Large-scale ligand-based virtual screening for SARS-CoV-2 inhibitors using deep neural networks

Markus Hofmarcher [1 2 *] **Andreas Mayr** [1 2 *] **Elisabeth Rumetshofer** [1 2 *] **Peter Ruch** [1 2 *] **Philipp Renz** [1 2 *] **Johannes Schimunek** [1 2 *] **Philipp Seidl** [1 2 *] **Andreu Vall** [1 2 *] **Michael Widrich** [1 2 *] **Sepp Hochreiter** [1 2 *] **Günter Klambauer** [1 2 *]

Abstract

Due to the current severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, there is an urgent need for novel therapies and drugs. We conducted a large-scale virtual screening for small molecules that are potential CoV-2 inhibitors. To this end, we utilized "ChemAI," a deep neural network trained on more than 220M data points across 3.6M molecules from three public drug-discovery databases. With ChemAI, we screened and ranked one billion molecules from the ZINC database for favourable effects against CoV-2. We then reduced the result to the 30,000 top-ranked compounds, which are readily accessible and purchasable via the ZINC database. We provide these top-ranked compounds as a library for further screening with bioassays at https://github.com/ml-jku/sars-cov-inhibitors-chemai.

Introduction. Due to the current world-wide crisis of SARS-CoV-2 virus infections, there is a strong need for new therapies. While many efforts are focused on repurposing existing drugs (Zhou et al., 2020; Wang et al., 2020; Ton et al., 2020), we suggest to test new molecules with potentially higher efficacy. Therefore, we performed a large-scale ligand-based virtual screening run, which resulted in 30,000 potential SARS-CoV-2 inhibitors with favorable properties. We actively outreach to the scientific community to test these molecules and consider them as a custom-designed chemical library.

Most current virtual screens are structure-based and use docking methods (Chen et al., 2020; Huang et al., 2020; Haider et al., 2020; Wang et al., 2020; Fischer et al., 2020; Chen et al., 2020; Ton et al., 2020; Senathilake et al., 2020; Ruan et al., 2020; Jin et al., 2020; Zhang et al., 2020) while one screen is ligand-based and uses a similarity-based approach (Zhu et al., 2020). The largest docking studies screen databases with sizes ranging from roughly 700 million (Fischer et al., 2020) to 1.3 billion (Ton et al., 2020) molecules. Also our study operates on databases of this size, concretely we perform a ligand-based virtual screening of a collection of one billion molecules from the ZINC database.

Deep ligand-based virtual screening. "ChemAI" is a deep neural network trained to simultaneously predict a large number of biological effects (Mayr et al., 2018; Preuer et al., 2019). In more detail, the network is trained on a data set comprised of ChEMBL (Gaulton et al., 2017), ZINC (Sterling & Irwin, 2015) and PubChem (Kim et al., 2016), and which is similar to the data set used by Preuer et al. (2018). ChemAI predicts 6,269 biological outcomes, such as binding to targets, inhibitory or toxic effects. The network was trained in a multi-task setting, in which data from other bioassays was used to enhance the predictive power for SARS-CoV inhibitory effects. Each modelled biological effect is represented by an output neuron of the neural network. We utilized a small set of output neurons associated with SARS-CoV inhibition and a set of output neurons associated with toxic effects to rank compounds.

We screen the ZINC database because it contains a large set of diverse molecules and additionally provides links to vendors from which to purchase and physically obtain those molecules. We downloaded 898,196,375 molecules from ZINC and converted them to canonical SMILES (Weininger, 1988) using RDKit (Landrum, 2006). We then performed inference with ChemAI to obtain predictions for each of those roughly one billion molecules.

Selecting bioassays for multiple targets of SARS-CoV. The SARS-CoV-2 has two main proteases that are critical for its replication, namely the 3CLpro (3C-like protease) and PLpro (Papain Like Protease), encoded in an open reading frame (Macchiagodena et al., 2020). A compound that inhibits both proteases could be promising drug candidates

[*]Equal contribution [1]ELLIS Unit at the LIT AI Lab, Johannes Kepler University Linz, Austria [2]Institute for Machine Learning, Johannes Kepler University Linz, Austria. Correspondence to: Günter Klambauer <klambauer@ml.jku.at>.

Technical Report

Assay ID	Source	#inact	#act	Description
1706	PubChem	193637	269	QFRET-based assay for SARS-CoV 3C-like Protease
1879	PubChem	167	86	QFRET-based assay for SARS-CoV 3C-like Protease (confirmation)
485353	PubChem	215030	390	Yeast-based Assay for SARS-CoV PLP
652038	PubChem	493	135	Yeast-based Assay for SARS-CoV PLP (validation)

Table 1. Overview of the main biological effects considered for ranking the molecules of the virtual screen. "#inact" and "#actAll" report the number of actives and inactives in the training set. All assays are based on inhibition of proteins of SARS-CoV-1.

(Ledford, 2009; Collison, 2019). The virus proteases are also strikingly similar to those in SARS-CoV-1 (Macchiagodena et al., 2020), which is also an implicit assumption by docking-based approaches. We therefore select two groups of assays, one of which measures the inhibition of 3CLpro and the other the inhibition of PLpro (see Table 1). For each of those four assays, ChemAI possesses an output unit, which models the ability of small molecules to exhibit the effect measured by the assay. Thus, using the predictions yielded by ChemAI, it is possible to rank compounds by their predicted ability to inhibit the two main proteases of SARS-CoV-1, which can be a proxy for the inhibitory potential for SARS-CoV-2.

Consensus ranking. We developed a library of compounds which is enriched for molecules with the ability to inhibit both proteases of the SARS-CoV-2. In order to score the multi-target effect, we calculated a consensus score for each molecule as the average rank of the predictions over the four selected assays (see Table 1). We then ranked all compounds by this consensus score. For each of the top-ranked compounds, we also calculated their minimal distance to actives in the training set to be able to identify novel chemical structures. Furthermore, for each compound we also report its number of potential toxic effects (Mayr et al., 2016).

For the distance metric, we used the Jaccard distance based on binary ECFP4 fingerprints folded to a length of 1024, which yields values in the interval $[0, 1]$. For potential toxic effects, we used 75 output units of ChemAI with high predictive quality, concretely an area under ROC-curve (AUC) larger than 0.80, and counted how many of those output units indicated a toxic effect. This value is reported in Table 2 (column "tox"). Furthermore, we report the clinical toxicity probability predicted by an independent multitask neural network fitted on the ClinTox dataset (Wu et al., 2018). These probability values are calibrated by Platt scaling (Platt et al., 1999) and reported in Table 2 (column "ct"). The additional information contained in these values can be used to obtain a refined ranking for testing the molecules.

We implemented the overall process as a two-step approach. In the first step, we reduced the ZINC database of one billion molecules to a smaller set, where we kept all molecules that exhibited some predicted activity on any of the four assays (precisely, at least one of the predictions had to reside in the top-1% quantile). In this way, we obtained an intermediate dataset of 5,672,501 molecules. For those molecules, the consensus score, the toxicity flags and the distance to known actives were calculated. In the second step, we reduced the dataset to the top-ranked 30,000 molecules by the consensus score.

Results. With the abovementioned approach, we assembled a library of potential inhibitors of SARS-Cov-2. We report three metrics for each compound: a) predicted inhibitory effect of SARS-CoV proteases b) potential toxicities and c) distance to known actives. This led to a ranked list of compounds of which we provide the top 30,000 as a screening library. The top-ranked molecules are given in Table 2 and Figure 1.

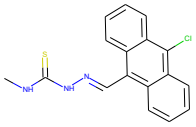
We also checked whether molecules suggested by other publications can be confirmed by ChemAI. Overall, some suggested molecules show at least mild predicted activity against SARS-CoV (see Table 3).

Discussion. In this work, we presented the construction of a screening library of small molecules that are potential inhibitors of SARS-CoV-2. Our ligand-based approach uses a neural network trained to predict the outcomes of bioassays. From this multi-task models, four tasks have been selected to predict the inhibitory potential against SARS-CoV-1. A consensus between these predictions was used to rank compounds from the ZINC database, of which the 30,000 top-ranked are reported.

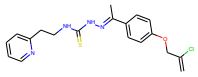
The approach is limited by the predictive quality of the underlying machine learning method, evaluated via AUC and leading to values in the range of 0.69 to 0.78. While these results are very promising, improved data quality, larger amount of data or machine-learning approaches could lead to increased predictive performance and quality of the library. We expect that the data for SARS-CoV-1 already has high predictive power for inhibitory effects of compounds on SARS-CoV-2. However, the current predictions can be further adjusted toward SARS-CoV-2 via transfer-learning and the incorporation of new data from SARS-CoV-2. In particular, few shot learning may be utilized for the first

ZINC ID	Canonical SMILES	dist	score	tox	ct
ZINC000254565785	CNC(=S)NN=Cc1c2ccccc2c(Cl)c2ccccc12	0.5455	0.8244	8	0.06
ZINC000726422572	C=C(Cl)COc1ccc(C(C)=NNC(=S)NCCc2ccccn2)cc1	0.5333	0.8232	7	0.05
ZINC000916265995	CNC(=S)NN=Cc1cc2cccc(C)c2nc1Cl	0.6111	0.8230	5	0.08
ZINC000916356873	N#CCCn1cc(C=NNC(=S)NCCc2ccc(Cl)cc2)c2ccccc21	0.6377	0.8221	17	0.07
ZINC000806591744	O=c1c(Br)nn(Cc2cnc3ccccc3c2)c2ccccc12	0.7258	0.8215	11	0.16
ZINC000178971373	O=c1c(Br)nn(Cc2nc3ccccc3s2)c2ccccc12	0.7288	0.8211	8	0.05
ZINC000000155607	CSC(=S)N/N=C/c1ccc2cc3ccccc3cc2c1	0.3902	0.8204	4	0.05
ZINC000016317677	C=CCNC(=S)NNC(=O)Cn1c(COc2ccc(Cl)cc2)nc2ccccc21	0.7000	0.8197	4	0.07
ZINC000193073749	O=C(Cn1cccc(Br)c1=O)c1ccc2ccccc2c1	0.6667	0.8197	1	0.13
ZINC000769846795	O=c1c(Br)nn(Cc2ccc3ncccc3c2)c2ccccc12	0.6949	0.8195	9	0.14
ZINC000755523869	CN(N=Cc1nc2ccccn2c1Br)C(=S)NCc1ccccc1	0.6452	0.8194	4	0.05
ZINC000763345954	C=CCNC(=S)NN=Cc1nc2ccc(Cl)cc2n1C	0.6508	0.8194	5	0.07
ZINC000001448699	CSC(=S)N/N=C/c1nc(-c2ccc(Cl)cc2)n2ccccc12	0.6866	0.8192	4	0.13
ZINC000016940508	C/C(=N\NC(=S)NNC(=S)N(C)c1ccccc1)c1nccc2ccccc12	0.6721	0.8191	11	0.06
ZINC000005486767	C/C(=N/NC(=S)NNC(=S)N(C)c1ccccc1)c1nccc2ccccc12	0.6721	0.8191	11	0.06
ZINC000005527649	CSC(=S)N/N=C/c1ccc2ccccc2n1	0.6327	0.8187	6	0.05
ZINC000755497029	C=CCNC(=S)NN=Cc1nc2cc(Cl)ccc2n1C	0.6719	0.8186	5	0.06
ZINC000746495682	FC(F)(F)CNC(=S)NN=Cc1cn(Cc2ccccc2)c2ccccc12	0.5690	0.8186	15	0.07
ZINC000005719506	CN(/N=C/c1ccc(Cl)cc1)C(=S)c1ccccc1	0.6818	0.8178	4	0.05
ZINC000002149503	S=C(NCc1ccccc1)N/N=C/c1cn(CCOc2ccc(Br)cc2)c2ccccc12	0.5625	0.8175	13	0.21

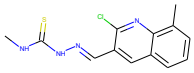
Table 2. Top-ranked molecules by ChemAI. All compounds have a high activity predicted on all four assays (column "score") and are relatively distant (column "dist") to current known inhibitors. The distance measure is the Jaccard distance based on binary ECFP4 fingerprints and resides in the interval $[0, 1]$. Some of the presented molecules might exhibit a number of toxic effects (column "tox"). Here the number of models indicating a toxic effect is reported, where the total number of toxicity models was 75. We also report the estimated probability to exhibit clinical toxicity (column "ct").



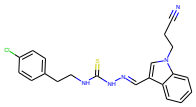
ZINC000254565785



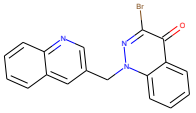
ZINC000726422572



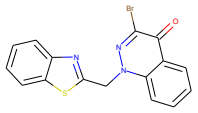
ZINC000916265995



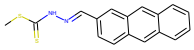
ZINC000916356873



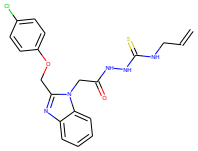
ZINC000806591744



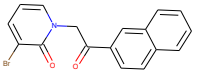
ZINC000178971373



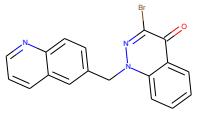
ZINC000000155607



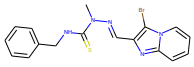
ZINC000016317677



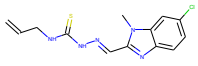
ZINC000193073749



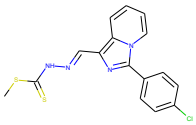
ZINC000769846795



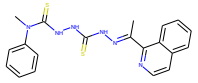
ZINC000755523869



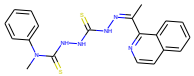
ZINC000763345954



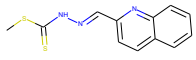
ZINC000001448699



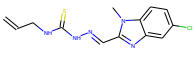
ZINC000016940508



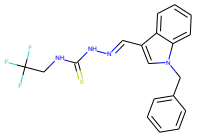
ZINC000005486767



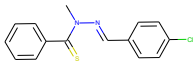
ZINC000005527649



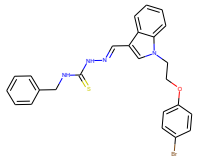
ZINC000755497029



ZINC000746495682



ZINC000005719506



ZINC000002149503

Figure 1. Graphical representation of the top-ranked molecules by ChemAI.

measurements for SARS-CoV-2 thus adjusting the multi-task model toward SARS-CoV-2.

Availability The library of molecules is available at https://github.com/ml-jku/sars-cov-inhibitors-chemai.

Acknowledgements

Funding by the Institute for Machine Learning (JKU). All authors contributed equally to this work.

References

Chen, Y. W., Yiu, C.-P. B., and Wong, K.-Y. Prediction of the sars-cov-2 (2019-ncov) 3c-like protease (3cl pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*, 9, 2020.

Collison, J. Two targets are better than one. *Nature Reviews Rheumatology*, 15(7):386–386, 2019.

Fischer, A., Sellner, M., Neranjan, S., Lill, M. A., and Smieško, M. Inhibitors for novel coronavirus protease identified by virtual screening of 687 million compounds. 2020.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

Glantz-Gashai, Y., Meirson, T., Reuveni, E., and Samson, A. O. Virtual screening for potential inhibitors of Mcl-1 conformations sampled by normal modes, molecular dynamics, and nuclear magnetic resonance. *Drug Des Devel Ther*, 11:1803–1813, 2017.

Haider, Z., Subhani, M. M., Farooq, M. A., Ishaq, M., Khalid, M., Khan, R. S. A., and Niazi, A. K. In silico discovery of novel inhibitors against main protease (mpro) of sars-cov-2 using pharmacophore and molecular docking based virtual screening from zinc database. 2020.

Huang, A., Tang, X., Wu, H., Zhang, J., Wang, W., Wang, Z., Song, L., Zhai, M.-a., Zhao, L., Yang, H., et al. Virtual screening and molecular dynamics on blockage of key drug targets as treatment for covid-19 caused by sars-cov-2. 2020.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Duan, Y., et al. Structure-based drug design, virtual screening and high-throughput screening rapidly identify antiviral leads targeting covid-19. *bioRxiv*, 2020.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

Landrum, G. RDKit: Open-source cheminformatics, 2006. URL http://www.rdkit.org.

Ledford, H. One drug, two targets, 2009.

Lim, L., Roy, A., and Song, J. Identification of a zika ns2b-ns3pro pocket susceptible to allosteric inhibition by small molecules including qucertin rich in edible plants. *bioRxiv*, 2016. doi: 10.1101/078543. URL https://www.biorxiv.org/content/early/2016/10/01/078543.

Macchiagodena, M., Pagliai, M., and Procacci, P. Inhibition of the main protease 3cl-pro of the coronavirus disease 19 via structure-based ligand design and molecular modeling. *arXiv preprint arXiv:2002.09937*, 2020.

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80, 2016.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S., and Unterthiner, T. Interpretable deep learning in drug discovery. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 331–345. Springer, 2019.

Ruan, Z., Liu, C., Guo, Y., He, Z., Huang, X., Jia, X., and Yang, T. Potential inhibitors targeting rna-dependent rna polymerase activity (nsp12) of sars-cov-2. 2020.

Senathilake, K., Samarakoon, S., and Tennekoon, K. Virtual screening of inhibitors against spike glycoprotein of 2019 novel corona virus: a drug repurposing approach. 2020.

Sterling, T. and Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Ton, A.-T., Gentile, F., Hsing, M., Ban, F., and Cherkasov, A. Rapid identification of potential inhibitors of sars-cov-2 main protease by deep docking of 1.3 billion compounds. *Molecular Informatics*, 2020.

Wang, Q., Zhao, Y., Chen, X., and Hong, A. Virtual screening of approved clinic drugs with main protease (3clpro) reveals potential inhibitory effects on sars-cov-2. 2020.

Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X., Zheng, M., Chen, L., and Li, H. Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*, 2020. ISSN 2211-3835. doi: https://doi.org/10.1016/j.apsb.2020.02.008. URL http://www.sciencedirect.com/science/article/pii/S2211383520302999.

Wu, Z., Ramsundar, B., Feinberg, E., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. ISSN 2041-6520, 2041-6539. doi: 10.1039/C7SC02664A. URL http://xlink.rsc.org/?DOI=C7SC02664A.

Zhang, J.-J., Shen, X., Yan, Y.-M., Yan, W., and Cheng, Y.-X. Discovery of anti-sars-cov-2 agents from commercially available flavor via docking screening. 2020.

Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., and Cheng, F. Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell Discovery*, 6(1):1–18, 2020.

Zhu, Z., Wang, X., Yang, Y., Zhang, X., Mu, K., Shi, Y., Peng, C., Xu, Z., et al. D3similarity: A ligand-based approach for predicting drug targets and for virtual screening of active compounds against covid-19. 2020.

ZINC ID	Trivial name(s)	Canonical SMILES	Publications
ZINC00057060	Melatonin	COc1ccc2[nH]cc(CCNC(C)=O)c2c1	Zhou et al. (2020)
ZINC03869685	Meletin, Quercetin	O=c1c(O)c(-c2ccc(O)c(O)c2)oc2cc(O)cc(O)c12	Lim et al. (2016)
ZINC85537142	Aclarubicin	CC[C@@]1(O)C[C@H](O[C@H]2C[C@H](N(C)C)[C@H](O[C@H]3C[C@H](O)[C@H](O[C@H]4CCC(=O)[C@H](C)O4)[C@H](C)O3)[C@H](C)O2)c2c(cc3c(c2O)C(=O)c2c(O)cccc2C3=O)[C@H]1C(=O)OC	Senathilake et al. (2020)
ZINC03794794	Mitoxantrone	C1=CC(=C2C(=C1NCCNCCO)C(=O)C3=C(C=CC(=C3C2=O)O)O)NCCNCCO	Wang et al. (2020)
ZINC01668172	-	O=C(C[n+]1ccc2ccccc2c1)c1ccc2ccc3ccccc3c2c1	Glantz-Gashai et al. (2017)
ZINC03830332	E155	C1=CC=C2C(=C1)C(=CC=C2S(=O)(=O)O)NN=C3C=C(C(=O)C(=NNC4=CC=C(C5=CC=CC=C54)S(=O)(=O)O)C3=O)CO	Senathilake et al. (2020)
ZINC14879972	Gar-936	CN(C)c1cc(NC(=O)CNC(C)(C)C)c(O)c2c1C[C@H]1C[C@H]3[C@H](N(C)C)C(O)=C(C(N)=O)C(=O)[C@@]3(O)C(O)=C1C2=O	Wu et al. (2020)
ZINC00001645	Magnolol	C=CCc1ccc(O)c(-c2cc (CC=C)ccc2O)c1	Wu et al. (2020)
ZINC00014036	Piceatannol	Oc1cc(O)cc(/C=C/c2ccc(O)c(O)c2)c1	Wu et al. (2020)
ZINC16052277	Doxycycline	C[C@H]1c2cccc(O)c2C(=O)C2=C(O)[C@]3(O)C(=O)C(C(=N)O)=C(O)[C@@H](N(C)C)[C@@H]3[C@@H](O)[C@@H]21	Wu et al. (2020)
ZINC3920266	Idarubicin	CC(=O)[C@]1(O)Cc2c(O)c3c(c(O)c2[C@@H](O[C@H]2C[C@H](N)[C@H](O)[C@H](C)O2)C1)C(=O)c1ccccc1C3=O	Wu et al. (2020)

Table 3. Compounds suggested in related publications for potential activity against SARS-CoV-2 and which also exhibit at least mild predicted activity against SARS-CoV proteases by ChemAI.