Novel Development of Predictive Feature Fingerprints to Identify Chemistry-Based Features for the Effective Drug Design of SARS-CoV-2 Target Antagonists and Inhibitors Using Machine Learning

Kelvin Cooper,* Christopher Baddeley, Bernie French, Katherine Gibson, James Golden, Thiam Lee, Sadrach Pierre, Brent Weiss, and Jason Yang

ABSTRACT: A unique approach to bioactivity and chemical data curation coupled with random forest analyses has led to a series of target-specific and cross-validated predictive feature fingerprints (PFF) that have high predictability across multiple therapeutic targets and disease stages involved in the severe acute respiratory syndrome due to coronavirus 2 (SARS-CoV-2)-induced COVID-19 pandemic, which include plasma kallikrein, human immunodeficiency virus (HIV)-protease, nonstructural protein (NSP)5, NSP12, Janus kinase (JAK) family, and AT-1. The approach was highly accurate in determining the matched target for the different compound sets and suggests that the models could be used for virtual screening of target-specific compound libraries. The curation-modeling process was successfully applied to a SARS-CoV-2 phenotypic screen and could be used for predictive bioactivity estimation and prioritization for clinical trial selection; virtual screening of drug libraries for the repurposing of drug molecules; and analysis and direction of proprietary data sets.

INTRODUCTION

The zoonotic transmission and emergence of the new coronavirus, severe acute respiratory syndrome due to coronavirus 2 (SARS-CoV-2), in the Hubei Province of China in late November of 2019 has triggered a global pandemic that has infected over 84 million people and killed over 1.8 million people. The sequence of the SARS-COV-2 virus was published in January 2020, and since then much of the genome function has been elaborated and published. The worldwide scientific community has responded to the COVID-19 pandemic with an intense focus on discovering and testing of therapeutic modalities and that focus is reflected in the rapid vaccine development, a large volume of publications, and multiple clinical studies. Much of the global effort to address COVID-19 by many organizations and research and academic institutes has been accompanied by free and open access to publications and data from the scientific efforts.

In concert with this theme of open access, CAS released a database of 50,000 compounds based on 100 known antivirals. The CAS COVID-19 Antiviral Candidate Compounds Dataset contained CAS REGISTRY numbers, chemical structures, and chemical properties associated with each compound. As a follow up to that data release, we have undertaken a program to conduct analyses of the 50,000 compounds with the goal of providing a curated resource to assist drug discovery and development researchers in the hunt for COVID-19 therapeutics. Specifically, the team aimed to provide access to curated CAS chemical and biological data with potential links to COVID-19 targets and also provide artificial intelligence (AI)-based analysis of the data sets to create predictive connections and associations.

Quantitative structure−activity relationship (QSAR) development has been employed as a critical tool in medicinal chemistry for almost 70 years with the pioneering work of Corwin Hansch as the first exploration of the association of biological activity with physicochemical properties of molecules. The field of QSAR has evolved and expanded considerably since then and has been reviewed in multiple articles and books with the 2013 review by Cherkasov et al. being a useful look at the history and future of QSAR. Since that review, the application of machine learning techniques has allowed the rapid analyses of large data sets to build predictive models and a wide range of approaches to chemoinformatics classifications have been used.

Indeed, artificial intelligence (AI) has been applied to the field of drug discovery for drug repurposing, for multiobjective lead
optimization, in the design-make-test-analyze cycle for reducing cycle times, and in predicting protein−protein interactions using machine learning techniques such as classification or supervised learning approaches. In addition, several groups have employed AI-based approaches to the COVID-19 pandemic, in particular, with the aim of repurposing existing drugs to treat the disease, including most recently a machine learning QSAR study of 3-chymotrypsin-like protease (3CLpro) and RdRp inhibitors.

In this study, the initial target selection was driven by the desire to map small molecule to target interactions across multiple stages of SARS-CoV-2 infection. Thus, we focused on the selection of targets that were involved in (1) viral entry; (2) viral replication, budding, and release; (3) host cell immune responses related to the late-stage cytokine storm; and (4) organ damage. Additional consideration was given to the quality of the biological data available, the size of the data sets for each potential target, the range of bioactivity in those data sets, and whether there were ongoing clinical trials related to the target.

■ RESULTS AND DISCUSSION

Kallikrein. SARS-CoV-2 gains entry to the host cells through the docking of the spike (S) protein on the surface of the virus with the ACE-2 receptor of host cells. The S-protein is then cleaved at the S1−S2 site by the serine protease transmembrane serine protease type 2 (TMPRSS2) and kallikrein 13, allowing the virus to unfold and enter the host cell. Bioactivity values were available for the kallikrein family with plasma kallikrein having the largest and best spread of activity data. Since the serine protease active site of the kallikreins is conserved, a data set of 1050 unique compounds active against plasma kallikrein was compiled from the ChEMBL database and submitted to the model process. The biological activities were converted through the series of manipulations to activity "grades" that could be cross-compared (see Scheme 1).

The activity classification models were trained on physicochemical features from CAS and ChEMBL and engineered features from simplified molecular-input line-entry system (SMILES) strings as well as structural features such as aozoles, sulfonamides, nitriles, trifluoro groups, and carboxyl groups to provide a total of 110 properties for inclusion. A binary classification was the best at distinguishing between active and weakly active compounds with an accuracy of 73% at predicting inactive compounds vs 100% at predicting active compounds (see Figure 1a). Features that predict for "druglike" properties (hydrogen donor and acceptor counts, Alog P, and quantitative

---

Scheme 1. Bioactivity Data Curation Process

For a detailed description of the curation process, see the Materials and Methods section.

Figure 1. Binary confusion matrix (a) and the predicted property fingerprint (b) for plasma kallikrein inhibitors. The confusion matrix labels are 0 for inactive and 1 for active with true label on the y-axis and predicted label on the x-axis. The feature importance plot shows the top 15 features in increasing order of importance from top to bottom.
estimate of drug likeness (QED) weighted) were key characteristics for the plasma kallikrein model (see Figure 1b), with predicted densities, predicted pKs, and polar surface area (PSA) among the most important structural features.

In a structure-based cluster analysis of the test data set, the 210 compounds were clustered into 10 clusters, with 6 clusters having less than 10 representatives (see Table 1 for the total counts in the 10 clusters and their respective successful prediction counts and percentages). The compounds associated with each cluster were mapped into the confusion matrix to assess the distribution of the cluster over true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) labels. This would potentially indicate the degree of bias of a cluster to yield false results (see Table 2).

The analysis shows that clusters 1 and 3 (center structures 1 and 2, respectively) have a high false-positive rate for inactive members of the clusters (58 and 67%, respectively), but the remaining clusters have a good to excellent predictability for low activity. For example, clusters 2, 4, and 6 (center structures 3, 4, and 5) have high predictability for low activity at 97, 69, and 58%, respectively. In addition, the model has an excellent predictivity for high activity at 100% for clusters 1 and 6 (center structures 1 and 5).

Table 1. Cluster Analysis for Plasma Kallikrein Compounds with Numbers of Unique Compounds in 10 Clusters and the Model’s Count of Successful Classification and Percentages

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------|---|---|---|---|---|---|---|---|---|----|
| Total # | 49 | 67 | 39 | 29 | 37 | 9 | 6 | 5 | 4 | 1  |
| Accurate Prediction # | 23 | 65 | 1 | 20 | 9 | 23 | 5 | 4 | 5 | 1  |
| %        | 47%| 97%| 33% | 69%| 100%| 62%| 83%| 100%| 100%| 100% |

“Shaded coloring provided to assist in interpretation of Tables 1−18. Orange and green shading represent the different clusters in the analysis and correlate with similarly shaded clusters in other tables within this work.

Table 2. Cluster vs Confusion Matrix Analysis for Plasma Kallikrein Compounds with Distributions of Classifications for Each

| Plasma Kallikrein Cluster to Confusion Plot |
|-------------------------------------------|
| N = 148 (0.73)                          |
| 19 61 1 20 9 26 2 9 0                  |
| 6 7 8 9 10                              |
| 6 7 8 9 10                              |
| 6 7 8 9 10                              |
| 6 7 8 9 10                              |
| N = 54 (0.27)                           |
| 19 61 1 20 9 26 2 9 0                  |
| 6 7 8 9 10                              |
| 6 7 8 9 10                              |
| 6 7 8 9 10                              |
| 6 7 8 9 10                              |

Predicted Label

“Shaded coloring provided to assist in interpretation of Tables 1−18. Blue shading represents N = x (y) fraction size, with darker shades indicating higher fractions provided in parentheses. Orange and green shading represent the different clusters in the analysis and correlate with similarly shaded clusters in other tables within this work.

The biological activities were converted through the series of manipulations to activity "grades" that could be cross-compared (see Scheme 1). The activity classification models were trained on physiochemical features from CAS and ChEMBL and engineered features from SMILES strings as well as structural features such as azoles, sulfonamides, nitriles, trifluoro groups, and carboxyl groups to provide a total of 110 properties for inclusion.

The ternary model was the best at distinguishing inactive, moderately active, and highly active compounds in the HIV protease family. The model provides a 73% success rate at predicting poorly active compounds; a 57% probability of predicting moderate activity; and an 81% success rate at predicting the highly active compounds (see Figure 2a). Features that predict for "druglike" properties (Alog P, CX log D, QED weighted, hydrogen donor−acceptor counts, and freely rotatable bonds) were among the most important characteristics for the HIV protease model (see Figure 2b), with predicted densities, molar volumes, and PSA among the most important structural features.

In a structure-based cluster analysis of the hold-out data set, the 769 compounds can be clustered into 22 distinct clusters with 4 clusters containing over 35 unique compounds (see Table 3).

Human Immunodeficiency Virus (HIV) Protease. Once the SARS-CoV-2 virus has gained entry to the cell, the viral RNA is translated to a set of large polyproteins that are then processed into the proteins required for viral replication. Nonstructural protein (NSP)3 also known as papain-like proteinase (PLpro) and NSP5, also known as 3-chymotrypsin-like proteinase (3CLpro), are proteinases responsible for the processing of polyprotein ORF-1. Inhibition of one or both of these enzymes should result in reduced viral translation and consequently a lower viral load and potentially milder disease. Although the HIV-protease inhibitors were designed to inhibit the HIV aspartyl protease, there were indications that some members of the class are thought to interact with NSP3 and/or NSP5.25,26 Accordingly, we assembled a data set from the ChEMBL database with 3844 unique compounds and bioactivity data points.
Figure 2. Ternary confusion matrix (a) and predicted feature fingerprints (PFF) (b) for HIV protease inhibitors. The confusion matrix labels are 1 for weakly active, 2 for moderately active, and 3 for highly active with true label on the y-axis and predicted label on the x-axis. The feature importance plot shows the top 15 features in increasing order of importance from top to bottom.

Table 3. Cluster Analysis for HIV Protease Compounds with Numbers of Unique Compounds in 10 Major Clusters and the Model’s Count of Successful Classification and Percentages

| Cluster | 1 | 2 | 3 | 4 | 6 | 7 | 10 | 13 | 16 | 20 |
|---------|---|---|---|---|---|---|-----|-----|-----|-----|
| Total # | 471 | 135 | 5 | 3 | 75 | 11 | 37 | 8 | 3 | 3 |
| Accurate Prediction # | 312 | 98 | 2 | 1 | 54 | 11 | 31 | 6 | 3 | 0 |
| % | 66% | 73% | 40% | 33% | 72% | 100% | 84% | 75% | 100% | 0% |

The compounds associated with each cluster were mapped into the confusion matrix to assess the distribution of the cluster over true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) labels. This would potentially indicate the degree of bias of a cluster to yield false results (see Table 4).

The largest cluster, cluster 1 with 471 unique compounds, is the prototypical protease inhibitor with center structure 6 and, as with the other highly active clusters, clusters 2, 6, and 10 (center structures 7, 8, and 9); the model was highly predictive across the activities with moderate activity being the most difficult to accurately predict.

Table 4. Cluster vs Confusion Matrix Analysis for HIV Protease Compounds with Distributions of Classifications
As discussed previously, NSP5, also known as 3-chymotrypsin-like proteinase (3CLpro), is one of the proteinases responsible, in part, for the processing of polyprotein ORF-1. Inhibition of this enzyme should result in reduced translation and consequently a lower viral load and potentially milder disease. Accordingly, we assembled a data set of 303 unique compounds and bioactivity data points from the CAS REGISTRY. The biological activities were converted through the series of manipulations to activity “grades” that could be cross-compared (see Scheme 1). The activity classification models were trained on physiochemical features from CAS and ChEMBL and engineered features from SMILES strings as well as structural features such as azoles, sulfonamides, nitriles, trifluoro groups, and carboxyl groups to provide a total of 110 properties for inclusion.

A binary classification was the best at distinguishing between active and weakly active compounds with an accuracy of 83% at predicting inactive compounds vs 100% at predicting active compounds (see Figure 3a). Features that predict for “druglike” properties (Alog P and QED weighted) were among the most important characteristics for the NSP5 model (see Figure 3b), with predicted molar and mass solubilities at pH 1, 7, and 10 among the most important structural features.

In a structure-based cluster analysis of the hold-out data set of 61 compounds, nine distinct clusters were formed with two clusters containing the most compounds (see Table 5, which shows the total counts for the nine major clusters and their respective successful prediction counts and percentages).

The compounds associated with each cluster were mapped into the confusion matrix to assess the distribution of the cluster over true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) labels. This would potentially indicate the degree of bias of a cluster to yield false results (see Table 6).

In general, the model makes highly accurate predictions across all structural classes indicating that there is no structural bias. The most populated clusters, 3 and 6 (center structures 10 and 11), have a high predictivity for their low activity, and cluster 5 (center structure 12) has excellent predictivity for low and high activity.
**NSP12—RNA-Dependent RNA Polymerase.** The SARS-CoV-2 virus genome encodes an RNA-dependent RNA polymerase (RdRp), named NSP12, that is essential for viral replication. The NSP12 protein requires a docking interaction with the viral proteins NSP7 and NSP8 for enzymatic function, and X-ray data for remdesivir bound into the active site has been published.\textsuperscript{27} Although bioactivity data for the SARS-CoV-2 RdRp has not been published, the sequence, functionality, and active site configuration of several other RdRps overlap with the SARS-CoV-2 sequence/functionality.\textsuperscript{28} We assembled a data set of 2888 unique compounds with bioactivity data points from the CAS database. The biological activities were converted through the series of manipulations to activity ”grades” that could be cross-compared (see Scheme 1). The activity classification models were trained on physiochemical features from CAS and ChEMBL and engineered features from SMILES strings as well as structural features such as azoles, sulfonamides, nitriles, trifluoro groups, and carbonyl groups to provide a total of 110 properties for inclusion.

A binary classification was the best at distinguishing between active and weakly active compounds with an accuracy of 85% at predicting inactive compounds vs 83% at predicting active compounds (see Figure 4a). Features that predict for “druglike” properties (AlogP and CX log P) were among the most important characteristics for the NSP12 model (see Figure 4b), with predicted densities molar volumes, and PSA among the most important structural features. Interestingly, the predicted molar and mass solubilities at pH 7 were also in the top 15 predictive features of the model.

In a structure-based cluster analysis of the hold-out data set of 575 compounds, 13 distinct clusters were formed with 4 clusters containing 20 or more unique compounds (see Table 7, which shows the total counts for the top 10 major clusters and their respective successful prediction counts and percentages). In general, the model makes very accurate predictions.

The compounds associated with each cluster were mapped into the confusion matrix to assess the distribution of the cluster over true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) labels. This would potentially indicate the degree of bias of a cluster to yield false results (see Table 8). Clusters 1, 2, 5, and 10 (center structures 13–16) have excellent predictivity with at least a 73% rating. Disappointingly the data set contained a disproportionate number of weakly actives, and the structural class that is closest to the antiviral remdesivir was weakly active and, not surprisingly, the model predicted weak bioactivity for remdesivir. This probably indicates a lack of suitable bioactivity data for the remdesivir-like class of molecules that could strengthen the predictability of the NSP12 model.

**Angiotensin II Type-1 Receptor (AT-1).** SARS-CoV-2 gains entry to its target cell by binding to the angiotensin-converting enzyme type-2 receptor. Angiotensin levels are elevated in animal models, where the SARS-CoV S-protein is injected and the high levels could contribute to the severe lung injury in patients.\textsuperscript{29} A set of 1186 AT-1 receptor antagonists with a wide range of activities were selected for modeling from the ChEMBL database and included six known FDA-approved drugs. The biological activities were converted through a series of manipulations to activity ”grades” that could be cross-compared.
The activity classification models were trained on physiochemical features from CAS and ChEMBL and engineered features from SMILES strings as well as structural features such as azoles, sulfonamides, nitriles, trifluoro groups, and carboxyl groups. In the final model, 110 molecular features were chosen for each compound.

A binary model is highly predictive for highly active and poorly active compounds with the combination of CAS and ChEMBL properties giving the best predictive power: poorly active compounds were predicted with 84% success and highly active compounds were predicted with an 81% success rate (see Figure 5a). Ternary evaluation accurately predicts for a decision tree that distinguishes poorly active (88%), moderately active (60%), and highly active compounds (75%) (see Figure 5c).

The top 15 properties from the 110 that were initially selected as the most important for the model are a mixture of physicochemical and structural features (see Figure 5b,d). Features that are linked to acidic pH predictions (predicted pKas most acidic, log \(D\) at pH1, \(K_{oc}\) at pH1) were important, as were features linked to being “druglike” (CX log \(P\), CX log \(D\), and QED weighted).

In a structure-based cluster analysis of the hold-out data set of 239 compounds, 10 discrete clusters were formed with 4 of the clusters having just a single representative (Table 9 shows the total counts for the top 10 clusters and their respective successful prediction counts and percentages).

The compounds associated with each cluster were mapped into the confusion matrix to assess the distribution of the cluster

---

**Table 8. Cluster vs Confusion Matrix Analysis for NSP12 Compounds with Distributions of Classifications for Each Cluster with True Label on the y-axis and Predicted Label on the x-axis**

---

**Figure 5.** Confusion matrices and predictive feature fingerprint characteristics for AT-1 binary classification, panels (a) and (c), and AT-1 ternary classification, panels (b) and (d). The binary confusion matrices (a) labels are 0 for inactive and 1 for active with true label on the y-axis and predicted label on the x-axis. The ternary confusion matrix (b) labels are 3 for highly active, 2 for moderately active, and 1 for weakly active. The feature importance plots (b) and (d) show the top 15 features in increasing order of importance from top to bottom.
over true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) labels. This would potentially indicate the degree of bias of a cluster to yield false results. The three largest clusters follow the general confusion matrix outcome with clusters 1, 3, and 5 (center structures 17–19 respectively)

Table 9. Cluster Analysis for AT-1 Compounds with Numbers of Unique Compounds in the Top 10 Clusters and the Model’s Count of Successful Classification and Percentages

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------|---|---|---|---|---|---|---|---|---|----|
| Total # | 21| 1 | 89| 19| 90| 1 | 1 | 8 | 7 | 1  |
| Accurate Prediction # | 16| 1 | 72| 9 | 55| 0 | 1 | 8 | 1 | 1  |
| %       | 76%|100%|81%|47%|61%|0%|100%|100%|14%|100%|

Table 10. Cluster vs Confusion Matrix Analysis for AT-1 Compounds with Distributions of Classifications for Each Cluster with Actual Classification on the y-axis and Predicted Classification on the x-axis

Figure 6. Predicted vs actual activity grades for the six FDA-approved AT-1 antagonists.
being studied in the clinic. Inhibitors including pacritinib, baricitinib, and tofacitinib are in process. In seriously ill patients, there is an overwhelming inflammatory response, also known as the cytokine storm, which can lead to multiple organ failure and death. Several JAK inhibitors including pacritinib, baricitinib, and tofacitinib are being studied in the clinic.

A set of six Food and Drug Administration (FDA)-approved drugs (fimasartan, pratosartan, saprisartan, losartan, candesartan, and zolasartan), not included in the model’s training set, were run through the model to determine whether it could predict their bioactivities. The model was quite successful in providing a predicted activity score for each compound with a match of ±1 to the true values, with four of the six compounds having exact predicted values (see Figure 6).

Figure 7. Distribution of bioactivity for the 6638 compounds across the JAK family.

Janus Kinases (JAK) 1–3. The Janus kinases are intracellular, nonreceptor tyrosine kinases that are intimately involved in the signal transduction of cytokine-mediated signals and therefore central to many immunologic and inflammatory processes. In seriously ill patients, there is an overwhelming inflammatory response, also known as the cytokine storm, which can lead to multiple organ failure and death. Several JAK inhibitors including pacritinib, baricitinib, and tofacitinib are being studied in the clinic.

A set of 6638 JAK inhibitors with a wide and well-distributed range of activities was selected from the ChEMBL database (see Figure 7 for a Venn diagram showing the distribution of bioactivities across the three kinases). The biological activities were converted through a series of manipulations to activity “grades” that could be cross-compared (see Scheme 1). For each of the three kinases, the activity classification models were trained on physiochemical features from CAS and ChEMBL and engineered features from SMILES strings including structural features such as azoles, sulfonamides, nitriles, trifluoro groups, and carboxyl groups. In the final model, 110 molecular features were chosen for each compound.

Three models were built, and ternary classifications provided the best predictions for each of the kinases. Thus, for JAK-1, the model accurately predicts for a decision tree that distinguishes poorly active (90%), moderately active (60%), and highly active compounds (74%) (see Figure 8a); for JAK-2, the model accurately predicts for a decision tree that distinguishes poorly active (82%), moderately active (66%), and highly active compounds (81%) (see Figure 8b); and for JAK-3, the model accurately predicts for a decision tree that distinguishes poorly active (77%), moderately active (64%), and highly active compounds (84%) (see Figure 8c).

The predictive feature fingerprints of each kinase are unique, emphasizing druglike properties as the best predictors. Thus, for JAK-1, -2, and -3, 7 of the top 15 properties are the same but have different factor levels (polar surface areas, monoisotopic molecular weight, predicted pKₐs most acidic, molar volumes, predicted pKₐ, most basic, Alog P, and QED weighted). JAK-1 has three unique factors (f count, hydrogen donor–acceptor sums, and CX log D), four shared with JAK-2 (densities, heavy atoms, hydrogen donor–acceptor counts, and molar intrinsic solubilities), and none shared with JAK-3. JAK-2 has two unique properties, predicted log Ds 25 °C pH 1 and n-count, and has four shared with JAK-1 (see above) and one shared with JAK-3 (boiling point). JAK-3 has five unique features (predicted log Ds 25 °C pH 10, and SMILES features “-“,”\[“,”\], and “H”). In a structure-based cluster analysis of the data set, the 2111 hold-out compounds (623 for JAK-1, 1236 for JAK-2, 459 for JAK-3, and 45 for JAK selectivity) can be clustered into 50 discrete clusters, with 10 of the clusters containing 91% of the data set. The compounds associated with each cluster were mapped into the confusion matrix to assess the distribution of the cluster over true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) labels. This would potentially indicate the degree of bias of a cluster to yield false results.

For JAK-1, the 10 top clusters (Table 11 shows the total counts for the clusters and their respective successful prediction counts and percentages) follow the general confusion matrix outcome with all clusters showing better than 67% predictability success, with the exception of clusters 6 and 9 (center structures 20 and 21). Notably, the most active compounds against JAK-1 are in clusters 5 and 15 (center structures 22 and 23), but the model is the best at predicting most active bioactivity for cluster 15 (center structure 23) (see Table 12).
counts and percentages) follow the general confusion matrix outcome with only cluster 26 (center structure 24) showing less than 60% predictivity. Notably, the most active compounds against JAK-2 are in clusters 2, 5, 6, 11, 15, and 19 (center structures 25, 22, 20, 26, 23, and 27, respectively), but the model is the best at predicting most active bioactivity for clusters 5, 15, and 19 (center structures 22, 23, and 27) (see Table 14).
For JAK-3, the 10 top clusters (Table 15 shows the total counts for the clusters and their respective successful prediction counts and percentages) follow the general confusion matrix outcome with clusters 1, 5, 10, 15, 26, and 29 (center structures 28, 22, 29, 23, and 30) showing 74% or higher success at predicting bioactivity for the compounds. Notably, the most active compounds against JAK-3 are in clusters 5, 9, 11, 15, and 29 (center structures 22, 21, 26, 23, and 30) and the model predicts their bioactivity with >85% accuracy (see Table 16), with the exception of cluster 9 (center structure 21), which is at 33%.

A smaller set of the JAK inhibitors, 971 compounds, had published bioactivity data for all three kinases, and this subset of compounds was explored as a validation tool to determine whether selectivity for one or more kinase vs the others could be modeled (see Figure 9a). Compounds were considered to be selective for one kinase over the other two if the compound showed at least 30-fold selectivity over both of the other kinases. Using this curation protocol, 198 inhibitors were chosen, with
76 compounds selective for JAK-1, 43 compounds selective for JAK-2, 54 compounds selective for JAK-3, and 25 compounds with no selectivity for one single kinase over the other two. Seventy compounds with partial selectivity for two of the kinases over the other kinase were excluded from the model. The model was built with 80% of the compounds and a hold-out set of 45 compounds was used to test the model.

The model built on this data set was able to classify compounds based on whether they were selective for JAK-1, 2, or 3 with weak ability to predict for nonselectivity (see Figure 9a). Thus, the model was able to predict JAK-1 selectivity with an 83% success rate, JAK-2 selectivity with an 89% success rate, JAK-3 selectivity with a 100% success rate, and nonselectivity with a 33% success rate. Attempts to further define the partially selective inhibitors were not successful, in part, at least, to the small n-numbers.

The set of features that the model selects to generate the quaternary plot comprise both physicochemical and structural elements. Physicochemical features that are consistent with druglike features, such as Alog P, CX Log D, QED weighted, and CX log P, are extremely important with n-count; chiral count and aromatic rings being key structural features (see Figure 9b).
Given the high homology of the active sites of the JAK family, the degree of predictive success that the model achieves is remarkable.

For the selectivity data set, only eight clusters were formed with cluster 5 (center structure 22) having the highest number of compounds (see Table 17 for the total counts for the clusters and their respective successful prediction counts and percentages).

The cluster analysis follows the general confusion matrix outcome with clusters 5, 6, 11, 12, and 26 (center structures 22, 20, 26, 31, and 24) showing 75, 70, 100, and 100% success at predicting the correct selectivity for the compounds (see Table 18).

**Table 16. Cluster vs Confusion Matrix Analysis for JAK-3 Compounds with Distributions of Classifications for Each Cluster with True Label on the y-axis and Predicted Label on the x-axis**

![Confusion matrix and feature fingerprint](image)

**Figure 9.** Confusion matrix (a) and predictive feature fingerprint characteristics (b) for JAK selectivity. The confusion matrix labels are 1 for JAK-1, 2 for JAK-2, 3 for JAK-3, and 4 for nonselective with true label on the y-axis and predicted label on the x-axis. The feature importance plot shows the top 15 features in increasing order of importance from top to bottom.

**Table 17. Cluster Analysis for JAK Selectivity with Numbers of Unique Compounds in the Eight Clusters and the Model’s Count of Successful Classification and Percentages**

|   | 1  | 2  | 5  | 6  | 11 | 12 | 15 | 26 |
|---|----|----|----|----|----|----|----|----|
| Total # | 5  | 1  | 25 | 1  | 3  | 2  | 2  | 6  |
| % | 40% | 0% | 80% | 0% | 100% | 0% | 100% | 50% |

A set of four FDA-approved drugs (ruxolitinib, fedratinib, midostaurin, and nintedanib) and seven JAK clinical candidates (lestaurtinib, pelitinib, abrocitinib, PF-06651600, KW-2249, ...
CGP-52421, and BMS-911543), which were not included in the model’s training set, were run through the model to determine whether the model could predict their selectivities. The model was quite successful in providing a predicted selectivity score for each compound. Table 18. Cluster vs Confusion Matrix Analysis for JAK-1−3 Selective vs Nonselective Compounds with Distributions of Classifications for Each Cluster

![Table 18](image)

“The confusion matrix labels are 1 for JAK-1, 2 for JAK-2, 3 for JAK-3, and 4 for nonselective with true label on the y-axis and predicted label on the x-axis.

Figure 10. Predicted vs actual selectivity for the four FDA-approved drugs and eight clinical candidates with activity against the JAK family.
each compound with a match ± 1 to the true. The model correctly predicts that lestaurtinib interacts with JAK-1, correctly predicts that pelitinib and midostaurin interact with JAK-2, correctly predicts that ruxolitinib interacts with JAK-3, and correctly predicts KW-2249 and BMS-911543 as non-selective (see Figure 10).

Activity Classification for AT-1, JAK-1−3, and HIV Protease. After having conducted the model building for the AT-1, JAK, and HIV-protease data sets, we wanted to determine whether the three target data sets can be distinguished from each other using a random forest model. Accordingly, a ternary classification model was trained on physiochemical features from the full chemistry feature list of 110 properties. As shown in Figure 11a, the model easily distinguishes between the three data sets, accurately classifying each molecule into the correct target (98% success for AT-1, 99% for HIV protease, and 99% success for JAK family), with the primary features that distinguish the targets being heavily dominated by structural features that include the tetrazole ring, that features heavily in the AT-1 data set (see Figure 11b).

Phenotypic Screening. Several independent groups have performed screening of different sets of FDA-approved drugs in SARS-CoV-2 phenotypic assays, and their data outputs have been compiled in ChEMBL and released. Three such screening exercises that measure SARS-CoV-2-induced cytotoxicity and the ability of FDA-approved drugs to prevent the cell death were considered for inclusion in the modeling effort.34−36 Ultimately, we examined the SARS-CoV-2-induced cytotoxicity of CACO-2 (human colon cell line used extensively for in vitro permeability studies). In this study, approximately 5600 FDA-approved drugs were screened in a 48 h cytotoxicity study to measure the degree of inhibition of cell death. Compounds that displayed a greater than 75% inhibition were considered active, and from the large set of FDA-approved drugs, 267 were considered active.34

We used the bioactivity data from ChEMBL, and a subset of the data was trained on physiochemical features from the set of 110 chemical properties and a hold-out set of 868 compounds were then used to establish a binary classification model that could predict whether a molecule was active or inactive in the

Figure 11. Confusion matrix (a) and predictive feature fingerprint characteristics (b) for classification of AT-1, HIV protease, and JAK family targets. The confusion matrix labels are 1 for AT-1, 2 for HIV protease, and 3 for JAK family with true label on the y-axis and the predicted label on the x-axis. The feature importance plot shows the top 15 features in increasing order of importance from top to bottom.

Figure 12. Binary confusion matrix (a) and predictive feature fingerprint characteristics (b) for classification of activity in a SARS-CoV-2-induced CACO-2 cell cytotoxicity assay. The confusion matrix labels are 0 for inactive and 1 for active with true label on the y-axis and predicted label on the x-axis. The feature importance plot shows the top 15 features in increasing order of importance from top to bottom.
phenotypic assay; the model was able to identify bioactivity with 65% accuracy (35% false-negative rate for inactive compounds) (see Figure 12).

**DISCUSSION/CONCLUSIONS**

We have used a highly methodical approach to addressing the COVID-19 pandemic challenge by dividing the life cycle of the disease into phases and identifying targets that address each stage. It was important in building the data sets on which to base the modeling to extract high-quality bioactivity data, followed by careful curation to ensure comparability of the data both within specific targets and across the targets. In addition, it was critical to include a uniform set of chemical features that are also curated to avoid repetition and redundancy in the model building. In this model building, we used highly curated chemical features available from CAS and ChEMBL and included a unique, curated SMILES feature set designed for this type of analysis. Once the data sets were assembled, we used a combination of random forest classification and structure-based cluster analyses to interpret the results.

We have found that the unique approach to data curation coupled with RF analyses produces target-specific and cross-validated predictive feature fingerprints (PFFs) that have high predictability across multiple targets: plasma kallikrein, HIV protease, NSP5, NSP12, AT-1, and the JAK family. This wide applicability to different target types (protease, RNA polymerase, G protein-coupled receptor (GPCR), tyrosine kinase, and a phenotypic assay) as well as to different chemotypes within a target is a major strength and differentiator of our methodology. Most QSAR methods, including regression analyses and principal component analysis (PCA) analyses, require a highly specific and tailored approach based on the specific target and are primarily based on a single chemotype that is active against the selected target. In addition, our methodology provides the medicinal chemist with a set of key features (the PFF) that can be used as a guide to the next round of compound design.

Our approach was also highly accurate in determining the matched target for the different compound sets, and this ability to distinguish compounds on the basis of a target fingerprint suggests that the models could be used for AI-based sorting of chemical structures with the potential for virtual screening of compound libraries for assessing compounds for clinical trial potential.

The ability of each target model to create a binary decision tree for inactive or active compounds or a ternary decision tree for weakly active, moderately active, and highly active compounds also suggests that the models could be used for virtual screening of target-specific compound libraries to select the most active compounds for synthesis and or clinical testing. The PFFs have wide structural applicability as shown by cluster analyses of each of the data sets, and the PFFs could also be used to guide additional synthesis or virtual library creation for screening through the models.

The ability to distinguish clinical candidates in the AT-1 and JAK data sets also suggests that this methodology could be used for honing virtual libraries toward compounds with druglike characteristics.

In addition, the model process is applicable to phenotypic screening and could be used for large-scale virtual screening of compound libraries given a training set of bioactivity values. This model building process could be used for predictive bioactivity estimation and prioritization for clinical trial selection; virtual screening of drug libraries for repurposing of drug molecules; and analysis and direction of proprietary data sets.

The model building process appears to be highly reproducible and should be applicable to a wider range of targets and therapeutic areas. The successful application of this unique curation-modeling methodology could effectively reduce medicinal chemistry cycle times by 18–24 months and reduce the overall drug discovery timelines by similar amounts. Additional studies to explore the applicability of the model building process are underway and will be targeted toward establishing the complete rules-based set of medicinal chemistry...
Table 19. Summary of the Selected Targets, Synonyms, Organism, UniProt ID (If Available), the Stage of the Disease, the Unique Compound Data Set Size and a Related Clinical Candidate

| target                        | synonyms            | organism | UniProt ID         | stage of disease            | # of unique compounds | clinical candidate |
|-------------------------------|---------------------|----------|--------------------|-----------------------------|-----------------------|--------------------|
| plasma kallikrein             | human               |          | KLKB1_HUMAN       | viral entry                 | 1050                  | camostat[^1^]      |
| NSP3                          | PLpro, HIV Protease | viral    | unassigned         | viral replication/Host innate immune response | 3844                  | lopinavir[^5^]     |
| NSP5                          | 3CLpro, Mpro        | viral    | unassigned         | viral replication           | 303                   | danoprevir[^6^]    |
| NSP12                         | RdRp                | viral    | R1AB_SARS2         | viral replication           | 2888                  | remdesivir[^7^]     |
| AT-1                          | human               |          | AGTR1_HUMAN       | lung injury                 | 1186                  | losartan[^8^]       |
| JAK-1–3                       | human               |          | JAK1_HUMAN        | host immune response        | 6639                  | Xeljanz[^9^]        |
|                               |                     |          | JAK2_HUMAN        |                             |                       |                    |
|                               |                     |          | JAK3_HUMAN        |                             |                       |                    |
| phenotypic                    | human and viral     | not applicable |                 |                             | 4337                  |                    |

Descriptors for each target to aid in compound design, plus a network tool to link targets, structures, and diseases that will be used to construct a medicinal chemistry knowledge graph.

**MATERIALS AND METHODS**

**Workflow.** The process workflow for the development of the models is illustrated in Scheme 2. The first step of target selection was followed by identifying molecules with bioactivity data from a variety of sources. The bioactivity data, molecule descriptors, and predicted physicochemical properties were then curated and compiled in a consolidated data set for each target. The data was partitioned into a training set for model development and a testing set for model evaluation. Next was model development, which involved designing and tuning a model to optimize its cross-validated performance on the training set. The tuned model was then further tested using a one-off hold-out test on the testing data set. A more detailed description for each step of the process follows.

**Derivation of the CAS COVID-19 Antiviral Candidate Compounds Data Set.** CAS scientists selected a seed list of 100 known antiviral agents with known clinical activity from the CAS REGISTRY that was expanded by a combination of a similarity search based on the Tanimoto algorithm[^17^,^30^] and based on a substructure search of the seed list compound base structure. The resulting set of 65,000 compounds was trimmed to eliminate overlapping structures from the two approaches, and the resulting data set of 50,000 unique compounds has been released for open access.[^10^]

**Target Selection.** Target selection was driven by the desire to map small molecule to target interactions across multiple stages of SARS-CoV-2 infection. SARS-CoV-2 host cell entry is mediated by the viral spike protein (S-protein), which docks to the angiotensin-converting enzyme type-II receptor (ACE-2). The spike protein—ACE-2 complex is then processed by transmembrane serine protease type 2 (TMPRSS2), which enables membrane fusion and the release of the viral contents into the cell.[^39^-^41^] The serine protease Kallikrein 13 has also been linked to SARS-CoV-2 cell entry in cleaving the S1–S2 region of the viral spike protein.[^41^]

The virus RNA encodes at least 14 open reading frames (ORFs) that are released into the cell for translation and processing. The ORF1a/ORF1b complex, which encodes a large polyprotein, is self-processed into a series of 16 proteins that control the viral replication (nonstructural protein [NSP] 1–12, 14, and 16) and downregulation of the host interferon response (NSP13, NSP15, and ORF9b).[^21^]

Viral replication, budding, and release are controlled through a number of viral proteins that include NSP3, a papain-like protease (PLpro); NSP5, a chymotrypsin-like protease (3CLpro);[^22^] and NSP12, the RNA-dependent RNA polymerase (RdRp).[^43^-^46^] Although HIV-protease inhibitors are designed as inhibitors of the HIV aspartyl protease, there are indications that some members of the class may interact with NSP3 and/or NSP5.[^25^,^26^]

The release of multiple viral particles prompts the host to initiate an immune response[^47^], which in many patients becomes an overwhelming inflammatory response that can lead to cytokine storm, precipitating multiple organ failure and death.[^30^-^32^] Among the many inflammatory mediators involved in this process, the Janus kinases are heavily implicated as they are obligatory partners in cytokine signaling.[^44^,^45^]

Additional consideration was given to the quality of the biological data available, size of the data sets for each potential target, range of bioactivity in those data sets, and whether there were ongoing clinical trials related to the target. Ideally, data sets of at least 1000 unique compounds with bioactivity spread evenly from mM to nM are required for the model generation to comply with an appropriate data point to feature ratio. Accordingly, we selected the following targets: plasma kallikrein; NSP3/HIV protease; NSP5 (3CLpro); NSP12 (RdRp); angiotensin II receptor type 1 (AT-1);[^50^] and JAK-1, -2, and -3. In addition, we sought to develop a model based on phenotypic screen results as a holistic approach to the problem. A summary of the selected targets, the stage of the disease targeted, the size of the unique compounds with bioactivity data sets, and clinical candidates is given in Table 19.

**Biodata Sources and Curation.** Biological data was collected from three main sources: GoStats[^51^], Liceceptor[^52^], and ChEMBL[^5^], using the target descriptor as the primary filter for compound selection. The bioactivities from the three databases were a complex mix of activity type, activity unit, and activity value and were subjected to a curation process prior to incorporation into the modeling analysis. In consideration of the different evaluation standards that each primary source of bioactivity data uses, we used single sources of bioactivity data for each of the targets, and this is noted in each target described in the Results and Discussion section.

Each data set was narrowed to bioactivities represented by EC[^50^], IC[^50^], K[^50^], K[^50^], p[^50^], and p[^50^], and, to use as much of the data as possible, the individual values were converted through the series of manipulations to values that could be cross-compared (see Scheme 1). We used the statistically validated methodology...
published by the ChEMBL group to make the conversions of the different activity types to a single cross-comparable value.\textsuperscript{60} Thus, EC\textsubscript{50} and IC\textsubscript{50} values were converted to their − log 10 values; \( K_a \) and \( K_b \) values were divided by 2.3 and then converted to their − log 10 values; and \( pK_a \) values were adjusted by subtracting 0.32. All values were then converted to activity grades starting with a value of \( p\text{IC}_{50} \) \( p\text{EC}_{50} \) \( pK_a \) \( pK_b < 5.5 \) assigned a grade of 1, stepping up by half log units and finishing with activity values of >9.5 being assigned a grade of 10 (see Scheme 1).

### Chemical Property Sources and Curation

The primary sources of chemical property data were taken from the CAS REGISTRY and from the ChEMBL data set and comprised both physiochemical and structural properties.

#### Physicochemical Properties

The CAS REGISTRY properties correspond to the physiochemical characteristics of each small molecule. They include: polar surface area (PSA); hydrogen bond acceptor (HBA) count; hydrogen bond donor (HBD) count; bio-concentration factors (predicted for pH 1–10); mass solubilities (predicted for pH 1–10); predicted molar solubilities (pH 1, 7, and 10); \( pK_a \) (under acidic and basic conditions); predicted log \( D \) (pH 1, 7 and 10); predicted boiling point; predicted enthalpies of vaporization, predicted densities; predicted molar volumes; predicted mass intrinsic solubilities; and predicted molar intrinsic solubilities (pH 1, 7, and 10). From the ChEMBL set, we selected the following physical properties: molecular formula; estimated lipophilicity (\( \log D \)); Lipinski’s rule of five violations (ROS violations); QED weighted; \( CX \log D \); \( CX \log D \); aromatic rings; heavy atoms, HBA Lipinski; HBD Lipinski; and monoisotopic molecular weight.

#### Structural Properties

The structural properties included in the data are the following: hydrogen bond acceptor counts; hydrogen bond donor counts; hydrogen donor–acceptor sums; freely rotatable bond counts; and polar surface areas. The SMILES strings of each molecule were used to engineer features, generating the set of unique characters across all of the SMILES strings across data for each target. A distributed representation of each symbol in the SMILES was calculated. Each feature vector consists of 74 features, 17 of which are atomic symbols and the remaining 57 are SMILES symbols.

#### AI Methods. Classification Model: Random Forests

In this study, we used the random forest (RF) methodology, which is a tree-based learning method initially introduced by Breiman in 2001.\textsuperscript{61} The RF algorithm generates multiple decision trees from randomly selected feature sets. This results in unpruned classification trees that reduce overfitting to training data. Activity classification predictions are made by each decision tree and a majority vote rule is used to combine the results. Here, we implement random forests using scikit-learn.\textsuperscript{62} Random forests have been used in a variety of benchmark studies. In 2003, Feuston et al. used random forests to build regression and classification models to predict quantitative and categorical biological activity based on quantitative descriptions of compounds’ molecular structures.\textsuperscript{63} Further, in 2019, Gitter et al. published a study using random forests to recover active molecules assay against Pria-SSB from a library of 22 000 molecules.\textsuperscript{64} In another study, Gitter et al. looked at using random forests and molecular fingerprints to classify drug functions.\textsuperscript{65}

#### Model Development

Model parameter selection was done using 10-fold cross validation on the training sets. Separate hyperparameter grid searches were performed for each model. The best set of hyperparameters were selected based on cross-validation accuracy (see Table 20).

### Table 20. Random Forest Classifier Hyperparameters Explored by Grid Search

| parameter          | value         |
|--------------------|---------------|
| no. estimators     | 50, 100, 200, 400, 800, 1000 |
| max features       | “auto,” “sqrt,” “log2” |
| max depth          | 10, 35, 60, 85, 110, None |
| min sample leaf    | 1, 2, 4       |
| min sample split   | 2, 5, 10      |
| bootstrap          | true, false   |

The training and test sets for each model were generated using randomized 80/20 train/test splits. The hold-out test sets were left out of training and model selection steps and reserved for model evaluation. To avoid overfitting, we did our best to satisfy the “1 in 10 rule,” which states that values of events per variable (EPV) greater than or equal to 10 reduces bias in model training.\textsuperscript{66} Trained models were evaluated using the hold-out test set. For each model, we report the following performance metrics from scikit-learn 0.21.3: confusion matrix metrics, accuracy, balanced accuracy, Matthew’s correlation coefficient, receiver operating characteristic area under the curve (ROC AUC), and average precision. For multiclass classification models, we report the one-vs-rest comparisons for average precision and ROC AUC for each class (see Supporting Information).\textsuperscript{67}

For the purpose of training and testing the models, we built binary and ternary classification models for predicting bioactivity with each target (see Table 21 for activity score classifications).

### Table 21. Bioactivity Grade\textsuperscript{64} to Activity Score Conversion Table for Binary and Ternary Classifications Used in the Model Generation

| activity score | bioactivity grade | activity score | bioactivity grade |
|----------------|-------------------|----------------|-------------------|
| 1              | \( \geq 8 \)      | 3              | \( \geq 8 \)      |
| 0              | \(< 8 \)          | 2              | 4–7               |
| 1              | \(< 4 \)          |                |                   |

\textsuperscript{64}See Scheme 1 for explanation of the bioactivity grade process.

### Cluster Analysis

Each of the data sets for the six selected targets was subjected to a cluster analysis using Molsoft-Chemist\textsuperscript{64}. Each set of unique compounds was loaded into MolSoft-Chemist\textsuperscript{64} as an SDF file, and the clustering was performed on the structures using the MolSoft fingerprint technology,\textsuperscript{69} based on the molecular structure alone and clustered using an unweighted pair group method with arithmetic mean.\textsuperscript{70} The distance range was adjusted to give at least five large clusters, and the centers for each cluster were noted.

### ASSOCIATED CONTENT

#### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c05303.

Tables for each target divided into the training and test sets with ChEMBL ID’s, SMILES, binary or ternary
classifier, and chemical structure cluster membership; the performance metrics from scikit-learn 0.21.3: confusion matrix metrics, accuracy, balanced accuracy, Matthew’s correlation coefficient, ROC AUC, and average precision; for multiclass classification models, we report the one-vs-rest comparisons for average precision and ROC AUC for each class; and the feature list (XLSX)

**AUTHOR INFORMATION**

**Corresponding Author**

Kelvin Cooper — KC Pharma Consulting, Sarasota, Florida 34239, United States; orcid.org/0000-0002-3196-9194; Email: kelvincooper@kcconsulting.org

**Authors**

Christopher Baddeley — CAS, A Division of the American Chemical Society, Columbus, Ohio 43210-3012, United States

Bernie French — Tassogen Inc., Westerville, Ohio 43018, United States

Katherine Gibson — CAS, A Division of the American Chemical Society, Columbus, Ohio 43210-3012, United States

James Golden — WorldQuant Predictive, New York 10017, United States

Thiam Lee — WorldQuant Predictive, New York 10017, United States

Sadrach Pierre — WorldQuant Predictive, New York 10017, United States

Brent Weiss — CAS, A Division of the American Chemical Society, Columbus, Ohio 43210-3012, United States

Jason Yang — WorldQuant Predictive, New York 10017, United States

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c05303

**Author Contributions**

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

**Funding**

The authors thank CAS and WorldQuant Predictive for their financial support of this work.

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We acknowledge the support of WorldQuant Predictive and CAS organizations. Specifically, we thank Steven Watkins, Angela Daniels, Albert Ihochi, Brendon Pittman at CAS for data compilation and curation, as well as Michael Dennis, Dana Albaiu, and Mark Grabau at CAS for their support and guidance of this work.

**ABBREVIATIONS**

3CLpro, 3-chymotrypsin-like protease; ACE-2, angiotensin-converting enzyme type 2; Alog_P, calculated log octanol to water partition coefficient; AT-1, angiotensin receptor type 2; CACO-2, human colon carcinoma cell line; ChEMBL, chemistry database of European Molecular Biology Laboratory; COVID-19, coronavirus infectious disease 2019; CX log_P, calculated log octanol to water partition coefficient; CX log_D, calculated log octanol to buffer partition coefficient; EC50, half-maximal effective concentration; FDA, Food and Drug Administration; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; HIV, human immunodeficiency virus; IC50, half-maximal inhibitory concentration; JAK, Janus kinase; Ki, binding coefficient; K_i, inhibition constant; log_D, log octanol to buffer partition coefficient; NSP, nonstructural protein; ORF, open reading frame; pEC50, negative log of the half-maximal effective concentration; pIC50, negative log of the half-maximal inhibitory concentration; pK_i, negative log of the inhibition constant; PLpro, papain-like protease; PFF, predictive feature fingerprint; PSA, polar surface area; QED, quantitative estimate of drug likeness; QSAR, quantitative structure–activity relationship; RdRp, RNA-dependent RNA polymerase; RF, random forest; RNA, ribonucleic acid; RO5, rule of five; ROCAUC, receiver operating characteristic area under the curve; S-protein, spike protein; SARS-CoV-2, severe acute respiratory syndrome due to coronavirus 2; SMILES, simplified molecular-input line-entry system; TMPRSS2, transmembrane serine protease 2

**REFERENCES**

(1) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Hu, Y.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W. A Novel Coronavirus from Patients with Pneumonia in China. N. Engl. J. Med. 2020, 382, 727–733.

(2) WHO COVID-19 Dashboard, 2021. https://covid19.who.int/ (accessed Jan 5, 2021).

(3) Wu, A.; Peng, Y.; Huang, B.; Ding, X.; Wang, X.; Ni, P.; Meng, J.; Zhang, Z.; Wang, J.; Sheng, J.; Qian, L.; Xia, Z.; Tan, W.; Cheng, G.; Jiang, T. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 2020, 27, 325–328.

(4) Khaiyani, R. A.; Safdar, M.; Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 2020, 9, No. 100682.

(5) Milken Institute FasterCures COVID-19 Treatment and Vaccine Tracker, 2020. https://covid-19tracker.milkeninstitute.org/ (accessed Sept 2, 2020).

(6) Zhou, Q.; A.; Kato-Weinstein, J.; Li, Y.; Deng, Y.; Granet, R.; Garner, L.; Liu, C.; Polschakov, D.; Gessner, C.; Watkins, S. Potential therapeutic agents and associated bioassay data for COVID-19 and related human coronavirus infections. ACS Pharmacol. Transl. Sci. 2020, 3, 813–834.

(7) Liu, C.; Zhou, Q.; Li, Y.; Garner, L. V.; Watkins, S. P.; Carter, L. J.; Smoot, J.; Gregg, A. C.; Daniels, A. D.; Jervey, S.; Albaiu, D. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. ACS Cent. Sci. 2020, 6, 315–331.

(8) Fragkou, P. C.; Belhadi, D.; Peofer-Smadja, N.; Moschopoulos, C. D.; Lescure, F. X.; Janocha, H.; Karofylakis, E.; Yordanpanah, Y.; Méntre, F.; Skevaki, C.; Laouénan, C.; Tsiaodras, S. Review of trials currently testing treatment and prevention of COVID-19. Clin. Microbiol. Infect. 2020, 26, 988–998.

(9) McKee, D. L.; Stemberg, A.; Stange, U.; Laufer, S.; Naujokat, C. Candidate Drugs against SARS-CoV-2 and COVID-19. Pharm. Res. 2020, 157, No. 104859.

(10) CAS Download CAS COVID-19 Antiviral Candidate Compounds Dataset, 2020. https://www.cas.org/covid-19-antiviral-compounds-dataset (accessed Sept 2, 2020).

(11) Hansch, C.; Maloney, P.; Fujita, T.; Muir, R. Correlation of Biological Activity of Phenoxycetic Acids with Hammet Substituent Constants and Partition Coefficients. Nature 1962, 194, 178–180.

(12) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. L.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz’min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? J. Med. Chem. 2014, 57, 4977–5010.
(13) Mathes, M.; Klingpohm, W.; Baumann, K. Chemoinformatic Classification Methods and their Applicability Domain. Mol. Inf. 2016, 35, 160–180.

(14) Bosc, N. ; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large Scale Comparison of QSAR and Conformal Prediction Methods and their Applications in Drug Discovery. J. Cheminf. 2019, 11, No. 4.

(15) Lee, K.; Lee, M.; Kim, D. Utilizing Random forest QSAR Models with Optimized Parameters for Target Identification and its Application to Target-Fishing Server. BMC Bioinf. 2017, 18, No. 567.

(16) Ma, D.-L.; Chan, D. S.-H.; Leung, C.-L. Drug repositioning by structure-based virtual screening. Chem. Soc. Rev. 2013, 42, 2130.

(17) Schneider, P.; Walters, W. P.; Ploewright, A. T.; Sieroka, N.; Lüftl, K.; Siewert, T.; Goffin, K.; Schiergens, T. S.; Herrler, G.; Wu, N. H.; Nitsche, A.; Müller, M. A.; Drosten, C.; Pohlmeyer, S.; SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 is blocked by a clinically proven protease inhibitor. Cell 2020, 181, 271–280.

(18) Kamat, M.; Yamanoto, N. HIV protease inhibitor nelfinavir inhibits replication of SARS-associated coronavirus. Biochim. Biophys. Acta. 2004, 1719, 719–725.

(19) Savarino, A. Expanding the frontiers of existing antiviral drugs: Possible effects of HIV-1 protease inhibitors against SARS and avian influenza. J. Clin. Virol. 2005, 34, 170–178.

(20) Yin, W.; Maa, C.; Luan, X.; Shen, D.-D.; Shen, Q.; Su, H.; Wang, X.; Dai, F.; Zhao, W.; Gao, C.; Cui, Y.; Xie, Y.-C.; Tian, G.; Jiang, H.-W.; Tao, S.-C.; Shen, J.; Jiang, Y.; Jiang, H.; Xu, Y.; Zhang, S.; Xu, X. E. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science 2020, 368, 1499–1504.

(21) Venkataramanan, S.; Prasad, B. V. L. S.; Selvarajan, R. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. Viruses 2018, 10, No. 76.

(22) Kuba, K.; Imai, Y.; Rao, S.; Gao, H.; Guo, F.; Guo, Y.; Huang, Y.; Yang, P.; Zhang, Y.; Deng, W.; Bao, L.; Zhang, B.; Liu, G.; Wang, Z.; Chappell, M.; Liu, Y.; Zheng, D.; Leibbrandt, A.; Wada T.; Slutsky, A. S.; Liu, D.; Qin, C.; Jiang, C.; Penninger, J. M. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. Nat. Med. 2005, 11, 875–879.

(23) Swayamsiddha, S. Application of Artificial Intelligence in COVID-19 with Optimized Parameters for Target Identification and its structure-based virtual screening. J. Cheminf. 2020, 12, 1031.

(24) Jose, R. J.; Manuel, A. COVID-19 cytokine storm: the interplay between inflammation and coagulation. Lancet Respir. Med. 2020, 8, E46–E47.

(25) Song, P.; Li, W.; Xie, J.; Hou, Y.; You, C. Cytokine storm induced by SARS-CoV-2. Clin. Chem. Acta. 2020, 509, 280–287.

(26) Bikdeli, B.; Madhavan, M. V.; Jimenez, D.; Chaich; T.; Dreyfus, I.; Driggin; E.; Déng; M. V.; Majid, M.; Guo, Y.; Tang, L. V.; Hu, Y.; Giri, J.; Cusmano, M.; Quré; I.; Dimakakos, E. P.; Gibson, M.; Lippi, G.; Favaloro, E. J.; Fareed, J.; Caprini, J. A.; Tafur, A. J.; Burton, J. R.; Francesc, D. P.; Wang, E. Y.; Falanga, A.; McIntock, C.; Hunt, B. J.; Spyropoulos, A. C.; Barnes, G. D.; Eikelboom, J. W.; Wei, Y.; Weinberg, I.; Schulman, S.; Carrier, M.; Piazza, G.; Beckman, J. A.; Steg, G.; Stone, G. W.; Rosenkranz, S.; Golub, S. Z.; Parikh, S. A.; Monreal, M.; Krumholz, H. M.; Konstantinides, S. V.; Weitz, J. I.; Lij; G. Y. COVID-19 and Thorbomtic or Thromboembolic Disease: Implications for Prevention, Antithrombotic Therapy, and Follow-Up. J. Am. Coll. Cardiol. 2020, 75, 2950–2973.

(27) Schneider, P.; Walters, W. P.; Ploewright, A. T.; Sieroka, N.; Lüftl, K.; Siewert, T.; Goffin, K.; Schiergens, T. S.; Herrler, G.; Wu, N. H.; Nitsche, A.; Müller, M. A.; Drosten, C.; Pohlmeyer, S.; SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 is blocked by a clinically proven protease inhibitor. Cell 2020, 181, 271–280.

(28) Kamat, M.; Yamanoto, N. HIV protease inhibitor nelfinavir inhibits replication of SARS-associated coronavirus. Biochim. Biophys. Acta. 2004, 1719, 719–725.
