Supporting Information: Repositioning of 8565 Existing Drugs for COVID-19

Kaifu Gao,[†] Duc Duy Nguyen,[‡] Jiahui Chen,[†] Rui Wang,[†] and Guo-Wei Wei[*,†,¶,§]

[†]*Department of Mathematics, Michigan State University, MI 48824, USA*

[‡]*Department of Mathematics, University of Kentucky, KY 40506, USA*

[¶]*Department of Biochemistry and Molecular Biology Michigan State University, MI 48824, USA*

[§]*Department of Electrical and Computer Engineering Michigan State University, MI 48824, USA*

E-mail: wei@math.msu.edu

Supporting analyses and models

Main protease sequence identity and 3D structure similarity analysis

The sequence identity is defined as the percentage of characters that match exactly between two different sequences. Calculated by SWISS-MODEL,[S1] the sequence identities between the SARS-CoV-2 3CL protease and that of SARS-CoV, MERS-CoV, HKU-1, OC43, HCoVNL63, 229E, and HIV are 96.1%, 52.0%, 49.0%, 48.4%, 45.2%, 41.9%, and 23.7%, respectively.

It is seen that the SARS-CoV-2 3CL protease is very close to the SARS-CoV 3CL protease, but distinguished from other proteases. SARS-CoV-2 has a strong genetic relationship

with SARS-CoV, the sequence alignment in Figure S1 further confirms their relationship.

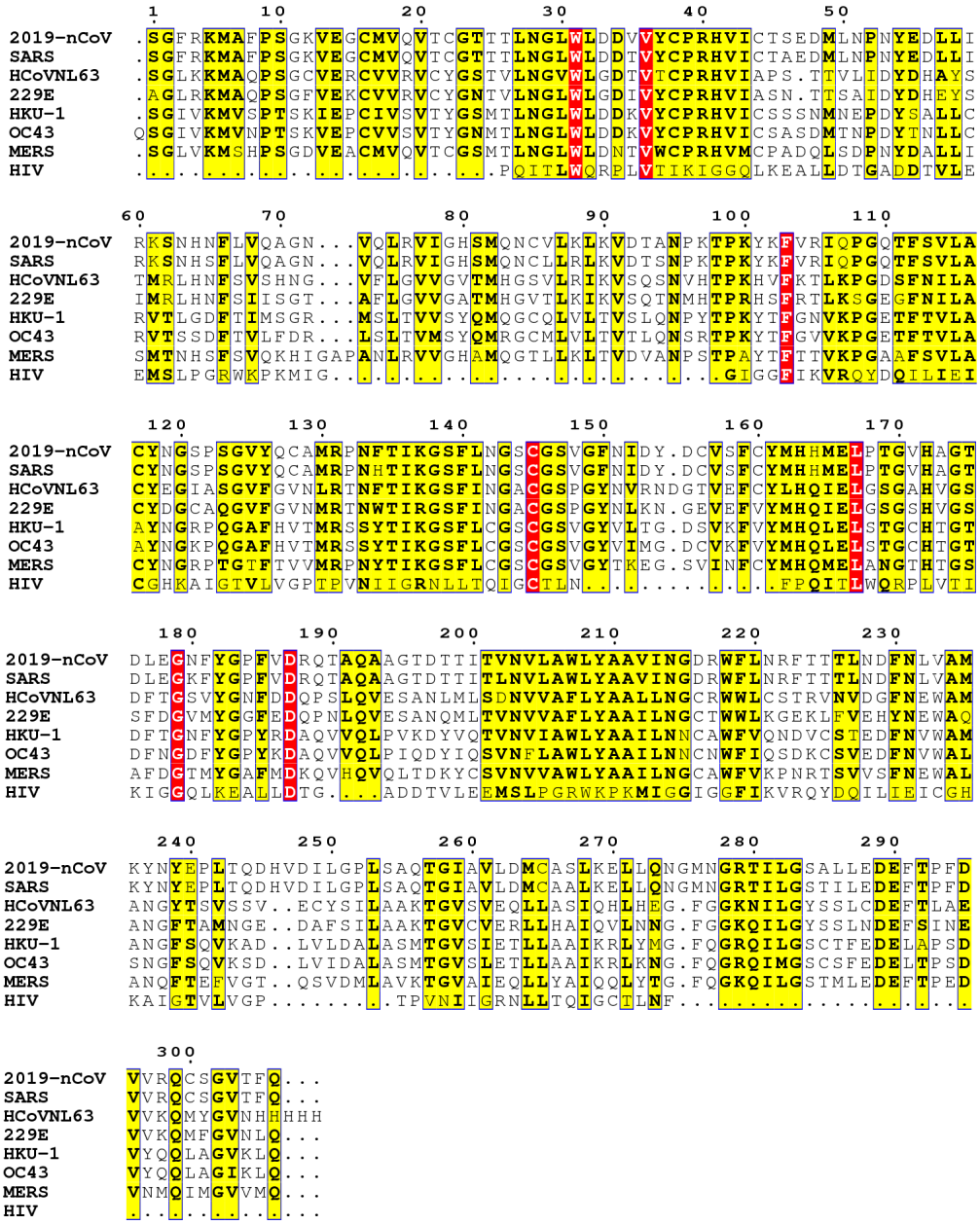

Figure S1: The 3CL protease sequence alignment between SARS-CoV-2, SARS-CoV, MERS, OC43, HCoVNL63, HKU-1, 229E, and HIV.

Not only are the sequences highly identical, but also, as shown in Fig. S2 the 3D crystal structures of the SARS-CoV-2 3CL protease is also substantially similar to that of SARS-CoV 3CL protease. Particularly, the RMSD of two structures at the binding site is only 0.42 Å.

The high sequence and structure similarity between the two proteases suggests that anti-SARS-CoV chemicals can be equally effective for the treatment of SARS-CoV-2. It means the available experimental data of SARS-CoV protease inhibitors can also be used as the training set to discover new inhibitors of SARS-CoV-2 protease. Our SARS-CoV-2 BA training set contains 314 compounds with their binding affinities to the SARS-CoV or SARS-CoV-2 3CL protease from single-protein experiments available.

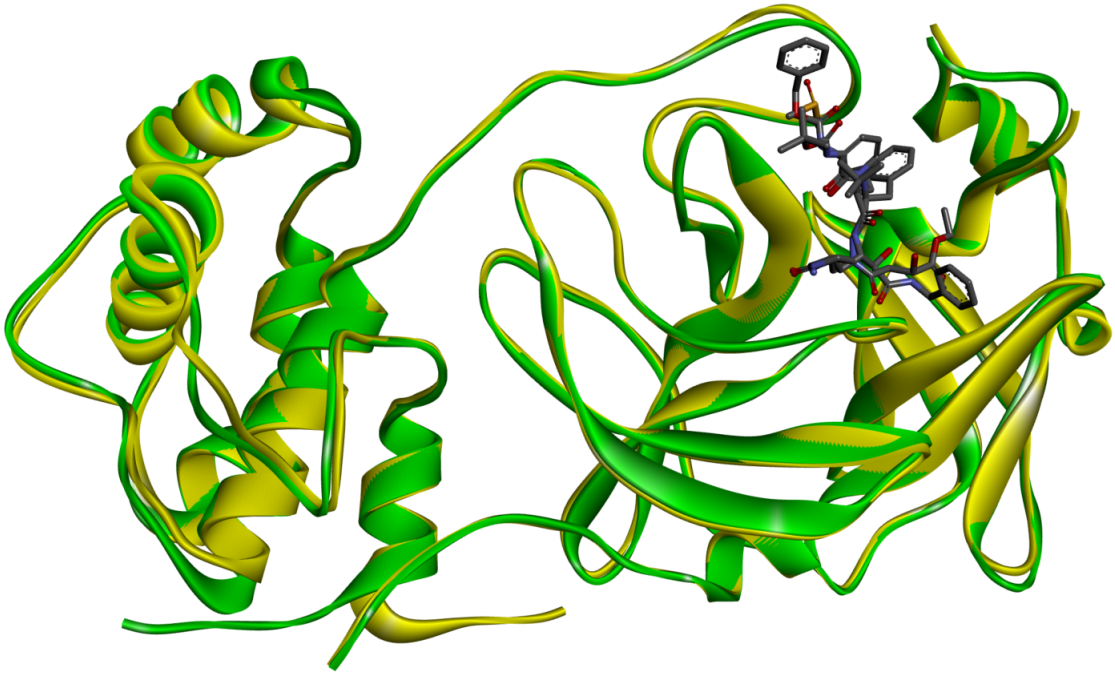

Figure S2: Illustration of the 3D structure similarity between the SARS-CoV-2 3CL protease (PDB ID: 6Y2F, in gold) and SARS-CoV 3CL protease (PDB ID: 2A5I, in green). The anti-SARS inhibitors in dark color indicate the binding site.

The experimental ΔG distribution of the training set

The binding affinities of the SARS-CoV-s main protease inhibitor training set range from -3.68 kcal/mol to -11.08 kcal/mol. Their distribution is depicted in Figure S3.

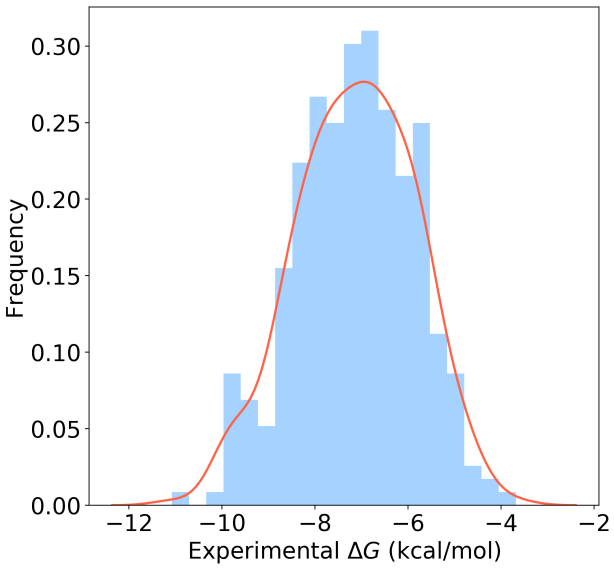

Figure S3: The experimental ΔG distribution of the training set of SARS-CoV-2 3CL protease inhibitors.

The hyperparameters in the machine learning-based binding affinity predictor

The GBDT parameters in our predictor are, for ECFP4, n_estimators=10000,max_depth=7, min_samples_split=3,learning_rate=0.01, subsample=0.3,max_features='sqrt';

for Estate1 and Estate2, n_estimators=2000,max_depth=9,min_samples_split=3, learning_rate=0.01,subsample=0.3,max_features='sqrt'.

The 10-fold cross-validation of the binding affinity predictor

The 10-fold cross-validation is carried out using 50 random splittings. Results in terms of Pearson correlation coefficient (R_p) the Kendall's τ (τ), and RMSE are given in Table S1.

MathDL

MathDL, designed for predicting various druggable properties of 3D molecules,[S2] is capable of efficiently and accurately encoding the high-dimensional biomolecular interactions into low-dimensional representations. Algebraic graph theory,[S3] differential geometry,[S4]

Table S1: The 10-fold cross-validation test of the machine learning model on the SARS-CoV-2 BA training set.

	R_p	τ	RMSE (kcal/mol)		R_p	τ	RMSE (kcal/mol)
Fold 1 (Train)	0.997	0.972	0.095	Fold 6 (Train)	0.997	0.972	0.096
Fold 1 (Test)	0.794	0.600	0.778	Fold 6 (Test)	0.777	0.582	0.777
Fold 2 (Train)	0.997	0.972	0.095	Fold 7 (Train)	0.997	9.971	0.096
Fold 2 (Test)	0.759	0.571	0.818	Fold 7 (Test)	0.781	0.588	0.780
Fold 3 (Train)	0.997	0.971	0.095	Fold 8 (Train)	0.997	0.971	0.096
Fold 3 (Test)	0.791	0.607	0.783	Fold 8 (Test)	0.770	0.576	0.811
Fold 4 (Train)	0.997	0.971	0.096	Fold 9 (Train)	0.997	0.972	0.094
Fold 4 (Test)	0.780	0.579	0.781	Fold 9 (Test)	0.762	0.578	0.801
Fold 5 (Train)	0.997	0.972	0.095	Fold 10 (Train)	0.997	0.972	0.095
Fold 5 (Test)	0.782	0.591	0.789	Fold 10 (Test)	0.770	0.585	0.811
Average (Train)	0.997	0.972	0.095				
Average (Test)	0.777	0.586	0.792				

and algebraic topology methods[S2] are applied to generate three mathematical representations of data in MathDL. These data representations can be integrated with well-designed deep learning models, such as gradient-boosted trees (GBTs) and convolutional neural networks (CNNs), for pose ranking and binding affinity predictions. In D3R Grand Challenges (https://drugdesigndata.org/about/grand-challenge), a worldwide competition series in computer-aided drug design, MathDL had been proved as the top competitor in free energy prediction and ranking in the past three years.[S2,S5] Figure S4 illustrates the framework of the MathDL model, which combined the aforementioned mathematical representations with the CNN architecture for druggable properties predictions. The PDBbind 2018 general set,[S6] along with the SARS 3CL protease related dataset is used in our training process. To address the reliability of the MathDL model, we did the 10-fold cross-validation on the various PDBbind refine sets with the average Pearson correlation coefficients and the root mean square error (RMSE) being 0.771 and 1.78 kcal/mol, respectively.[S7]

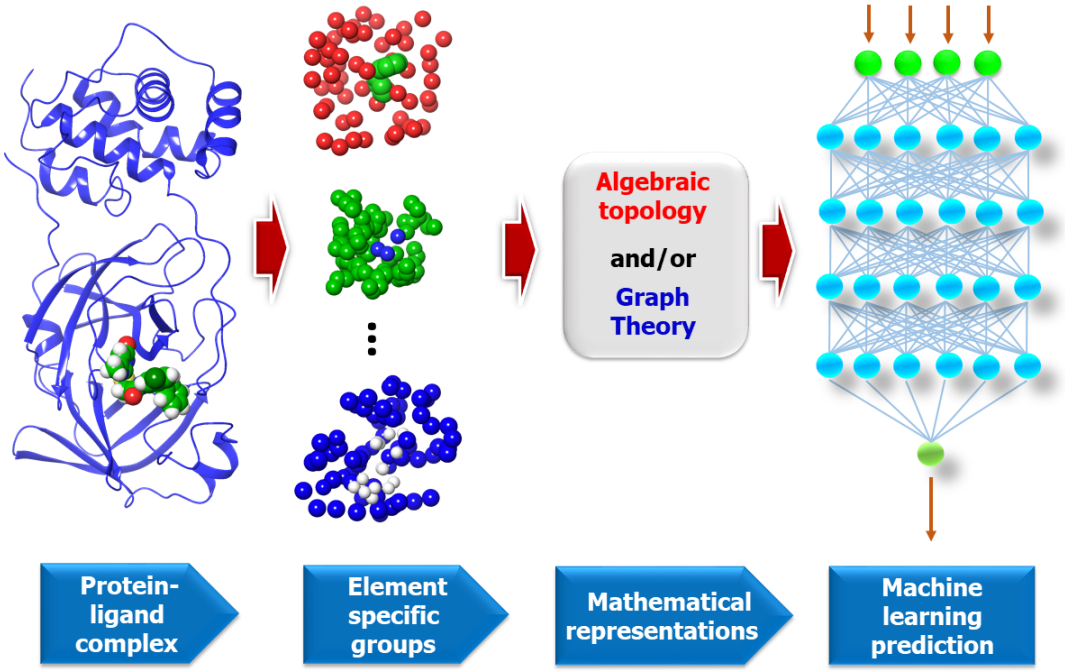

Figure S4: A framework of MathDL energy prediction model which integrates advanced mathematical representations with sophisticated CNN architectures

Nonpolar binding site residues in SARS-CoV-2 Main protease

In the main protease's binding site, there are some nonpolar residues, namely Ala193, Gly143, Leu27, Leu141, Met165, Met49, Phe140, and Pro168.

Supporting tables

Supporting tables are available in SupportingTables.xlsx for follows.

The table of experimental binding affinities, predicted synthesizability scores, predicted log P, and predicted log S for 314 SARS-CoV-2 3CL protease inhibitors in the training set

The table of the predicted binding affinities, predicted synthesizability scores, predicted log P, and predicted log S for 1553 FDA-approved drugs

The table of the predicted binding affinities for 7012 investigational or off-market drugs

References

(S1) Bienert, S.; Waterhouse, A.; de Beer, T. A.; Tauriello, G.; Studer, G.; Bordoli, L.; Schwede, T. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.* **2017**, *45*, D313–D319.

(S2) Nguyen, D. D.; Gao, K.; Wang, M.; Wei, G.-W. Mathdl: Mathematical deep learning for d3r grand challenge 4. *J. Comput. Aided Mol. Des.* **2020**, *34*, 131–147.

(S3) Nguyen, D. D.; Wei, G.-W. AGL-Score: Algebraic Graph Learning Score for Protein–Ligand Binding Scoring, Ranking, Docking, and Screening. *J. Chem. Inf. Model.* **2019**, *59*, 3291–3304.

(S4) Nguyen, D. D.; Wei, G.-W. DG-GL: Differential geometry-based geometric learning of molecular datasets. *Int. J. Numer. Method. Biomed. Eng.* **2019**, *35*, e3179.

(S5) Nguyen, D. D.; Cang, Z.; Wu, K.; Wang, M.; Cao, Y.; Wei, G.-W. Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *J. Comput. Aided Mol. Des.* **2019**, *33*, 71–82.

(S6) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: The CASF-2016 update. *J. Chem. Inf. Model.* **2018**, *59*, 895–913.

(S7) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **2018**, *14*, e1005929.