Vector Machines rather than the deep neural nets of "deep docking[9]") was developed on purpose and used to process the 5M compounds of the Merck AMS library. The approach uses our evolutionary model builder[10] to select best-suited descriptors (out of 100 diverse ISIDA fragmentation schemes) and SVM parameters defining top-performing SVM regression models trained to predict Ligand Efficacy LE=(docking score)/(number of heavy atoms), on an initial training set of 3000 randomly picked Merck compounds. Then, a next pool of 3K random "test set" compounds are being docked, all while having their LE values predicted by the calibrated SVM model. Should the prediction not be in very close agreement to actual LE scores (R^2<0.8), then test and previous training set are merged, in order to refit a new predictive model. As soon as the current predictive model successfully predicts five further test sets with R^2≥0.8, the model is deemed trustworthy, and is used to run a LE prediction for the entire Merck library (except, of course, already processed training and test sets). Top5% of candidates with best predicted LE are selected and docked. Since ISIDA fragment-based SVM models turned out to

achieve trustworthiness status at $R^2 \geq 0.9$ in a few iterations, the SVMdock procedure of the Merck AMS database completed in a few days on a regular desktop Linux workstation.

The program PLANTS[11, 12] was used for docking (in "precision mode" speed1, with the ChemPLP score) – both for Targeted docking and SVMdock. In terms of targeted docking, ZINC-RAS, ZINC-VIR and ZINC-3CL were all docked into the active site of the 6W63 structure of 3CL protease. ZINC-RAS and ZINC-VIR were also targeted against the papain-like protease 6W9C (ZINC-3CL being designed as 3CL-specific was not submitted to the latter). SVMdock of the Merck AMS collection was performed against both 6W63 and 6W9C, independently. In order to select the top 10K dockers to report, both PLANTS ChemPLP docking scores and ligand efficacies (docking score/number of heavy atoms) were estimated for all docked candidates. It appears that these entities are strongly anticorrelated: larger molecules with low efficacy are nevertheless the ones with the most favorable scores, because of their size. Therefore, rational selection should have been performed in agreement with the wider strategical aims of the project. In view of a hit-to-lead optimization campaign, smaller molecules with high efficacy should have been prepared. Otherwise, molecules with optimal global docking scores should be privileged (albeit their large size and complexity may be the source of both synthesis and bioavailability problems). As the utility of these compounds and the role of the JEDI challenge within a long-term drug design program are not clearly defined, a compromise selection strategy was chosen. As the PDB ligands of 6W63 and 6W9C typically contain about 30 heavy atoms, candidates with less than 20 or more than 40 heavy atoms were discarded. These typically represented <10% of the pools of preselected molecules, as the preselection *per se* favors "antiviral-like candidates", but they were the ones to reach top absolute docking scores (the largest) or, respectively, top ligand efficiencies (the smallest, nucleotide-like in ZINC-VIR, which need not be tested on proteases – top ligand efficiency notwithstanding). For the remaining compounds, enlarged "Pareto[13] fronts" of compounds featuring best efficacies at given global docking scores were selected for both 3CL and papain-like proteases (Figure 1). These included a controlled amount of "second-line", dominated molecules, such as to ensure selection of the imposed 10K compounds/target.

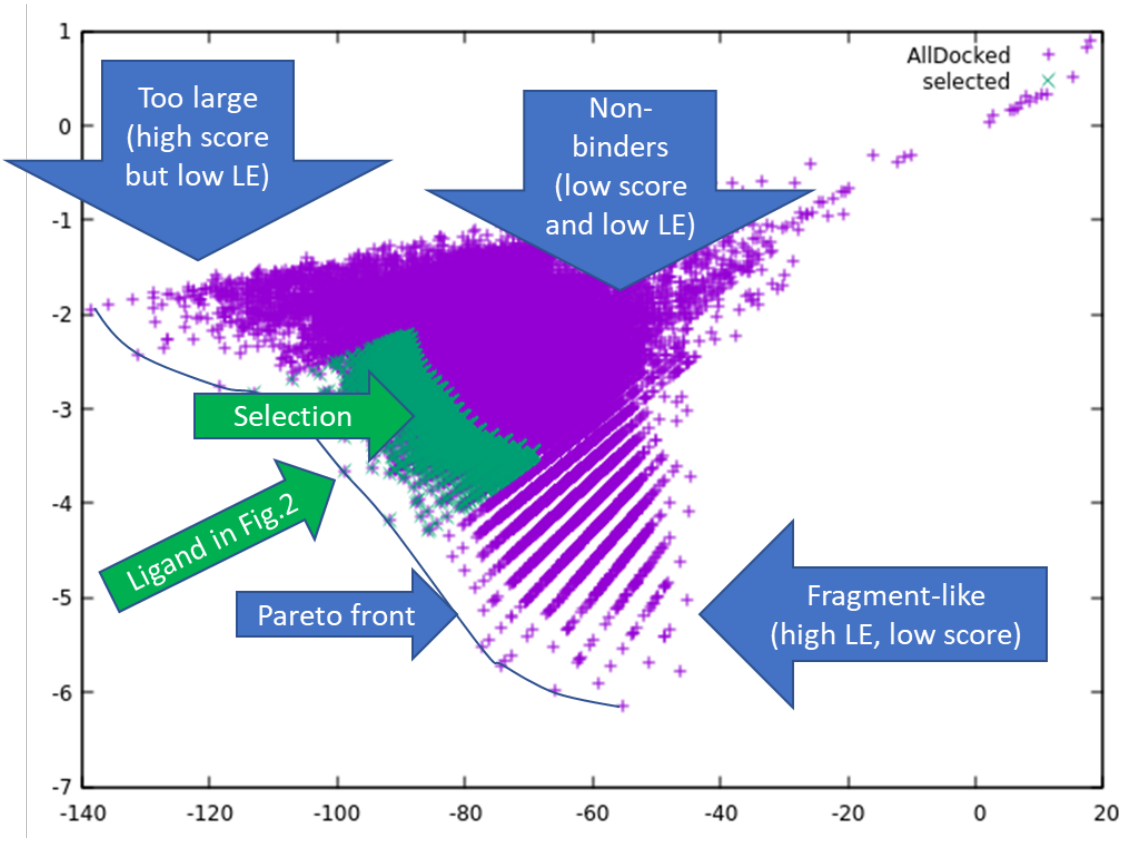

Figure 1: Enlarged Pareto front (Global ChemPLP docking score:X, ligand efficiency:Y) for the 108K (ZN-RAS+ZINC-VIR) compounds docked into 6W9C. In green: selected 10K compounds, of intermediate sizes (between 20 and 40 heavy atoms)

They were ranked by the geometric means of their global docking score and ligand efficacies, respectively – species with good scores and good efficacies were ranked first. The "strict" pareto fronts (compounds that were never dominated) represent 19 molecules for 3CLpro and 20 molecules for the papain-like protease. These compounds were redocked several times and their interactions with the binding sites were studied.
Eventually, we consider that no docking tool is pertinent for docking into the RNA polymerase site, because of the extreme impact of electrostatic and ionic atmosphere effects that are notoriously failed to account for in any classical scoring functions. Therefore, nucleotide-likeness of candidates, as captured by the similarity search was deemed as useful a selection criterion: the top 10K nearest neighbors of RNA polymerase inhibitors from ZINC-RNApol were reported, ranked by the mean, over the seven used descriptor spaces, of their Tanimoto similarity score to their nearest of reference RNA polymerase inhibitors.

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:

6W63 (PDB) was chosen because it featured a reversible, non-covalent ligand structure. This option is the only one supported by classical docking scores (and is also preferable to reactive ligands from a safety/toxicity perception).

Target 2:

6W9C is the only papain-like protease structure for SARS-CoV-2 to our knowledge, crystallized without ligand, while the highly homologous 4OW0 of the SARS-CoV virus does feature a non-covalently bound ligand. The latter was transferred into the 6W9C site after protein alignment and served for active site definition.
Target 3:
7BV2 was analyzed, but as mentioned in Methods we do not think that high-throughput docking into this type of electrostatics-controlled enzymes is meaningful.

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.

Library 1: ZINC
Library 2: Merck AMS

Section 4: results
Briefly describe you 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 Results section of a Virtual Screening article is typically written *after* experimental results (in)validate the used strategy. The only thing that can be claimed so far is that ligands binding with seemingly strong and meaningful non-covalent interactions to the protease sites were found. There are many methods for a further refinement of virtually screened selections – using interaction fingerprints to pick only compounds fulfilling hydrogen bonds and hydrophobic contacts to known key residues. Yet, since this is supposed to be an exploratory screening, forcing known interaction rules may limit discovery of new compounds to already observed interaction patterns. Hit rates would be certainly much higher, but not much new would be learned. Furthermore, given the paucity of known SARS-CoV2 protease inhibitors, it may be premature to attribute “key” roles to any active site residue (except for the catalytic triad). An example of “interesting” binding pattern into the active site of the 3CL protease is illustrated in Figure 2 below.

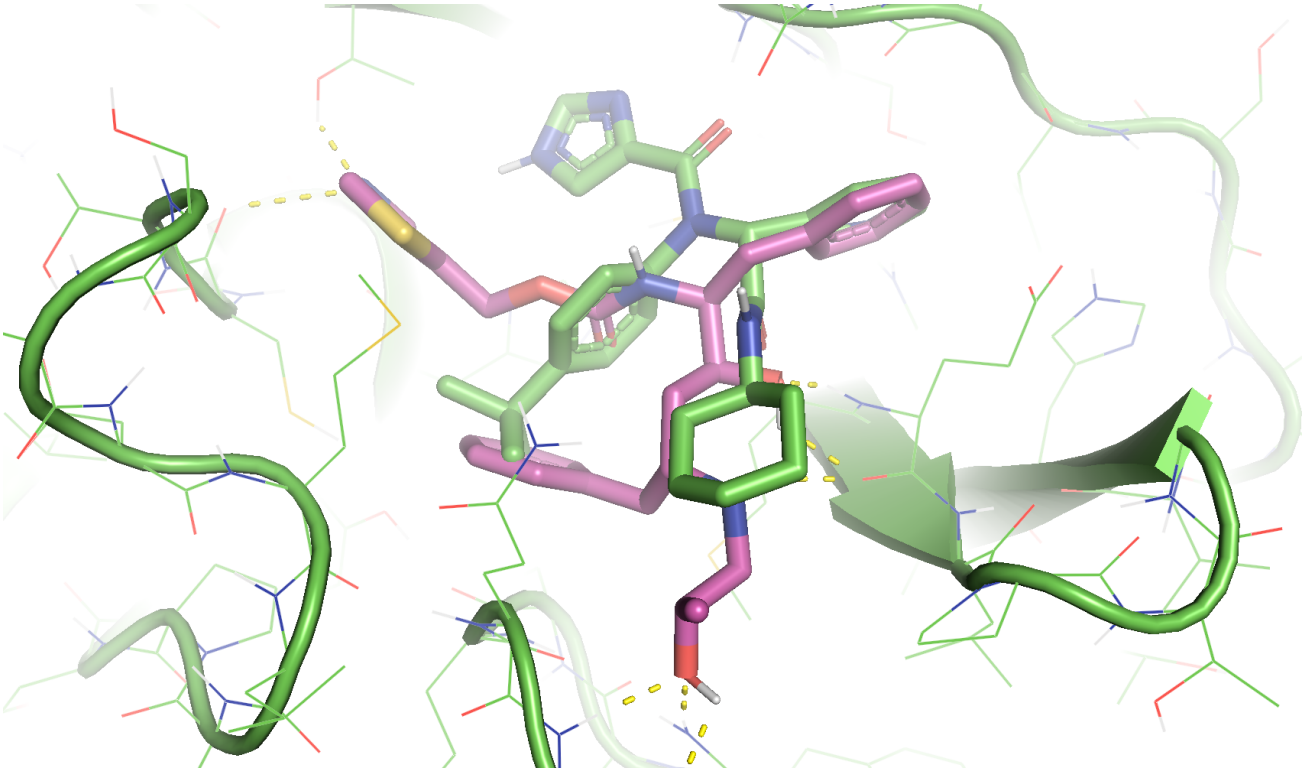

Figure 2: Docked ligand (in magenta) filling in the space occupied by the native ligand of 6W63 (green) and partially overlapping with it, forms however a much denser hydrogen bonding network with the protein (yellow dashed lines)

Other comments:

1. Irwin, J. J.; Shoichet, B. K., ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model* **2005**, 45, 177-182.
2. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *Journal of Computer-Aided Molecular Design* **2015**, 29, 1087-1108.
3. Horvath, D.; Orlov, A.; Osolodkin, D. I.; Ishmukhametov, A. A.; Marcou, G.; Varnek, A., A Chemographic Audit of anti-Coronavirus Structure-Activity Information from Public Databases (ChEMBL). *Molecular Informatics* **2020**, n/a.
4. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., Isida Property-labelled Fragment Descriptors. *Molecular Informatics* **2010**, 29, 855-868.
5. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. v.; Marcou, G., Isida - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, 4, 191-198.
6. Horvath, D.; Marcou, G.; Varnek, A., Do Not Hesitate to Use Tversky-and Other Hints for Successful Active Analogue Searches with Feature Count Descriptors. *J Chem Inf Model* **2013**, 53, 1543-62.
7. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2011**, 40, D1100-D1107.
8. Zhenin, M.; Bahia, M. S.; Marcou, G.; Varnek, A.; Senderowitz, H.; Horvath, D., Rescoring of docking poses under Occam's Razor: are there simpler solutions? *Journal of Computer-Aided Molecular Design* **2018**, 32, 877-888.

9. Ton, A.-T.; Gentile, F.; Hsing, M.; Ban, F.; Cherkasov, A., Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Molecular Informatics* **2020**, n/a.
10. Horvath, D.; Brown, J.; Marcou, G.; Varnek, A., An Evolutionary Optimizer of libsvm Models. *Challenges* **2014**, 5, 450-472.
11. Korb, O.; Stützle, T.; Exner, T. E., Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *Journal of Chemical Information and Modeling* **2009**, 49, 84-96.
12. Korb, O.; Stützle, T.; Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In Ant Colony Optimization and Swarm Intelligence, Berlin, Heidelberg, 2006//, 2006; Dorigo, M.; Gambardella, L. M.; Birattari, M.; Martinoli, A.; Poli, R.; Stützle, T., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp 247-258.
13. Cottrell, S. J.; Gillet, V. J.; Taylor, R., Incorporating partial matches within multiobjective pharmacophore identification. *Journal of Computer-Aided Molecular Design* **2006**, 20, 735-749.