Turning high-throughput structural biology into predictive inhibitor design – Supplementary Information

Kadi L. Saar,[a,b] William McCorkindale,[a] Daren Fearon,[c] Melissa Boby,[d] Haim Barr,[e] Amir Ben-Shmuel,[f] The COVID Moonshot Consortium, Nir London,[g] Frank von Delft,[c,h,i,j] John D. Chodera[d] & Alpha A. Lee[a,k,1]

[a] *Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK*

[b] *Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, UK*

[c] *Diamond Light Source Ltd., Harwell Science and Innovation Campus, Didcot, UK*

[d] *Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA*

[e] *Wohl Institute for Drug Discovery of the Nancy and Stephen Grand Israel National Center for Personalized Medicine, The Weizmann Institute of Science, Rehovot, 7610001, Israel*

[f] *Israel Institution of Biological Research, Ness-Ziona, Israel*

[g] *Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel*

[h] *Centre for Medicines Discovery, University of Oxford, Oxford, UK*

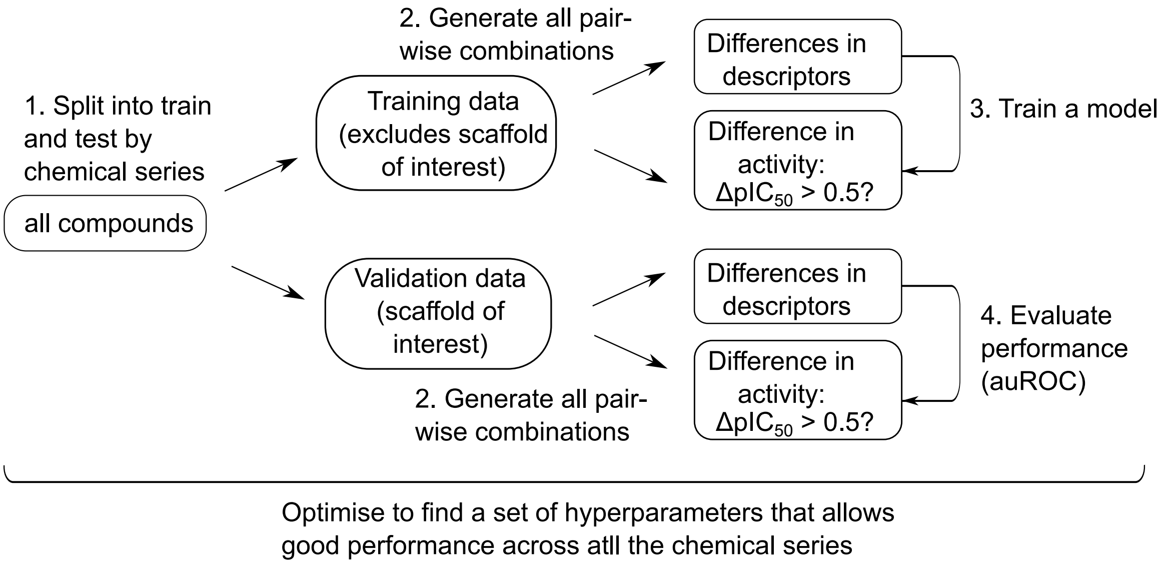
[i] *Faculty of Science, University of Johannesburg, Johannesburg, South Africa*

[j] *Research Complex at Harwell, Harwell Science and Innovation Campus, Didcot, OX11 0FA, UK*

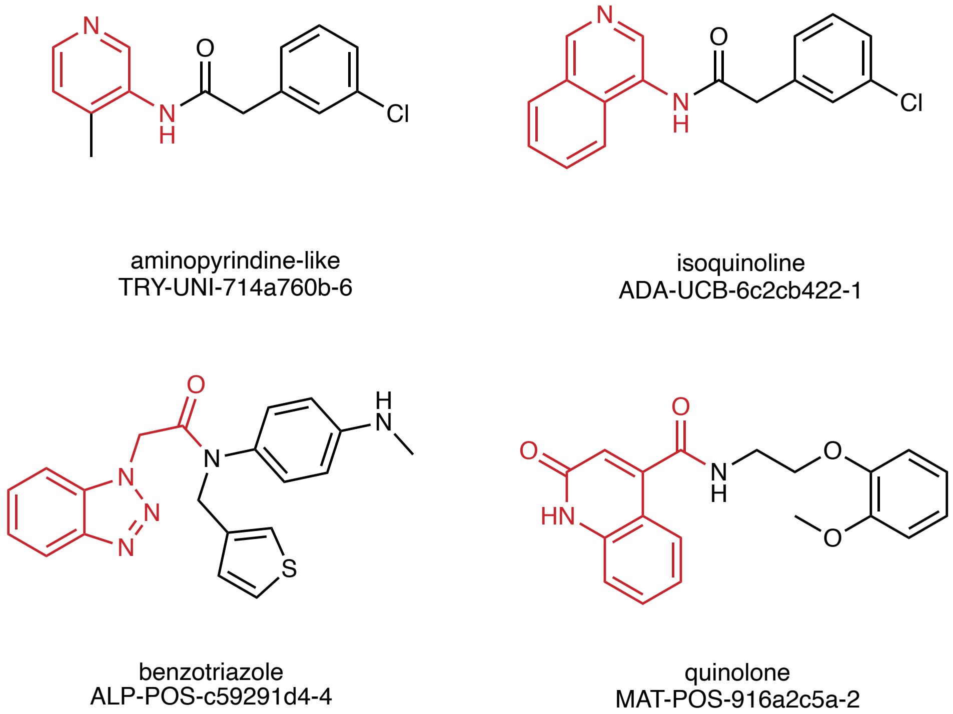
[k] *PostEra Inc, 2 Embarcadero Center, San Francisco, CA 94111, USA*

[*] Full consortium membership available at: `https://tinyurl.com/y3r7redd`

[1] To whom correspondence should be addressed: aal44@cam.ac.uk


1. Split into train and test by chemical series
all compounds
2. Generate all pair-wise combinations
Training data (excludes scaffold of interest)
Differences in descriptors
Difference in activity: ΔpIC50 > 0.5?
3. Train a model
Validation data (scaffold of interest)
2. Generate all pair-wise combinations
Differences in descriptors
Difference in activity: ΔpIC50 > 0.5?
4. Evaluate performance (auROC)
Optimise to find a set of hyperparameters that allows good performance across atll the chemical series

Supplementary Figure S1. Overview of the model development process. 1. The data was split into a training and a validation set in a scaffold-split manner, each time excluding one of the chemical series (Supplementary Figure S2) from the training set. 2. Pair-wise differences between the descriptors of the protein-ligand structure and the measured activity value were estimated for all pairs within the training and the validation sets. 3. A model was trained to learn if two compounds differed by more than a half pIC_{50} unit as a function of the differential fingerprint. 4. The performance of the model was quantified on the left-out set using the area under the receiver-operator characteristic curve (auROC) as a metric. The process was repeated across all the four distinct chemical series and the set of hyperparameters that allowed for a good performance (highest mean auROC value) across all the four chemical series was chosen as the optimal model.



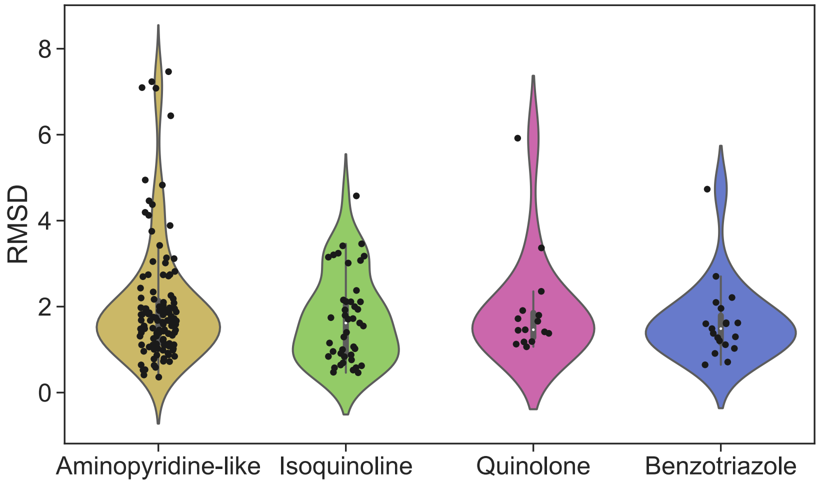
aminopyrindine-like
TRY-UNI-714a760b-6

isoquinoline
ADA-UCB-6c2cb422-1

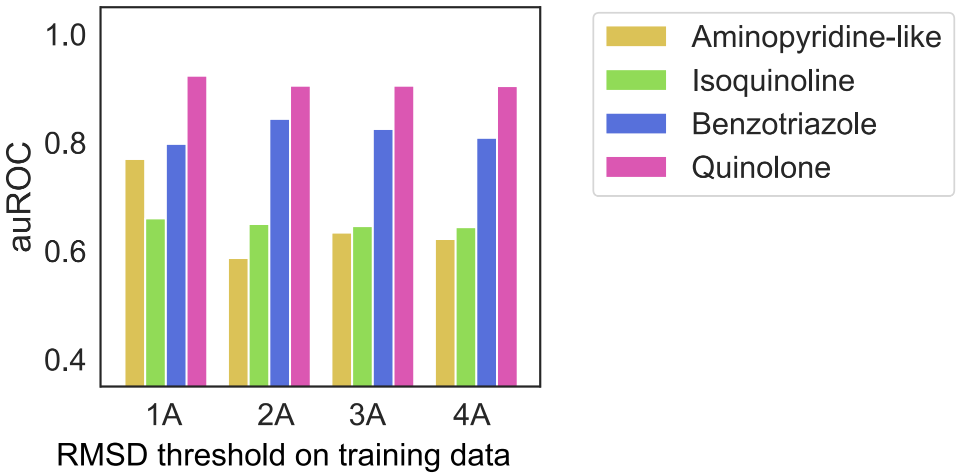
benzotriazole
ALP-POS-c59291d4-4

quinolone
MAT-POS-916a2c5a-2

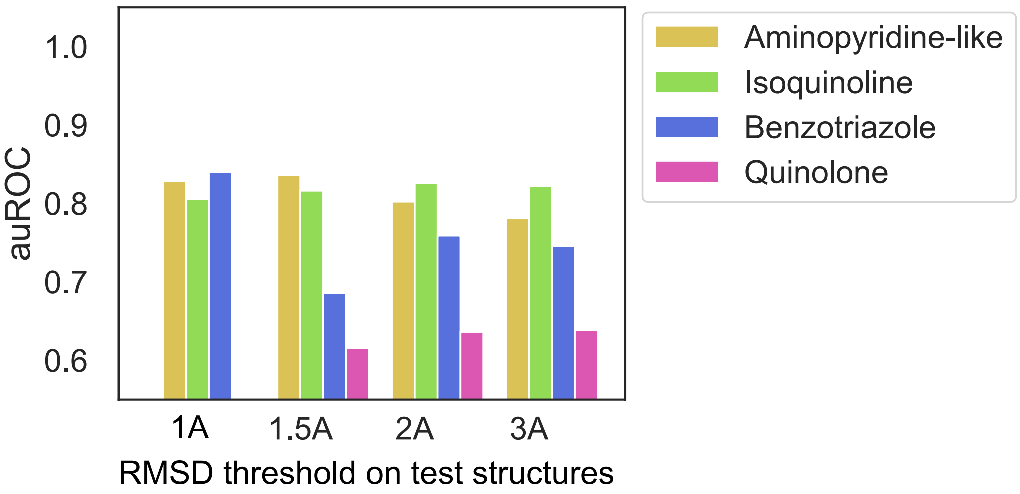
Supplementary Figure S2. Division of the molecules into chemical series. Representative example from each of the four chemical series with the salient chemical motif highlighted in red.



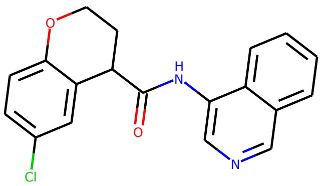
Supplementary Figure S3. The distribution of heavy atom root-mean squared distance (RMSD) between the docked structures and the experimental determined structures for the four chemical series.



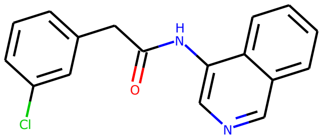
Supplementary Figure S4. auROC value achieved with the docking-based model when training data is restricted to the cases where the heavy atom root mean squared displacement (RMSD) between the crystal and the docked structure is below a specified threshold. The number of structures with heavy atom RMSD values below each threshold is shown in Supplementary Figure S3.



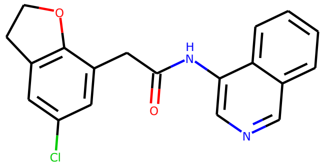
Supplementary Figure S5. auROC value for the structure-based learning approach with the validation data being restricted to ligands for which the root mean square deviation (RMSD) between the crystal and the experimental structure is below a specific threshold.



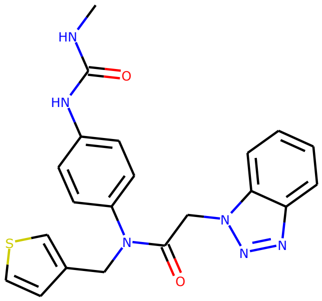
MAT-POS-bbbbc21a-3;
IC50 = 1500 nM



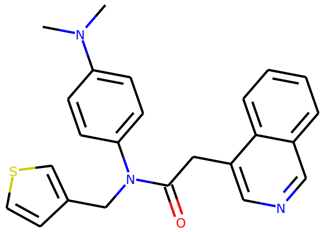
ADA-UCB-6c2cb422-1;
IC50 = 700 nM



VLA-UCB-1dbca3b4-15;
IC50 = 360 M



PET-UNK-1901c25b-1;
IC50 = 280 nM



ROB-IMP-e811baff-1;
IC50 = 1420 nM

Supplementary Figure S6. Structures of the five highly potent non-covalent compounds from the COVID moonshot campaign that were used as the reference when evaluating the relative rankings of each of the compounds in the virtual library.

Supplementary Table S1. Area under the receiver operator characteristic curve (auROC) for the learn-to-rank task on SARS-CoV-2 MPro binding. Each column shows performance for the case when the specific chemical series is left out as test data. The values correspond to the mean and standard error of the mean when a 10-fold boostrapping process was performed on the test data. All models were built using Autodock Vina descriptors using the structure-based learning approach where experimental crystal structures were used for the training the model and docked structures

auROC	**Aminopyridine-like (n = 123)**	**Isoquinoline (n = 44)**	**Benzotriazole (n = 19)**	**Quinolone (n = 15)**	**Mean**
Logistic regression	0.76 ± 0.04	0.73 ± 0.01	0.82 ± 0.02	0.73 ± 0.02	0.76 ± 0.02
Extra Tree classifier	0.71 ± 0.04	0.68 ± 0.02	0.82 ± 0.01	0.72 ± 0.04	0.73 ± 0.03
Random forest	0.78 ± 0.01	0.73 ± 0.01	0.81 ± 0.01	0.79 ± 0.02	0.78 ± 0.02
k-nearest neighors	0.71 ± 0.05	0.70 ± 0.01	0.81 ± 0.01	0.72 ± 0.03	0.73 ± 0.03

Supplementary Table S2. Same data as in Supplementary Table S1 with the performance evaluated using area under the precision-recall curve (auPRC) as the performance metric.

auPRC	Aminopyridine-like (n = 123)	Isoquinoline (n = 44)	Benzotriazole (n = 19)	Quinolone (n = 15)	Mean
Logistic regression	0.66 ± 0.06	0.70 ± 0.01	0.82 ± 0.01	0.72 ± 0.03	0.73 ± 0.03
Extra Tree classifier	0.60 ± 0.04	0.67 ± 0.02	0.81 ± 0.02	0.70 ± 0.05	0.70 ± 0.04
Random forest	0.73 ± 0.02	0.70 ± 0.01	0.82 ± 0.02	0.77 ± 0.02	0.76 ± 0.03
k-nearest neighors	0.74 ± 0.07	0.73 ± 0.01	0.84 ± 0.03	0.78 ± 0.02	0.77 ± 0.03

Supplementary Table S3. Hyperparameters considered for the different model architectures.

Model	**Hyperparameters**

Logistic regression	penalty: { l_1, l_2 }
	C: {0.001, 0.001, 0.01, 1, 10}
	solver: {lbfgs, liblinear}

Extra tree classifier	$n_{estimators}$: { 20, 50, 80 100, 150, 200, 400}
	max depth: {3, 4, 5, 6, 7}
	min samples split: {1, 2, 4}
	criterion: {gini, entropy, logloss}

Random forest	$n_{estimators}$: {20, 50, 100, 200, 500}
	max depth: {2, 3, 4, 5, 6, 7}
	min samples leaf: {1, 2, 3, 4}
	max features: {sqrt, auto }
	k-nearest neighbors $n_{neighbors}$: {2, 5, 10, 30, 80, 200, 500}
	weights: {uniform, distance

Supplementary Table S4. The performances of structure-based modelling strategy in comparison to the ligand-based and docking-based learning strategies for each of the four scaffolds as quantified by auROC value. The values correspond to the average and the standard deviation of the auROC scores on the left-out data using a 10-fold bootstrapping process.

auROC	**Aminopyridine-like (n = 123)**	**Isoquinoline (n = 44)**	**Benzotriazole (n = 19)**	**Quinolone (n = 15)**	**Mean**
Ligand-based learning	0.64 ± 0.05	0.46 ± 0.07	0.42 ± 0.01	0.56 ± 0.10	0.76 ± 0.02
Docking-based learning	0.51 ± 0.09	0.62 ± 0.02	0.79 ± 0.02	0.82 ± 0.03	0.73 ± 0.03
Docking with structure-based learning	0.80 ± 0.01	0.71 ± 0.01	0.81 ± 0.01	0.78 ± 0.02	0.78 ± 0.02