

1 **Machine Learning Models Identify Inhibitors of SARS-CoV-2**

2

3 Victor O. Gawriljuk¹, Phyto Phyto Kyaw Zin², Daniel H. Foil², Jean Bernatchez³, Sungjun
4 Beck³, Nathan Beutler⁴, James Ricketts⁴, Linlin Yang⁴, Thomas Rogers^{4,5}, Ana C. Puhl²,
5 Kimberley M. Zorn², Thomas R. Lane², Andre S. Godoy¹, Glaucius Oliva¹, Jair L.
6 Siqueira-Neto³, Peter B. Madrid⁶ and Sean Ekins^{2*}

7

8 ¹São Carlos Institute of Physics, University of São Paulo, Av. João Dagnone, 1100 -
9 Santa Angelina, São Carlos - SP, 13563-120, Brazil

10

11 ²Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC
12 27606, USA.

13

14 ³Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California
15 San Diego, La Jolla, California, 92093, USA.

16

17 ⁴The Scripps Research Institute, La Jolla, California, 92093, USA.

18

19 ⁵School of Medicine, University of California San Diego, La Jolla, California, 92093,
20 USA.

21

22 ⁶SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA.

23 *Corresponding Author sean@collaborationspharma.com

24 **Key words:** SARS-CoV-2, COVID-19, Machine learning, drug discovery, Bayesian

25

26 **Abstract**

27

28 With the ongoing SARS-CoV-2 pandemic there is an urgent need for the
29 discovery of a treatment for the coronavirus disease (COVID-19). Drug repurposing is
30 one of the most rapid strategies for addressing this need and numerous compounds
31 have been selected for *in vitro* testing by several groups already. These have led to a
32 growing database of molecules with *in vitro* activity against the virus. Machine learning
33 models can assist drug discovery through prediction of the best compounds based on
34 previously published data. Herein we have implemented several machine learning
35 methods to develop predictive models from recent SARS-CoV-2 *in vitro* inhibition data
36 and used them to prioritize additional FDA approved compounds for *in vitro* testing
37 selected from our in-house compound library. From the compounds predicted with a
38 Bayesian machine learning model, CPI1062 and CPI1155 showed antiviral activity in
39 HeLa-ACE2 cell-based assays and represent potential repurposing opportunities for
40 COVID-19. This approach can be greatly expanded to exhaustively virtually screen
41 available molecules with predicted activity against this virus as well as a prioritization
42 tool for SARS-CoV-2 antiviral drug discovery programs. The very latest model for
43 SARS-CoV-2 is available at www.assaycentral.org.

44

45

46

47

48 **Introduction**

49 In December 2019, several cases of pneumonia with unknown etiology started to
50 arise in Wuhan, China. A new betacoronavirus was identified and named SARS-CoV-2
51 due to high similarity with previous SARS-CoV^{1,2}. This virus causes the disease which
52 has been called COVID-19³. Since then, SARS-CoV-2 has rapidly spread worldwide
53 prompting the World Health Organization to declare the outbreak a pandemic, with more
54 than 1.5 million cases confirmed in less than 100 days.⁴ The high infection rate has also
55 caused considerable stress on global healthcare systems leading to more than 400,000
56 deaths from COVID-19.

57 The SARS-CoV-2 pandemic started a worldwide effort to discover a treatment
58 that could prevent further COVID-19 deaths and decrease the number and length of
59 hospitalization⁵. Drug repurposing is one of the main strategies being used to accelerate
60 this as most preclinical stages are removed and a promising drug could move directly
61 into phase II clinical studies or beyond by using an approved, safe drug^{6,7}. So far, most
62 SARS-CoV-2 inhibition studies rely on small to medium scale assays with high
63 throughput screens (HTS) campaigns testing specific FDA-approved drugs and
64 compounds that have previously shown inhibition against different betacoronaviruses or
65 specific antiviral targets⁸⁻¹⁶.

66 Quantitative Structure Activity Relationship (QSAR) analyses from previous *in*
67 *vitro* data has been widely used to assist drug discovery in both industry and
68 academia¹⁷. In the past few years the rise of machine learning has also expanded to
69 drug discovery, with different methods being implemented in a wide range of areas from
70 predicting synthetic routes to biological activity^{18,19}. Many examples show that

71 prioritizing compounds from machine learning and QSAR models can increase the
72 success rate and save resources¹⁷. Here we have implemented several machine
73 learning methods to develop predictive models from recent SARS-CoV-2 *in vitro*
74 inhibition data and used them to prioritize compounds for *in vitro* testing of different
75 compound libraries. These efforts will add to the list of >200 drugs and vaccines under
76 assessment elsewhere and which is continually growing²⁰.

77

78 **Materials and Methods**

79

80 **Data Curation**

81 Data from recent drug repurposing campaigns for SARS-CoV-2 were used to
82 build a dataset from whole cell inhibition assays^{8,9,12,14,15}. In assays with several
83 Multiplicity of Infection (MOI) the one closer to the whole dataset was chosen. In
84 machine learning model generation, duplicate compounds with finite activities are
85 averaged into a single entry. Due to the potential for diminished activity, when duplicate
86 compounds were present, only the most active one was retained in the dataset.
87 Additionally, compounds with ambiguous dose-response curves were discarded.
88 Datasets were built with Molecular Notebook (Molecular Materials Informatics, Inc). In
89 order to evaluate the model performance on an external testing set, a total of 30
90 molecules was collated from different studies^{11,21-25}.

91

92 **Assay Central™**

93 The Assay Central™ software (AC) has been previously described^{19,26-34}. AC
94 employs a series of rules for the detection of problem data for automated structure
95 standardization to generate high-quality data sets and Bayesian machine learning
96 models capable of predicting potential bioactivity for proposed compounds. AC was
97 used to prepare and merge data sets, as well as generate Bayesian models using the
98 ECFP6 descriptor and five-fold cross validation. During model generation, training
99 compounds are standardized (i.e. salts were removed, corresponding acids
100 neutralized), and thresholds for binary activity classification are applied to optimize
101 internal five-fold cross validation metrics. For predictions, AC workflows assign a
102 probability score and applicability score to prospective compounds according to a user-
103 specified model, with prediction scores greater than 0.5 considered active.

104

105 **Additional Machine Learning Methods**

106 Additional Machine learning algorithms including Bernoulli Naïve Bayes (bnb),
107 AdaBoost Decision trees (ada), Random Forest (rf), support vector classification (svc),
108 k-Nearest Neighbors (knn) and Deep Learning (DL) were also implemented with ECFP6
109 fingerprints and five-fold cross validation. Details for the development of these models
110 was previously described in detail in our earlier articles^{28,32,35}. Bayesian models were
111 also generated with Discovery Studio (Biovia, San Diego CA) using ECFP6 descriptors
112 where the top and bottom scoring fingerprints were selected for qualitative comparison.

113

114 **Model Performance**

115 Machine learning model performance was evaluated with different metrics:
116 accuracy, recall, precision, specificity, F1-score, area under receiver operating
117 characteristic curve, Cohen's kappa, and the Matthews correlation coefficient. The
118 statistics were calculated for both training data with five-fold cross validation, to evaluate
119 training performance, as well as in external testing set, to evaluate model performance
120 in predicting data outside the training set.

121

122 **Principal Component Analysis**

123 Principal Component Analysis (PCA) was computed for both the SARS-CoV-2
124 data set as well as SARS-CoV-2 with different compound libraries to assess its
125 chemical space. The scikit-learn³⁶ (0.22.2) PCA algorithm was used to reduce feature
126 dimensionality to three using different molecular descriptors (MW, MolLogP, NR, NArR,
127 NRB, HBA, HBD) and also with EFCP6 fingerprints. Molecular descriptors and
128 fingerprints were generated from the cheminformatics library RDkit (2020.03.1).

129

130 **Applicability and Reliability Domain Assessment**

131 In order to check if it is valid to apply the model for compounds being predicted
132 and how reliable the predictions are, an applicability and reliability domain assessment
133 was performed. First, the compound applicability within the model is assessed
134 comparing its similarity with the model's data using both molecular and fingerprint
135 descriptors. If the molecule satisfies both criteria it is considered within the applicability
136 domain and goes to the reliability domain assessment.

137 The first criterion for the applicability assessment is determined based on
138 whether it fits within the range of the key molecular descriptors of the training set (MW,
139 MolLogP, NRB, TPSA, HBA, HBD). If at least four properties lie within the maximum
140 and minimum values of the model's data, the molecule is considered similar and goes to
141 the next criterion. The second criterion relies on structural fragment-based similarity
142 measured with Tanimoto coefficient using MACCS fingerprints. The similarity of the
143 MACCS fingerprints for the query compound and all training data is computed using the
144 Tanimoto score. Only 5% of the training set compounds that are most similar to the
145 query compound is used for evaluation (i.e. if the training set has 100 molecules only 5
146 molecules with more similarity to the query compound are used for the next evaluation).
147 If the Tanimoto score exceeds 0.5 against the 5% of the training set compounds, the
148 model is considered to have enough structural fragments overlap with the query
149 compound and thus the compound goes onto the reliability assessment.

150 The reliability domain assessment implements k-means clustering methods
151 based on ECFC6 fingerprints to classify the predictions from very high to low reliability.
152 The reliability class depends on four criteria: distance from the major central point of the
153 training data, distance from the closest cluster, closest cluster density and closest
154 cluster distance within the chemical space. Each criterion has different weights and
155 scores, with the second and third having higher priority. If the compound scores 1 in
156 each criterion it is classified as very highly reliable, if that is not the case only the two
157 higher priority criteria are considered for the next classes. The compound is classified
158 as highly reliable if scores a total of 2, moderately reliable if it scores between -1 and 2
159 or low reliability if it scores less than or equal to -1 in the two higher priority criteria. The

160 scores for each criterion as well as its definition are extensively described in the
161 Supplemental Methods.

162

163 ***In vitro* testing**

164 Compounds were tested in a 10-point serial dilution experiment to determine the 50%
165 inhibitory concentration (IC_{50}) and 50% cytotoxicity concentration (CC_{50}). 1,000 HeLa-
166 ACE2 cells/well were added into 384-well plates with compounds in a volume of 25 nL.
167 The final concentrations of compound ranged from 78nM-40 M. 4 h post seeding 500
168 pfu SARS-CoV-2 (Washington strain USA-WA1/2020), BEI Resources NR-52281 were
169 added to each well at a MOI = 0.5. Twenty-four hours post infection cells were fixed with
170 4% formaldehyde solution. The cells were then treated with a Primary ab: human
171 polyclonal plasma (COVID-19 patient); Secondary ab: goat anti-human IgG coupled
172 with HRP. Images were acquired with ImageXpress MicroXL (bright field); Custom
173 Module developed in MetaXpress was used for automated count of total cells and
174 infected cells. Antiviral activity was assessed based on the infection ratio (number of
175 infected cells/total number of cells) in comparison with the average infection ration of
176 the untreated controls.

177

178 **Results**

179 **Data Curation**

180 *In vitro* SARS-CoV-2 data was initially collated from five drug repurposing studies
181 leading to a data set of 63 molecules with mean activity of $15.94 \pm 22.45 \mu\text{M}$ ^{8,9,12,14,15}.
182 The external testing set collated from different studies has 30 molecules and a mean

183 activity of $34 \pm 42 \text{ M}^{11,21-25}$. Most assays were performed with different Vero cell lines,
184 inhibition was measured with viral RNA quantification, cytopathogenic effects or
185 immunofluorescence methods with MOI and incubation time varying from 0.01-0.05 and
186 24-72 hrs respectively (Figure S1). The threshold set for activity classification by the
187 Bayesian model generated with AC was 6.65 M, with a final ratio of 52% actives in the
188 training set and 37% in the external test set. The molecules in both training and test set
189 are available in the supplemental data.

190

191 **Machine Learning Models**

192 Machine learning models were developed with AC as well as several other
193 methods available to us. This five-fold cross validation comparison shows the different
194 prediction statistics for all machine learning algorithms implemented with the training
195 data only (Table 1). AC outperformed all of them at the threshold of 6.65 μM with Rf
196 coming the closest. These models were chosen for further external testing predictions.

197

198 **Table 1 – Five-fold cross validation statistics for all SARS-CoV-2 machine**
199 **learning models implemented using ECFP6 fingerprints.**

	ACC	AUC	CK	MCC	Pr	Recall	Sp	F1
AC	0.81	0.78	0.62	0.64	0.78	0.88	0.73	0.83
rf	0.75	0.74	0.49	0.5	0.73	0.82	0.67	0.77
knn	0.71	0.71	0.43	0.42	0.71	0.76	0.67	0.74
svc	0.7	0.69	0.39	0.4	0.68	0.79	0.6	0.73
bnb	0.68	0.68	0.36	0.36	0.7	0.7	0.67	0.7

ada	0.64	0.63	0.27	0.26	0.65	0.67	0.6	0.66
DL	0.65	0.65	0.3	0.3	0.66	0.67	0.63	0.66

200

201 ACC: Accuracy, AUC: Area under curve, CK: Cohen's Kappa, MCC: Matthews
202 correlation coefficient, Pr: Precision, Sp: Specificity, F1: F1 Score. bnb: Bernoulli Naïve
203 Bayes, ada: AdaBoost Decision trees, rf: Random Forest, svc: support vector
204 classification, knn: k-Nearest Neighbors and DL: Deep Learning (DL)

205

206 **External Validation**

207 The performance of the machine learning models on the external testing data is
208 shown in Table 2. The external validation was used to measure model performance in
209 data from different studies outside the training set. svc and knn had slightly better
210 statistics compared to all other models, with the best balance between recall and
211 specificity.

212

213 **Table 2 – Prediction statistics with the external data for all SARS-CoV-2**
214 **machine learning models implemented**

	ACC	AUC	CK	MCC	Pr	Recall	Sp	F1
AC	0.62	0.58	0.17	0.17	0.50	0.40	0.76	0.44
rf	0.63	0.57	0.10	0.11	0.42	0.30	0.80	0.35
knn	0.67	0.6	0.21	0.21	0.50	0.40	0.80	0.44
svc	0.70	0.57	0.34	0.34	0.54	0.60	0.75	0.57
bnb	0.50	0.49	-0.09	-0.09	0.27	0.30	0.60	0.28

ada	0.53	0.49	0.00	0.00	0.33	0.40	0.60	0.36
DL	0.63	0.56	0.15	0.15	0.44	0.40	0.75	0.42

215

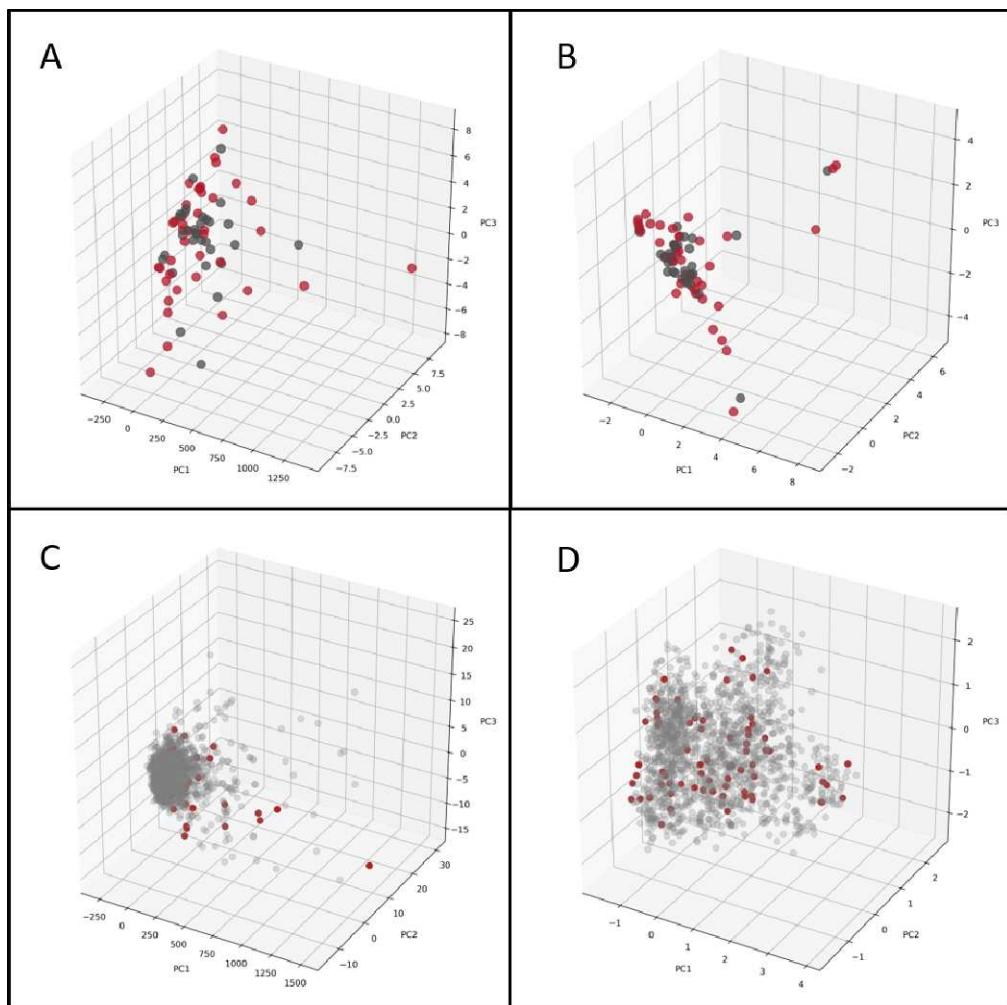
216

217 **Chemical Space**

218 The PCA of the model training set alone shows that the SARS-CoV-2 chemical
219 space is well distributed with active and inactive molecules well mixed when analyzed
220 using either molecular and fingerprint descriptors. When compared with Prestwick
221 Chemical Library (PwCL), a library of predominantly FDA approved drugs, the SARS-
222 CoV-2 data lie within a big cluster with molecular descriptors and is more widely
223 distributed when using the fingerprint descriptors.

224

225 **Figure 1 – PCA of the SARS-CoV-2 set with Molecular Descriptors (A), and**
226 **ECFP6 (B). Red Spheres – Active, Grey Spheres – Inactive. PCA of SARS-CoV-2 set**
227 **and Prestwick Chemical Library (PwCL) with molecular descriptors (C), and ECFP6 (D).**
228 **Red Spheres – SARS-CoV-2, Grey Spheres – PwCL**



229
230 **Applicability and Reliability Domain Assessment of External Test Set**
231 The applicability and reliability domain assessment of the external test set was
232 determined for each molecule as described in the methods to see how the test set
233 compares with the training data. Molecules in the applicability domain are considered
234 suitable for the model predictions due to similarity based on structural and molecular

235 properties with the training data, whereas the reliability value is a measurement of how
236 reliable the predictions are and uses different clustering metrics to determine its value.

237 From 30 molecules in the external test set, 22 were within the training data
238 applicability domain and had their reliability value calculated. Most molecules that fell
239 within the applicability domain had high or very high reliability values, with only 36%
240 showing moderate reliability, so, most molecules obey the similarity criteria and are not
241 far away from dense clusters. In comparison, with the Assay Central applicability score,
242 which accounts only for structural similarity of the query compound with the training
243 data, only 10 molecules were considered within the domain with a higher reliability,
244 suggesting it is likely more conservative. Indeed, with the external test and training set
245 PCA we can see that most molecules superimpose with few of them distant from each
246 other (Figure S1). Therefore, similarity together with clustering methods are more
247 suitable for applicability and reliability assessment compared with only structural
248 similarity, as seen by the PCA.

249

250 **Prospective Prediction**

251 A selection of FDA approved drugs available to us in our relatively small in-house
252 compound collection of hundreds of molecules was scored with the AC Bayesian model.
253 A selection of some of the best scoring molecules (Table 3) was used to identify and
254 prioritize compounds for *in vitro* testing. AC Applicability score is the similarity of the
255 compound with the training data, compounds are ranked by reliability which may
256 provide some degree of confidence in these predictions.

257

Name	Prediction Score	AC Applicability Score	Reliability
CPI1062	0.67	0.5	High
CPI1066	0.62	0.38	High
CPI1004	0.62	0.39	High
CPI1012	0.70	0.70	Moderate
CPI1155	0.70	0.40	Moderate
CPI1175	0.65	0.41	Moderate
CPI1153	0.7	0.7	Low

258 **Table 3** – Prospective prediction compounds predicted and prioritized for testing.

259

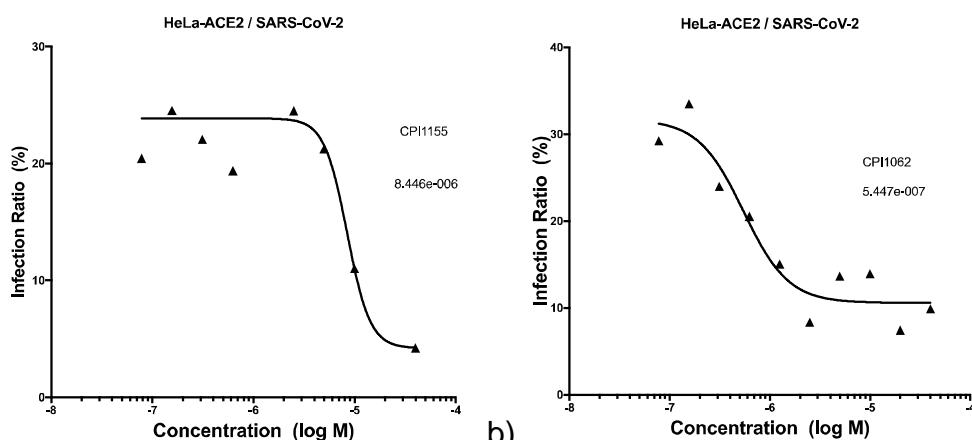
260 ***In vitro* Inhibition Assays of Predicted Compounds**

261 Antiviral activity testing in the HeLa-ACE2 cells demonstrated that CPI1155 and
262 CPI1062 have antiviral activity with IC₅₀ values of 8.4 M and 540nM (Figure 2),
263 respectively. The cell viability of these compounds was also tested, with both CC₅₀
264 higher than 40 μ M. Other compounds did not inhibit viral replication in HeLa cells or had
265 appreciable cytotoxicity.

266

267 **Figure 2** – Preliminary dose response curves for a) CPI1155 and b) CPI1062.

268



269 a)

270

271 **Discussion**

272 One of the challenges for addressing novel viral outbreaks is selection of drugs
273 to test. Testing capacity, even for *in vitro* antiviral activities is likely to be low at the
274 onset of an outbreak, making compound selection even more critical in this situation. In
275 the case of SARS-CoV-2, the initial focus was on molecules that had previously shown
276 activity against SARS or MERS^{37,38}. The training set for the current model is therefore
277 not a random sampling of drug property space. When compared with the PwCL, a
278 library of mostly FDA approved drugs, all molecules superimpose in the property space
279 highlighting the model suitability for drug repurposing. Even with a relatively small
280 training dataset the machine learning models evaluated have shown acceptable five-
281 fold cross validation statistics, with almost all metrics greater than random and ROC
282 >0.75 for AC (Table 1). When compared with different machine learning methods AC
283 outperforms all of them in the SARS-CoV-2 training set, but this may be due to the
284 threshold for all models being set as optimal for AC. However, choosing different values
285 could imbalance the training set and remove important compounds from the active
286 group.

287 More important than a good performance in the training set is the performance
288 on external data, since most prospective predictions will occur for molecules outside
289 training data. For external validation all models had intermediate performance, with
290 ROC of 0.6. Taking into account the small number of molecules and that some test set
291 molecules lie outside the applicability domain, the performance is acceptable. Different
292 from the training set performance, svc had the highest overall score, predicting 60% of
293 the active molecules despite its modest statistics in five-fold cross validation. The good
294 performance of svc in predicting biological activity is in accordance with several studies
295 that show good performance in different datasets ^{28,32,35,39}. Therefore, the models
296 described here are suitable for initial prospective predictions.

297 The applicability and reliability assessment shows that 73% of the test set
298 molecules lie within the model applicability domain with high to moderate reliability, so
299 poor performance in external validation occurs because there isn't a clear boundary in
300 the model's feature space that can correctly classify external data. Increasing the
301 number of molecules might include new features in both actives and inactive molecules
302 which can increase model performance in both training and external data.

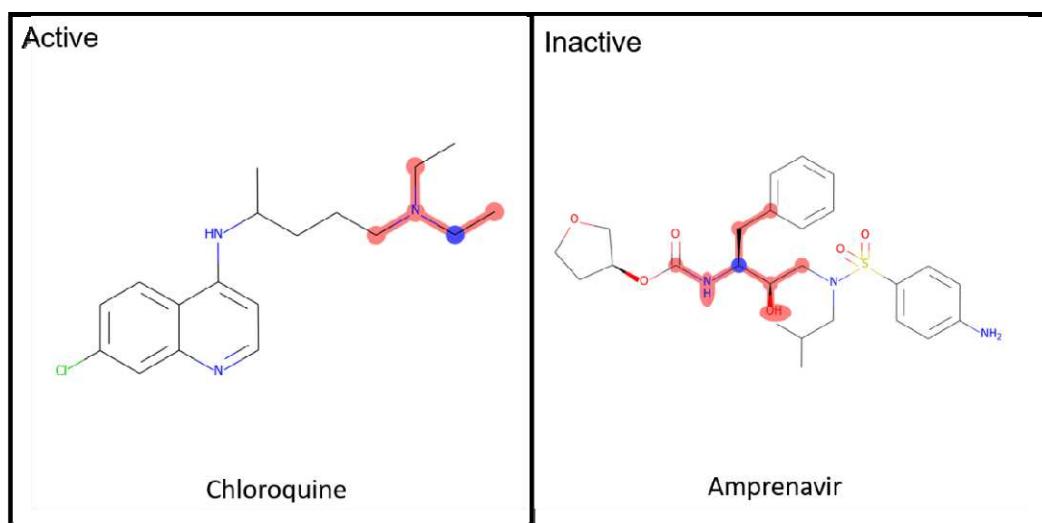
303 The training and test set described herein can be merged to increase data set
304 size and applicability domain. The AC model with merged training and test data has
305 slightly worse statistics (ACC: 0.76, AUC:0.79, CK: 0.53, MCC: 0.75, Pr: 0.76, Recall:
306 0.76, Sp: 0.77, F1: 0.76), but a higher applicability domain. The PCA confirms this wide
307 chemical property space (Figure S1), the PCA of this updated model is much more
308 balanced and broader than the previous one (Figure S2) versus Figure 1B. Without
309 some form of external validation, we cannot assess how predictions of compounds

310 outside the applicability domain perform, as model statistics were comparable it is
311 expected that compounds outside this would obviously have unreliable predictions,
312 however this may be offset by a higher domain which can increase reliability of some
313 compounds.

314 The molecules of the dataset do not have a common scaffold, but there are
315 several common structural features that occur in active/inactive molecules that can be
316 highlighted, such as tertiary amines and aliphatic chains in active molecules and phenyl
317 rings and peptide molecule features in inactive molecules (Figure S3). These most
318 common active features appear in chloroquine, tripanarol and tilorone, while the inactive
319 features appear in darunavir, amprenavir and ritonavir (Figure 3). The lack of common
320 scaffolds and features that appears in more than 30% of the active or inactive
321 molecules shows how different and diverse the active molecules are, which turn
322 classification models for these molecules into a relatively difficult task.

323

324 **Figure 3-** Common Active/Inactive structure features of the SARS-CoV-2 dataset



325

326

327 The performance of a predictive model is highly dependent on the curation and
328 data used. One of the main problems that comes from building models with biological
329 data from different laboratories is data reproducibility and assay standardization⁴⁰. Cell
330 based assays of viral infections have many parameters that can affect the compound
331 potency, e.g., cell lines, MOI, assay readout⁴¹. From all inhibition assays for SARS-CoV-
332 2 collated to date, most studies use MOI of 0.01-0.05 (73% of data), different Vero cell
333 lines (77% of data) and qRT-PCR (60% of data), however there is no clear definition of
334 compound addition time post infection (Figure S1).

335 Besides this, even assays with the same or similar conditions have differences in
336 'control' compounds such as chloroquine or remdesivir, showing a lack of data
337 reproducibility between laboratories, which can impact model building. If we keep only
338 studies with the most in common there is not enough data to build a model, while
339 merging all studies will have problems of different assay parameters. It was shown that
340 for Ebola infections in VeroE6 cells the change in the compound potency at different
341 time post infections are lower when using MOI of 0.01-0.1 therefore, merging different
342 assays with the same cell line and low MOI is a good choice to avoid data
343 inconsistency⁴¹.

344 It should be noted that most of the *in vitro* data collated to date uses Vero or Vero
345 E6 cells for inhibition assays. Although these cell lines have high ACE2 expression
346 levels, they lack a TMPRSS2 gene. Priming of viral S proteins can occur with the host
347 cell protease TMPRSS2 and Cathepsin L and is essential for SARS-CoV-2 entry^{42,43}.
348 Therefore, inhibition assays with cells that do not express TMPRSS2 should be avoided
349 as they might miss compounds that could inhibit the protein and instead find

350 compounds that prevent virus entry by inhibiting only Cathepsin L. In order to avoid
351 these problems with the TMPRSS2 and Cathepsin L gene, cell lines like Calu-3 or
352 modified Vero cell lines should be used instead.⁴⁴

353 From the 7 compounds prioritized for testing in our laboratory using the machine
354 learning model, CPI1155 and CPI1062 showed antiviral activity against SARS-CoV-2
355 infections in HeLa-ACE2 cells. Like Vero cells, HeLa does not express TMPRSS2,
356 therefore compounds might need to be retested in different cell lines to see whether
357 or not the expression of TMPRSS2 affects compound activity.⁴⁵

358 As new data is continually being published the machine learning models can be
359 updated to increase performance in terms of both training and external test set
360 validation. The very latest model for SARS-CoV-2 is available at www.assaycentral.org.
361 In the meantime, we have shown these models perform well with internal cross
362 validation, external validation as well as prospective prediction, enabling us to find
363 additional active molecules. These models should be used to prioritize compounds
364 which have both a high prediction score and reliability as described herein. This will be
365 expected to return more reliable predictions that together with drug discovery expertise
366 can help prioritize compounds in future for *in vitro* testing.

367

368 **Acknowledgements**

369 We would like to kindly acknowledge Dr. Nancy Baker and Ms. Natasha Baker for their
370 help in collating recently SARS-CoV-2 published data. We also thank Biovia for
371 supplying Discovery Studio. Per Subcontract: "This material is based upon work
372 supported by the Defense Advanced Research Projects Agency (DARPA) under

373 Contract No. HR001119C0108." Per DISTAR Form: "The views, opinions, and/or
374 findings expressed are those of the author(s) and should not be interpreted as
375 representing the official views or policies of the Department of Defense or the U.S.
376 Government."

377

378 **Funding**

379 We kindly acknowledge NIH funding: R44GM122196-02A1 from NIGMS (PI – Sean
380 Ekins) and support from DARPA (HR0011-19-C-0108; PI: P. Madrid) is gratefully
381 acknowledged. Distribution Statement "A" (Approved for Public Release, Distribution
382 Unlimited). The views, opinions, and/or findings expressed are those of the author and
383 should not be interpreted as representing the official views or policies of the Department
384 of Defense or the U.S. Government. FAPESP funding: 2019/25407-2 (PI – Glaucius
385 Oliva).

386

387 **Conflicts of interest**

388 SE is CEO and owner of Collaborations Pharmaceuticals, Inc. DHF, KMZ, TRL, AP are
389 employees of Collaborations Pharmaceuticals, Inc.

390

391 **References**

- 392 1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory
393 disease in China. *Nature*. 2020;579(7798):265-269. doi:10.1038/s41586-020-
394 2008-3
- 395 2. Gorbunova AE, Baker SC, Baric RS, et al. The species Severe acute respiratory
396 syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-
397 2. *Nat Microbiol*. 2020;5(March). doi:10.1038/s41564-020-0695-z

398 3. WHO. Naming the coronavirus disease (COVID-2019) and the virus that causes
399 it.

400 4. Practice BB. Coronavirus disease 2019. *World Heal Organ.*
401 2020;2019(April):2633. doi:10.1001/jama.2020.2633

402 5. Kupferschmidt K, Cohen J. Race to find COVID-19 treatments accelerates.
403 *Science (80-).* 2020;(March). doi:10.1126/science.367.6485.1412

404 6. Harrison C. Coronavirus puts drug repurposing on the fast track. *Nat Biotechnol.*
405 February 2020. doi:10.1038/d41587-020-00003-1

406 7. Baker NC, Ekins S, Williams AJ, Tropsha A. A bibliometric review of drug
407 repurposing. *Drug Discov Today.* 2018;23(3):661-672.
408 doi:10.1016/j.drudis.2018.01.018

409 8. Jeon S, Ko M, Lee J, et al. Identification of antiviral drug candidates against
410 SARS-CoV-2 from FDA-approved drugs. *bioRxiv.* 2020:2020.03.20.999730.
411 doi:10.1101/2020.03.20.999730

412 9. Weston S, Haupt R, Logue J, Matthews K, Frieman MB. FDA approved drugs with
413 broad anti-coronaviral activity inhibit SARS-CoV-2 in vitro. 2020;(3).

414 10. Sheahan TP, Sims AC, Zhou S, et al. An orally bioavailable broad-spectrum
415 antiviral inhibits SARS-CoV-2 and multiple endemic, epidemic and bat
416 coronavirus. *bioRxiv.* 2020;(153):2020.03.19.997890.
417 doi:10.1101/2020.03.19.997890

418 11. Caly L, Druce JD, Catton MG, Jans DA, Wagstaff KM. The FDA-approved Drug
419 Ivermectin inhibits the replication of SARS-CoV-2 in vitro. *Antiviral Res.*
420 2020:104787. doi:10.1016/j.antiviral.2020.104787

421 12. Jin Z, Du X, Xu Y, et al. Structure of Mpro from COVID-19 virus and discovery of
422 its inhibitors. *bioRxiv.* 2020:2020.02.26.964882. doi:10.1101/2020.02.26.964882

423 13. Zhang J, Ma X, Yu F, et al. Teicoplanin potently blocks the cell entry of 2019-
424 nCoV. *bioRxiv.* 2020.

425 14. Wang M, Cao R, Zhang L, et al. Remdesivir and chloroquine effectively inhibit the
426 recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.*
427 2020;30(3):269-271. doi:10.1038/s41422-020-0282-0

428 15. Liu J, Cao R, Xu M, et al. Hydroxychloroquine, a less toxic derivative of

429 chloroquine, is effective in inhibiting SARS-CoV-2 infection in vitro. *Cell Discov.*
430 2020;6(1):16. doi:10.1038/s41421-020-0156-0

431 16. Soares VC, Gomes S, Temerozo JR, et al. Atazanavir inhibits SARS-CoV-2
432 replication and pro-inflammatory cytokine production. 2020.

433 17. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: Where have you
434 been? Where are you going to? *J Med Chem.* 2014;57(12):4977-5010.
435 doi:10.1021/jm4004285

436 18. Lima AN, Philot EA, Trossini GHG, Scott LPB, Maltarollo VG, Honorio KM. Use of
437 machine learning approaches for novel drug discovery. *Expert Opin Drug Discov.*
438 2016;11(3):225-239. doi:10.1517/17460441.2016.1146250

439 19. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug
440 discovery and development. *Nat Mater.* 2019;18(5):435-441. doi:10.1038/s41563-
441 019-0338-z

442 20. A P. Vanquishing the Virus: 160+ COVID-19 Drug and Vaccine Candidates in
443 Development. <https://www.genengnews.com/a-lists/vanquishing-the-virus-160-covid-19-drug-and-vaccine-candidates-in-development/>. Published 2020.
444 Accessed May 3, 2020.

446 21. Riva L, Yuan S, Yin X, et al. A Large-scale Drug Repositioning Survey for SARS-
447 CoV-2 Antivirals. 2020.

448 22. Su H, Yao S, Zhao W, Li M, Liu J, Shang W. Discovery of baicalin and baicalein
449 as novel , natural product inhibitors of SARS-CoV-2 3CL protease in vitro. 2020:1-
450 29.

451 23. Sheahan TP, Sims AC, Zhou S, et al. An orally bioavailable broad-spectrum
452 antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple
453 coronaviruses in mice. *Sci Transl Med.* 2020;5883(April).
454 doi:10.1126/scitranslmed.abb5883

455 24. Touret F, Gilles M, Barral K, Nougairède A, Decroly E. In vitro screening of a FDA
456 approved chemical library reveals potential inhibitors of SARS-CoV-2 replication.
457 *bioRxiv.* 2020.

458 25. Xu T. Indomethacin has a potent antiviral activity against SARS CoV-2 in vitro and
459 canine coronavirus in vivo Abstract□: 2020;(December 2019).

460 26. Ekins S, Gerlach J, Zorn KM, Antonio BM, Lin Z, Gerlach A. Repurposing
461 Approved Drugs as Inhibitors of K(v)7.1 and Na(v)1.8 to Treat Pitt Hopkins
462 Syndrome. *Pharm Res.* 2019;36(9):137. doi:10.1007/s11095-019-2671-y

463 27. Dalecki AG, Zorn KM, Clark AM, et al. High-throughput screening and Bayesian
464 machine learning for copper-dependent inhibitors of *Staphylococcus aureus*.
465 *Metallomics.* 2019;11(3):696-706. doi:10.1039/c8mt00342d

466 28. Zorn KM, Lane TR, Russo DP, Clark AM, Makarov V, Ekins S. Multiple Machine
467 Learning Comparisons of HIV Cell-based and Reverse Transcriptase Data Sets.
468 *Mol Pharm.* 2019;16(4):1620-1632. doi:10.1021/acs.molpharmaceut.8b01297

469 29. Anantpadma M, Lane T, Zorn KM, et al. Ebola Virus Bayesian Machine Learning
470 Models Enable New in Vitro Leads. *ACS omega.* 2019;4(1):2353-2361.
471 doi:10.1021/acsomega.8b02948

472 30. Wang P-F, Neiner A, Lane TR, Zorn KM, Ekins S, Kharasch ED. Halogen
473 Substitution Influences Ketamine Metabolism by Cytochrome P450 2B6: In Vitro
474 and Computational Approaches. *Mol Pharm.* 2019;16(2):898-906.
475 doi:10.1021/acs.molpharmaceut.8b01214

476 31. Hernandez HW, Soeung M, Zorn KM, et al. High Throughput and Computational
477 Repurposing for Neglected Diseases. *Pharm Res.* 2018;36(2):27.
478 doi:10.1007/s11095-018-2558-3

479 32. Russo DP, Zorn KM, Clark AM, Zhu H, Ekins S. Comparing Multiple Machine
480 Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Mol
481 Pharm.* 2018;15(10):4361-4370. doi:10.1021/acs.molpharmaceut.8b00546

482 33. Lane T, Russo DP, Zorn KM, et al. Comparing and Validating Machine Learning
483 Models for *Mycobacterium tuberculosis* Drug Discovery. *Mol Pharm.*
484 2018;15(10):4346-4360. doi:10.1021/acs.molpharmaceut.8b00083

485 34. Sandoval PJ, Zorn KM, Clark AM, Ekins S, Wright SH. Assessment of Substrate-
486 Dependent Ligand Interactions at the Organic Cation Transporter OCT2 Using
487 Six Model Substrates. *Mol Pharmacol.* 2018;94(3):1057-1068.
488 doi:10.1124/mol.117.111443

489 35. Lane T, Russo DP, Zorn KM, et al. Comparing and Validating Machine Learning
490 Models for *Mycobacterium tuberculosis* Drug Discovery. *Mol Pharm.*

491 2018;15(10):4346-4360. doi:10.1021/acs.molpharmaceut.8b00083

492 36. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-
493 learn. *GetMobile Mob Comput Commun.* 2015;19(1):29-33.
494 doi:10.1145/2786984.2786995

495 37. Weston S, Haupt R, Logue J, Matthews K, Frieman M. FDA approved drugs with
496 broad anti-coronaviral activity inhibit SARS-CoV-2 in vitro. *bioRxiv*.
497 2020;(3):2020.03.25.008482. doi:10.1101/2020.03.25.008482

498 38. Coleman CM, Frieman MB. Coronaviruses: Important Emerging Human
499 Pathogens. *J Virol.* 2014;88(10):5209-5212. doi:10.1128/jvi.03488-13

500 39. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of Deep Learning with
501 Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery
502 Data Sets. *Mol Pharm.* 2017;14(12):4462-4475.
503 doi:10.1021/acs.molpharmaceut.7b00578

504 40. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to
505 Chemogenomics Data Curation. *J Chem Inf Model.* 2016;56(7):1243-1252.
506 doi:10.1021/acs.jcim.6b00129

507 41. Postnikova E, Cong Y, DeWald LE, et al. Testing therapeutics in cell-based
508 assays: Factors that influence the apparent potency of drugs. *PLoS One*.
509 2018;13(3):1-18. doi:10.1371/journal.pone.0194880

510 42. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry
511 Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease
512 Inhibitor. *Cell.* 2020:1-10. doi:10.1016/j.cell.2020.02.052

513 43. Ou X, Liu Y, Lei X, et al. Characterization of spike glycoprotein of SARS-CoV-2 on
514 virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun.*
515 2020;11(1):1620. doi:10.1038/s41467-020-15562-9

516 44. Matsuyama S, Nao N, Shirato K, et al. Enhanced isolation of SARS-CoV-2 by
517 TMPRSS2-expressing cells. *Proc Natl Acad Sci.* 2020;202002589.
518 doi:10.1073/pnas.2002589117

519 45. Shirato K, Kanou K, Kawase M, Matsuyama S. Clinical Isolates of Human
520 Coronavirus 229E Bypass the Endosome for Cell Entry. *J Virol.* 2017;91(1).
521 doi:10.1128/jvi.01387-16

522

523

524