

Phytochemical Drug Discovery for COVID-19 Using High-resolution Computational Docking and Machine Learning Assisted Binder Prediction

Zirui Wang^{1,2}, Theodore Belecciu^{1,2}, Joelle Eaves^{1,2}, Michael H. Bachmann^{1,3}, Daniel Woldring^{1,2,*}

¹Institute for Quantitative Health Science and Engineering, Michigan State University, 775 Woodlot Dr, East Lansing, MI 48824, USA

²Department of Chemical Engineering and Materials Science, Michigan State University, 428 S Shaw Lane, East Lansing, MI 48824, USA

³Department of Microbiology and Molecular Genetics, Michigan State University, 567 Wilson Rd, East Lansing, MI 48824, USA

* Email: woldring@msu.edu

KEYWORDS

COVID-19, SARS-CoV-2, Drug Discovery, Ligand Docking, Cheminformatics, Natural Products, Phytochemicals, Virtual Screening, Machine Learning

ABSTRACT

The COVID-19 pandemic has resulted in millions of deaths around the world. Although multiple safe and effective vaccines and some pharmaceuticals have been approved for use, the problem is still unsolved for individuals with underlying medical conditions and those living in underserved areas that lack vaccines and/or an adequate medical infrastructure. This is especially challenging as new variants of SARS-CoV-2 emerge. One possible approach to solving this problem lies in using naturally abundant phytochemicals generally regarded as safe that bind to and disrupt SARS-CoV-2. When used in conjunction with a polypharmacological approach, targeting multiple essential viral proteins can lead to stronger functional inhibition and provides a safeguard against escape mutations. Although finding the proper phytochemicals to accomplish a specific therapeutic task is challenging and costly, *in-silico* screening methods have made this a more tractable problem by expediting the initial lead compound discovery phase. Recent studies have gained mechanistic insights of drug interactions through computational docking against select SARS-CoV-2 proteins, yet several viral proteins remain unexplored as druggable targets. Here we investigate a wide range of drug products against a comprehensive array of SARS CoV-2 proteins using a high-resolution docking workflow. Our initial lead compound discovery phase consisted of a structure-based virtual screening (SBVS) wherein 10 types of structural and non-structural SARS-CoV-2 proteins were computationally docked against a panel of anti-viral phytochemicals from the USDA Phytochemical and Ethnobotanical Databases. In the second phase of the study, we employed ligand-based virtual screening (LBVS) by extracting chemical

features of 34 lead compounds from the SBVS using unsupervised clustering based on common motifs. Features among dominant ligand clusters were then used to prioritize subsets of additional phytochemical databases for drug discovery. Among the 53 newly identified phytochemicals generated via LBVS, high-resolution docking predicted that 28 elicit strong binding interactions with SARS-CoV-2 proteins. Thus, the inclusion of LBVS resulted in a 4-fold increase in the rate of lead discovery. Finally, drug-likeness of all lead compounds and phytochemical sourcing was evaluated. As a result, this three-phase workflow gave rise to 18 flavone, alkaloid, and anthraquinone phytochemicals with the greatest potential for therapeutic utility. Among these phytochemicals, multiple lead compounds with favorable drug-likeness can be derived from individual plants (e.g. *Camptotheca acuminata* and *Mahonia japonica*). Collectively, this study demonstrates the exciting potential of plant-based drug development for COVID-19 prevention and treatment using a polypharmacological approach. These findings further support the advantage of incorporating machine learning elements into a virtual screening workflow.

INTRODUCTION

Since its start in December 2019, the COVID-19 pandemic has caused more than five million deaths worldwide,¹ long term health effects in many who have recovered from acute infection,² and severe global economic damage. Moreover, zoonotic disease-driven pandemics are likely to become more prevalent over time.³ Thanks to the extraordinarily rapid development of vaccines, transmission of COVID-19 has been prevented and its symptoms have been greatly reduced in a large fraction of global societies, albeit large inequities remain. As the virus mutates, however, and modifies its structural and functional components, existing vaccines may become less effective.⁴ Vaccines developed against the early variants of SARS-CoV-2 are already less effective against the more infectious delta and omicron variants, hindering progress toward herd immunity.^{5,6} Identifying chemicals with therapeutic potential against SARS-CoV-2 and related viruses would (1) provide additional protection, for even the vaccinated members of the global community, (2) offer supplementary treatment options for individuals with medical conditions or personal beliefs that preclude vaccine use, and (3) provide treatment for less developed regions of the world. Some antiviral drugs, namely the polymerase inhibitors Remdesivir (GS-5734, Veklury) and β -D-N4-hydroxycytidine (NHC, Molnupiravir, Merck), as well as the protease inhibitor PF-07321332 (Nirmatrelvir, Pfizer), have received at least an EUA (emergency use authorization) from the FDA⁷. However, mono-drug therapy, as the HIV/AIDS epidemic taught us, carries in the risk of rapid development of drug resistance.⁸ Hence, combinational drug therapy that simultaneously targets several viral proteins and possibly also benefits host anti-viral and anti-inflammatory mechanisms is desirable as it would reduce viral replication and could delay, if not abrogate, the development of resistant variants.⁹

Thus, therapeutic drug regimens will ideally be (i) effective against multiple arising variants of SARS-CoV-2 as well as (ii) quickly accessible for disparate communities across the globe. Identifying such therapeutics is a critical, yet challenging endeavor due to the resource-intensive process of drug development and the rate at which many infectious disease agents mutate to evade these same treatments. A potential solution for the first challenge of developing a broadly effective therapeutic is through polypharmacology (i.e., using a cocktail of drugs that target multiple distinct protein functions of the virus).¹⁰ Polypharmacology has shown remarkable results for other devastating diseases such as HIV.¹¹ A key advantage of this combinatorial drug approach is that the virus would need to undergo multiple simultaneous mutations in order to become resistant to each individual drug in the combination.

Plant-derived phytochemicals that are generally regarded as safe (USDA GRAS), are an attractive resource for drug development. We focus our efforts on them because, provided they are used in physiological doses, they would not require pre-clinical animal testing nor phase I and II safety trials in humans and are widely accessible to many global communities. Phase II efficacy trials could be implemented rapidly upon development of a reproducible protocol. Among the many tens of thousands of diverse compounds produced by plants, hundreds of these phytochemicals have already been identified as having antiviral, antibacterial, and anti-inflammatory properties.¹² Thus, antiviral phytochemicals offer a promising starting point for the screening and discovering of specific drugs that are effective against SARS-CoV-2.

Computational 3D docking has produced a surge of advances, in part, due to the continued rise in processing power, refinement of score functions, and increased availability of high-resolution molecular structure data. Thanks to the fast and fierce response by the scientific community, more than 1500 structures of the structural and non-structural protein components of SARS-CoV-2 have been generated and made publicly available.¹³ Using a combination of crystallographic and modeled structures, recent studies have explored the use of computational simulations to identify small molecules that bind to SARS-CoV-2 proteins. Much of this work has focused on inhibition of the main protease (Mpro)^{14 15 16 17} as well as the RNA-dependent RNA polymerase¹⁸, spike protein¹⁹, and replicase²⁰. Further work describes the potential for phytochemicals to make a positive impact on treating COVID-19 and provide evidence for benefits elicited from flavonoids²¹, polyphenols²², and alkaloid drugs²³.

The advent of machine learning (ML) in drug discovery and development has also facilitated and accelerated predictive processes through the use of Bayesian models²⁴, structure-based algebraic topology²⁵, convolutional neural networks²⁶, and transfer learning²⁷. The application of ML has prevailed in various stages including target identification and validation, compound screening and lead discovery, preclinical development, and clinical development.²⁸

In this study, we use the extensive structural datasets in combination with a refined and annotated collection of anti-viral phytochemicals to evaluate which naturally derived medicines have the highest potential for evoking strong binding interactions to SARS-CoV-2 proteins to

preclude or disrupt the viral infection process. Previous docking studies have typically examined one or a few protein targets. Here, we screened several non-structural proteins (NSP1, NSP3, NSP5, NSP7, NSP8, NSP9, NSP10, NSP13, and NSP15) and two forms of the structural spike protein (the receptor binding domain and the full-length spike) because of their essential contributions to viral replication and infection. For instance, the main protease (NSP5) is responsible for cleaving individual SARS-CoV-2 protein chains from a translated polyprotein chain.²⁹ The helicase (NSP13) has an essential role in viral replication due to its function in unwinding RNA and DNA.³⁰ We used the modeling software Rosetta to conduct ligand docking simulations (structure-based virtual screenings or SBVS), to obtain the estimated docking free energies between anti-viral phytochemicals and proteins. Analyzing the distributions of the docking energy scores for each protein, we identified lead compounds with high affinity toward individual protein structures.

Because of the time-consuming nature of high-resolution docking simulations, it was infeasible to run SBVS for all phytochemicals of interest. Therefore, we implemented machine learning algorithms to predict potential leads from a second large phytochemical library. We used unsupervised learning to cluster already screened anti-viral phytochemicals, aiming to extract chemical features of identified leads. Then, we employed supervised learning to classify the un-screened phytochemicals from the large library into already-formed clusters. With computationally identified leads from the docking simulations, we recognized lead clusters by ranking the total or relative abundance of lead phytochemicals within each cluster. We then applied our clustering algorithm to the large unscreened library. Among the library compounds, only those classified into our lead clusters were subjected to docking simulations to evaluate their ability to bind SARS-CoV-2 protein targets (Figure 1). Overall, our study has identified 62 lead compounds that may inhibit one or more SARS-CoV-2 proteins. Eighteen of those leads show promising results in a SwissADME drug screening. This investigation also demonstrated that the use of machine learning significantly speeds up the ligand screening process, giving rise to a 4-fold increase in lead compound yield.

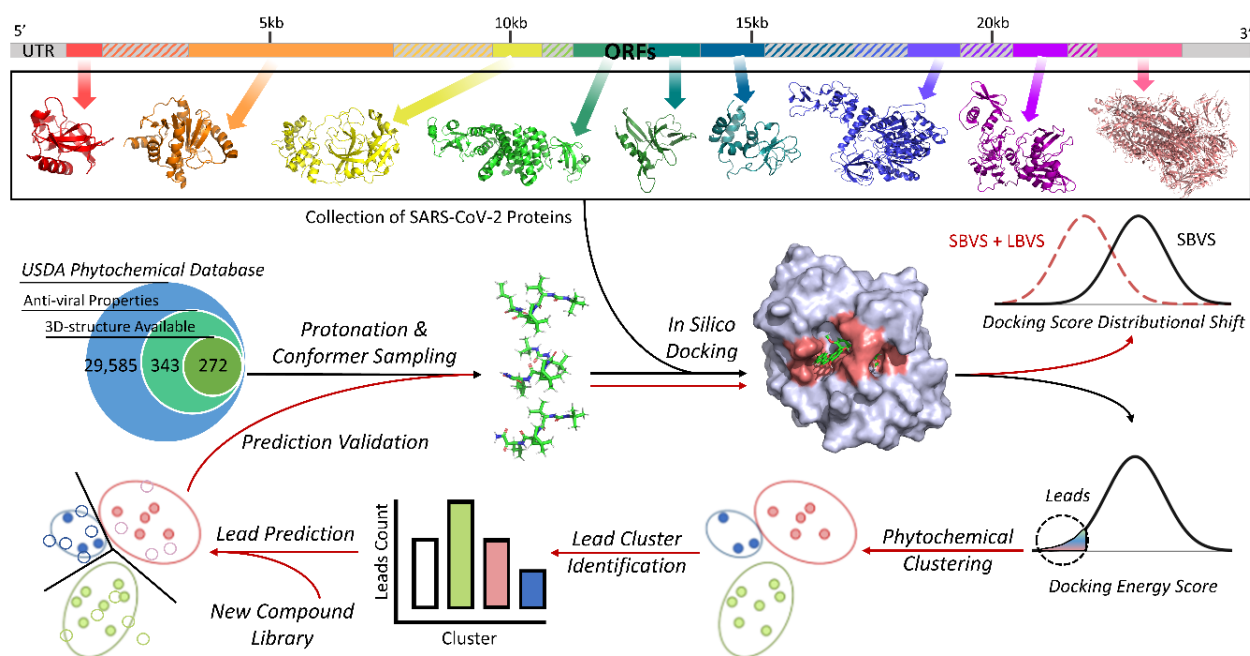


Figure 1. Overview of the structure and ligand-based virtual screening workflow. Numerous SARS-CoV-2 protein structures and 272 anti-viral phytochemicals were prepared for the Rosetta protein-ligand docking (SBVS). Lead phytochemicals were chosen based on the highest performing (lowest docking energy) simulations. The entire phytochemical library was clustered according to chemical similarity. Lead clusters were identified as clusters having the highest proportion of lead phytochemicals. The ligand-based virtual screening (LBVS, red lines within workflow) clustered phytochemicals from a second distinct database. New phytochemicals classified as belonging to lead clusters were identified and subjected to high-resolution structural docking. The ligand based virtual screening began only once the structural based virtual screening of the 272 initial phytochemicals was completed and the molecule clusters were established.

METHODS

Ligand Preparation for *In Silico* Docking

The Rosetta protein-ligand docking protocol requires two inputs: a PDB file containing the protein and ligand structures, and a params file. A list of 343 antiviral phytochemicals was obtained from the USDA Phytochemical and Ethnobotanical Databases.¹² Three-dimensional structures of 272 of these phytochemicals were downloadable from the ZINC and PubChem in SDF format for the initial SBVS. OpenBabel,³¹ a chemical file conversion and manipulation tool, was used to protonate ligand structures for their configuration at a physiological pH of 7.4. Ligand conformational space sampling was performed using the BCL::Conf application.³² This application generates 100 conformers for each ligand by segmenting the ligand into fragments

and recombining them based on information contained in a small molecule fragments database.³³ Afterwards, a Python script in the Rosetta package named “molfile_to_params.py” was used to generate a params file and a ligand PDB file.³⁴

Protein Preparation for *In Silico* Docking

All SARS-CoV-2 protein structures (Figure 6B) were obtained from the Protein Data Bank.¹³ When multiple structures existed for a single protein, priority was given to those with higher resolution. Structural files were cleaned by removing unnecessary components such as water molecules, solvated ions, and non-targeted oligomers. Lastly, the cleaned protein structures were concatenated with the ligand PDB files prior to docking.

Protein-Ligand Binding Site Prediction

To locate potential binding sites on our proteins prior to the docking runs, we utilized the CASTp (Computed Atlas of Surface Topography of Proteins) webserver to obtain pocket structural information and the center coordinates of each unit sphere that comprised the pockets.³⁵ CASTp applies geometric techniques to identify surface pockets and internal cavities within a protein structure (Figure 3A). Two metrics (pocket volume and surface area) were employed to sample CASTp-identified pockets, since sampling each of the numerous pockets during docking would have been computationally unfeasible. Once potential pockets were determined, the center coordinates of the spheres that made up the pockets were used as initial coordinates for high-resolution docking.

Binding pocket sampling criteria were established based on the statistics of binding pockets of protein-ligand complexes obtained from the CASF-2016 dataset containing 285 unique crystal structures (Figure S1, Table S1, and Table S2).³⁶ These criteria first rank all pockets for a particular protein structure by volume from largest to smallest, and then compare each pocket to the first, largest volume pocket. If any of the smaller binding pocket volumes were less than 10% of the largest pocket volume, then their surface areas were compared to the surface area of the previously ranked pocket. Such small binding pockets were only considered potential binding pockets if their surface areas were larger than that of the previously ranked pocket. All other pockets with volumes smaller than 10 % of the largest pocket’s volume were not considered and not sampled during docking.

Two separate methods for binding site coordinate extraction were developed: one optimized for smaller pockets (Figure 3A in green), and another for large pockets (Figure 3A in red). For small pockets, defined by volumes less than 1000 Å³, the center of the largest sphere within that pocket was extracted as a starting coordinate for the docking simulation. For large pockets, multiple starting coordinates were extracted (Figure 3B in red). These coordinates were the centers of spheres within the pocket whose volumes were larger than 5% of the total pocket volume. The distance between pairs of coordinates also had to be at least 30 Å to avoid sampling space overlap during docking simulations. All chosen coordinates within the potential binding pockets were embedded in the “start_from” mover in the Rosetta docking script.³⁷

Docking Data Analysis

One thousand models were generated for every binding pocket in each protein-ligand docking event, with each model supplying data that describes its docked structure. Among the data from the simulations, the index “Interface_delta_X” (energy score) was used to indicate the free energy of the binding event. Because binding likelihood is inversely related to the energy score, the lowest energy score from all model scores generated for a given protein-ligand pair was used to represent the binding favorability of that docking. We performed exploratory data analysis on all lowest scores for each protein structure, and we fit these scores to a normal distribution per protein structure. Phytochemicals with scores at least two standard deviations below the average of all the compounds’ scores for a specific protein were designated as lead candidates against that specific protein.

If a crystal structure was available to use for model error calculations, the index “ligand_rms_no_super_X” (the root mean square deviation or RMSD of atomic positions without superimposition) was used to measure the difference between the model structure and the crystal structure. This was applicable to our evaluation of different Rosetta score functions.

Rosetta Score Function Testing on SARS-CoV-2 Structures

Score functions are used to calculate the energies of proposed biomolecules during each step of the docking simulation. A score function is the sum of weighted energy terms that include both physical forces and statistical parameters. In order to determine which score function was best suited for this study, we tested multiple Rosetta score functions (RosettaLigand, Talaris2014, Ref2015, and Betanov16) on 10 SARS-CoV-2 main protease and NSP3 ligand-bound crystal structures from the Protein Data Bank (Table S3). Since Rosetta docking excludes water influence, an additional protocol, Rosetta-ECO (efficient consideration of coordinated waters) was also included.³⁸ The weights for each term in score functions were obtained from Smith et al.³⁹ Our docking script example is provided in Supporting Materials (Figure S3).

Based on the fact that model energy scores tend to correlate positively with RMSDs, we used two methods to evaluate the performance of different score functions. The first evaluation strategy involved comparing the Spearman correlation coefficients between energy scores and RMSDs. The second approach involved comparing the case percentages where the energy score of the lowest RMSD model was within the 10th and 20th percentiles of the entire energy score range for a specific protein-ligand pair docking.

Phytochemical Structure Embedding

To quantitatively cluster and classify molecules, phytochemicals were converted into numerical form. We used a circular molecular fingerprint method, extended-connectivity fingerprints (ECFPs),⁴⁰ to generate molecular descriptors that store the structural information of a given molecule. These descriptors are then mapped on a 1024-bit vector, each bit indicating the appearance of a specific feature within a molecule.

ECFPs treat each atom in a molecule as a center and iteratively examine immediate neighbors with increasing scope. A hash function is used to produce an identifier (hash value) that describes structural features. Identifiers from the previous iteration serve as the input for the subsequent generation of a new identifier that encompasses more of the molecular structure. For example, a single atom is examined during iteration zero and the input (i.e., the initial identifiers) are six properties of that atom, which are the daylight atomic invariants: the number of heavy atom connections, the number of hydrogen bonds, the atomic number, the atomic mass, the atomic charge, and the number of attached hydrogens.⁴⁰ These invariants are hashed into an identifier which stores information from the chosen atom. In the next iteration (iteration one), the identifiers of connecting atoms are hashed into a new identifier which describes the structural information of the whole expanded neighborhood. The list of identifiers is updated each time when progressively larger circular substructural neighborhoods are included. The iteration proceeds until it reaches a user-specified number of iterations, or until no new identifier is generated. Duplicated identifiers in the list were removed or counted.

Once all identifiers were obtained, the remainders from the division of each identifier by 1024 were the vector indexes where the bit is 1. By these means, we obtained a fixed-length vector (1024-bits) where 0 and 1 indicate the absence and presence, respectively, of identifiers.

Unsupervised Phytochemical Clustering

Unsupervised learning is a type of machine learning that identifies data patterns in unlabeled data. We used the algorithm from Sci-Kit library⁴¹ to cluster our structure-screened anti-viral phytochemicals by structural similarities given only their feature representations (0s and 1s). Four clustering methods were compared for our clustering analysis.

1. Agglomerative Hierarchical Clustering with the Ward linkage criterion. We utilized a bottom-up approach where each molecule starts in its own cluster and newly formed clusters are successively merged together until one root cluster is formed. Cluster centroids were computed to represent formed clusters and were used to calculate the Euclidean distance between clusters. The merging of the clusters is determined by the Ward linkage criterion, which minimizes the error sum of squares (ESS) for all clusters.^{41,42}

2. Spectral Clustering. This method first builds a graph $G(V, E)$ connecting vertices (data points) if the edge (similarity) is positive or above a certain threshold. Subsequently, the graph Laplacian matrix is computed by subtracting the adjacency matrix from the adjacency matrix. Using the eigenvectors and eigenvalues of the graph Laplacian, the graph is embedded into a low dimensional space, where the clustering algorithm is applied to partition the embedding by clustering the components of eigenvectors.^{41,43}

3. Affinity Propagation (AP). This clustering method is determined by messages (values) sent between data points. The first message is responsibility, $R(i, k)$, which is the evidence that data point k should be the cluster center (exemplar) for data point i . The second is availability,

$A(i, k)$, which is the evidence that i chooses k to be the exemplar. These messages are updated iteratively between pairs of points until convergence. Until then, the final exemplars are chosen, and clusters are formed.^{41,44}

4. Ordering Points to Identify Cluster Structure (OPTICS). This method shares many commonalities with Density-based Spatial Clustering of Applications with Noise (DBSCAN). However, unlike DBSCAN which assumes the constant density of clusters, OPTICS allows varying densities of clusters. The idea of density-based clustering is that an area with center p and radius ε has to contain a minimal number of objects (MinPts). The cluster order is represented by the core distance and the reachability distance. The core distance of an object p is the smallest radius ε' of the circular area that contains MinPts. The reachability distance is the higher value between the core distance and the direct distance between points, written as: $\max(\text{core distance}(o), \text{distance}(o, p))$.⁴⁵ OPTICS produces the ordering of the data points using the reachability distance to delineate cluster structures. The Tanimoto index was used in our OPTICS analysis for distance computation.⁴⁶

After molecule clustering, we used the Shannon entropy (Eq 1) to measure the distribution of phytochemicals among all clusters. A high Shannon entropy suggests that the sizes of clusters are relatively similar, and it can avoid false accuracy of imbalanced classification.

$$H(X) = - \sum_{i=1}^n P(x_i) \log_n P(x_i) \quad (\text{Eq.1})$$

In the above equation, n is the number of clusters and $P(x_i)$ is the fraction of molecules in cluster i over all clustered molecules.

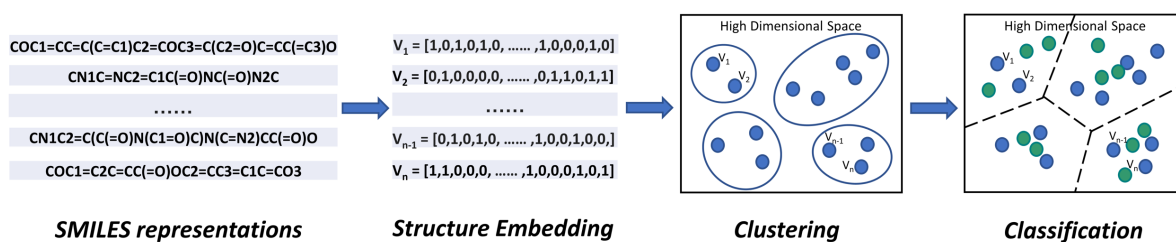


Figure 2. Phytochemical Clustering and Classification Scheme. The ECFP algorithm was used to encode molecule structural information to fixed-length vector representations. The molecule clustering is based on the distance calculation of vector representations of molecules. The un-clustered molecules (green) were classified into already-formed clusters by supervised learning.

Supervised Classification for Potential Lead Prediction

The fraction of identified lead phytochemicals in each cluster was determined and clusters with the highest fractions were labeled as lead clusters. Through supervised learning, a classifier was built to classify new phytochemicals that had not undergone high-resolution docking into already-formed clusters. We predicted that phytochemicals classified as belonging to lead clusters would be potential lead phytochemicals.

Supervised learning uses labeled datasets to learn the mapping function from inputs (features) to outputs (labels). In our case, the features were the 0s and 1s contained in each molecule-describing vector and the labels were their cluster IDs. The 272 structure-screened phytochemicals were split into 80% training and 20% testing sets. The classification accuracy rate was obtained from the testing sets only. In order to get a high accuracy rate, we compared four classification methods: K-nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Linear Discriminant Analysis (LDA). KNN uses distance metrics to compute the distance between data points and classify them based on the majority votes of their surrounding k neighbors.⁴⁷ In our model, k was chosen to be three, and the distance metric used was Tanimoto.⁴⁶ The weights in the weight function for points closer to neighbors are higher than the weights for points further away. SVM classifies data points by moving data to a high dimensional space, where the soft margin between classes is maximized. Hyperplanes were created to separate classes.⁴⁸ Radial Basis Function (RBF) was used to transform features to a high dimensional space. RF is an ensemble learning method that generates many classifiers (decision trees) and takes the majority votes of generated classifiers to predict the final outcome.⁴⁹ LDA is a Gaussian maximum likelihood classification method that assumes each class is under a Gaussian distribution. The estimate means and covariances were obtained directly from the data. LDA classifies new observations by creating a dimension where the means of projected classes are maximally separated and the variance within each class is minimized.⁵⁰

RESULTS AND DISCUSSION

Global docking is accurately guided by CASTp pocket identification. Prior to performing high-resolution docking between our phytochemical libraries and the individual SARS-CoV-2 protein structures, CASTp software was employed to identify concave regions of the protein surface that may facilitate ligand binding.³⁵ We hypothesized that limiting the docking search space to highly solvent exposed concave crevices (pockets) would sufficiently capture the true location of most small molecule binding interactions while significantly reducing the necessary computational time required for iterative high-resolution docking. To test this hypothesis, we used the CASF-2016 dataset³⁶ – composed of 285 crystal structures of reliably characterized protein-ligand complexes – to quantify how often the largest protein surface cavities are involved in ligand binding interactions. The numerous surface cavities of each CASF-2016 structure were

calculated and ranked by volume using the CASTp webserver. We calculated the frequency of ligands binding to the ranked surface cavities (Figures 3C, S1). This resulted in 87% (247 /285) of the ligand binding events occurring in the either the largest or second largest pocket by volume, whereas only 2% (7 /285) of the true binding pockets were not identified via CASTp. Following this validation, we analyzed each of our 15 SARS-CoV-2 protein structures (Figure 6B) using CASTp to obtain the configuration of each available pocket, described by an aggregation of small spheres (Figure 3A). We extracted the central coordinates from only the pockets that met our selection criteria mentioned in Methods, and used these coordinates for the initial ligand locations during high-resolution docking (Figure 3B).

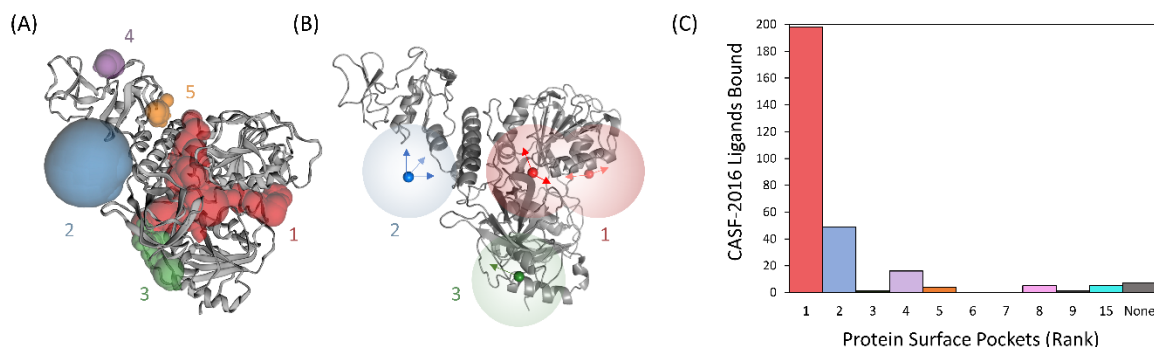


Figure 3. Binding pocket identification and ranking. (A) Concave crevices (pockets) along the protein surface are calculated using CASTp. Distinct pockets are individually colored throughout the SARS-CoV-2 helicase (PDB: 7NIO), shown here. Pockets are numbered according to pocket volume rank. (B) Central coordinates of the largest pockets determine the initial placement of phytochemical ligands during high-resolution docking. The shaded colored regions indicate the approximate space sampled by the ligand during docking. Multiple docking regions were explored for pockets having volumes greater than 1000 \AA^3 . In the example shown, pocket 1 (red) is subdivided into two spheres. (C) Method validation was conducted using 285 solved protein-ligand complex crystal structures from the CASF-2016 dataset (further details are given in the Supporting Information). The histogram shows the number of true binding events (y-axis) occurring at each CASTp ranked pocket (x-axis). As mentioned, 87% of the binding events occurred in the first largest (red) and second largest (blue) pockets by volume.

Rosetta Score Function Validation for SARS-CoV-2 Protein Docking.

During protein-ligand docking, score functions are responsible for accurately capturing the physicochemical contributions of macromolecular complexes. RosettaLigand (pre-talaris 2013) was chosen to be the score function for our SBVS after testing multiple score functions on two different SARS-CoV-2 protein structures. We first measured the fitness of the score functions using the Spearman correlation coefficients between the Rosetta energy scores and RMSDs. A high positive correlation indicates that the score function successfully captured the direct relationship between RMSD models and corresponding energy scores. The distribution of correlation coefficients for 20 tested SARS-CoV-2 structures showed that the holistic

distributions of RosettaLigand, Talaris2014, and Ref2015 are higher than the distributions of Betanov16 and Rosetta-ECO (Figure 4A). Next, we checked the energy score percentile of the lowest RMSD model for each docking event, assuming that good score functions generate models with both energy scores and RMSDs occurring in a low percentile. Figure 4B displays the case percentage of the lowest RMSD model whose energy score was within the lowest 10% and 20% of the entire energy score range. The RosettaLigand score function outperformed the others with 50% of cases where the lowest RMSD structures are within the lowest 20% of the entire energy score range, and 30% of cases where structures are within the lowest 10%. Smith et al. have performed a similar, but more comprehensive score function comparison on a large set of well-studied protein-ligand complexes.³⁹ Their results also indicate that RosettaLigand performs the best overall. (Detailed testing data available at: https://ziruiw.shinyapps.io/score_functions_on_sarscov2/).

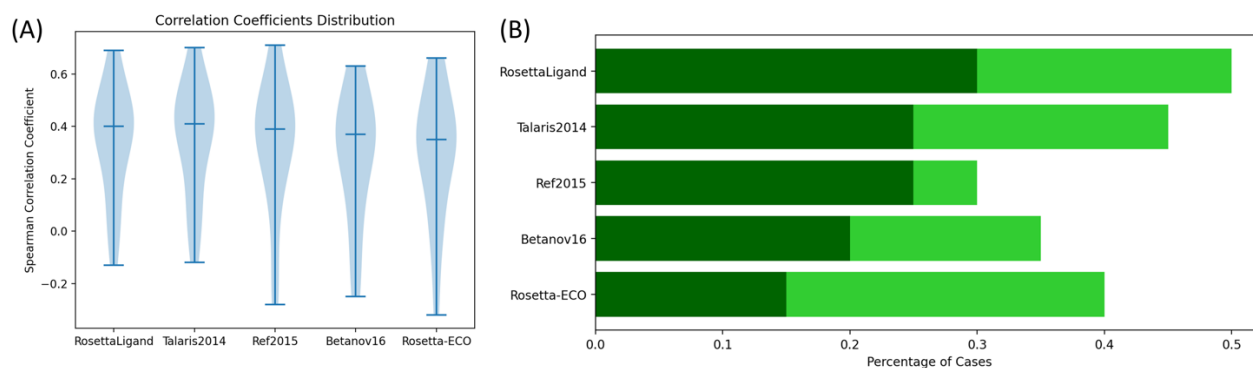


Figure 4. Score function testing results. (A) Distributions of Spearman correlation coefficients between energy score and RMSD produced by five different Rosetta score functions. (B) Percentage of cases for each analyzed score function where the energy scores of the lowest RMSD model structures are within the lowest 10% (dark green) and the lowest 20% (light green) of the energy score range.

Clustering and Classification of Phytochemical Ligands. The Ward Hierarchical Clustering method and Random Forest method were selected to cluster and classify phytochemicals. Because the prediction is largely determined by classifying molecules, the classification accuracy rate is a key indicator to measure the performance of different models. High Shannon entropy is another key attribute of a good model because it demonstrates evenly distributed classes that can avoid imbalanced classification. Model hyperparameters were tuned with different classification methods in order to obtain the best results (Figure 5A). Principal component analysis (PCA) was applied to reduce the 1024-dimensional molecule representation to a 2-dimensional representation for a visualization of clustering results (Figure 5B). The color of each data point in Figure 5B indicates its cluster. The molecule points colored in black were treated as noise, meaning they were in a group that did not belong to any of the clusters formed by the similarity search.

The hyperparameter tuned for Ward hierarchical and spectral clustering was the number of clusters. The best performing classification methods for hierarchical and spectral clustering are RF and KNN respectively. The accuracy rate decreased, while the Shannon entropy increased with an increasing number of formed clusters for both clustering methods. The accuracy rate dropped from 95% to 83% and from 96% to 71% for Ward hierarchical and spectral clustering, respectively. When increasing the number of clusters from 10 to 60, the Shannon entropy increased from 0.87 to 0.93 and from 0.1 to 0.57 for Ward hierarchical and spectral clustering, respectively. This trend supports the inference that a higher misclassification rate occurs when more clusters are formed. Because more clusters formed with a certain number of molecules, they were more evenly distributed among all clusters. However, the overall Shannon entropy for spectral clustering was low due to a large portion of molecules classified as noise.

We next tuned the damping factor for affinity propagation clustering. The damping factor is the degree to which the current value is maintained relative to incoming values and is used to avoid numerical oscillations when updating values.⁴¹ The overall accuracy of this model is not as good as the accuracy of Ward hierarchical method. Damping factors in the range [0.59, 1) had no effect on the clustering outcome, as was indicated by the constant Shannon entropy. When the damping factor was beyond 0.79, only one cluster formed; therefore, the multiclass classification could not be performed. Since the affinity propagation clustering depends on the values (availability and responsibility) sent between pairs of data points, the total cluster number is determined by the provided data rather than the users. Thus, we were not able to tune the number of clusters for this method.

For the last clustering method, OPTICS, the minimal samples parameter (MinPts) was tuned. MinPts is the minimal number of points in a neighborhood used to consider a point as a core point.⁴⁵ The KNN and RF classification methods generated a higher accuracy than SVM and LDA. With the increasing MinPts from two to nine, the accuracy rate increased from 0.62 to 0.8 and 0.62 to 0.75 for RF and KNN, respectively. However, the Shannon entropy decreased from 1 to below 0.3. This suggests that when more points are needed to decide a core point (cluster centroid), fewer clusters are formed which makes classification easier. However, using this density-based clustering method, many molecules are categorized as noise.

Comparing different methods, we concluded that Ward hierarchical clustering with random forest classification produced the best result with 52 clusters formed, an 88% accuracy, and 0.943 Shannon entropy. Spectral clustering and OPTICS treated many molecules as noise indicated by the black data points, and Affinity Propagation generated skewed cluster size indicated by its color distribution (Figure 5B). The details of the molecule clustering results are in the Supporting Information Table S9.

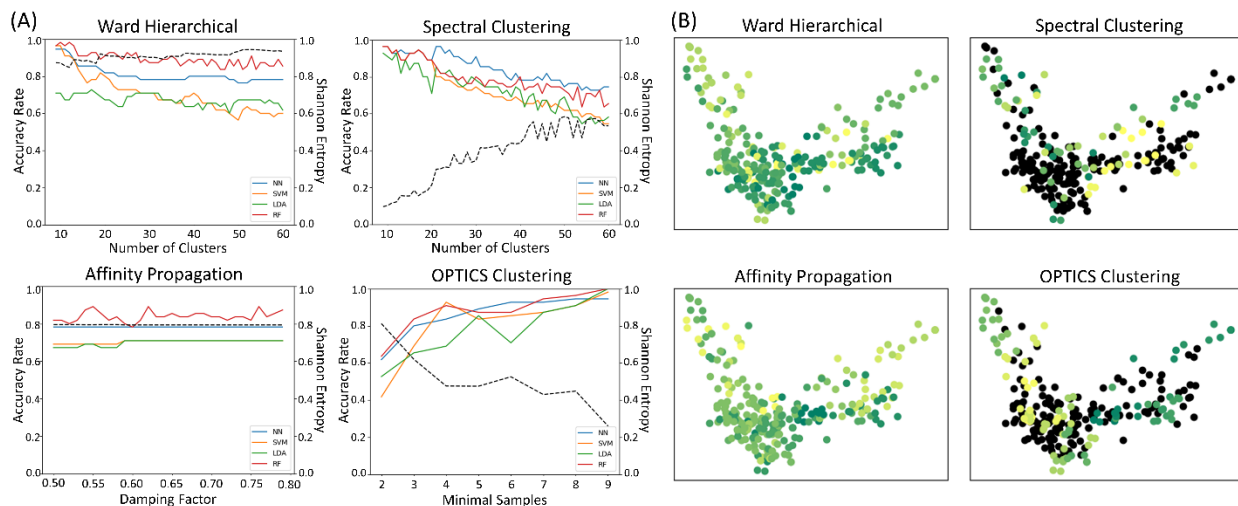


Figure 5. Comparison of phytochemical clustering and classification models. (A) Model hyperparameter tuning combined with classification methods. The left side of the y-axis indicates accuracy (colored solid lines) and the right side of the y-axis indicates Shannon entropy (dashed line --) (B) 2-Dimensional representations of clustered molecules using PCA. The color shows the distribution of molecules into different clusters. Black data points represent molecules that were treated as noise because the clustering algorithm is unable to group them based on similarities.

Identification of Lead Phytochemicals and Lead Clusters.

We identified 34 lead phytochemicals and 8 lead clusters by combining clustering and SBVS results. Because different SARS-CoV-2 protein structures generated different energy score distributions, all energy scores were standardized by using z-scores to compare the binding ability of phytochemicals across different structures. The z-scores indicate the number of standard deviations from the sample means, which, in this study, are the averages of all lowest energy scores for the dockings of the initial 272 anti-viral phytochemicals (in SBVS) with specific protein structures. In the heatmap of z-scores (Figure 6A left), each column represents a different protein structure and, therefore, has a different mean and standard deviation. The dark blue and purple cells indicate significantly greater-than-average binding affinities of phytochemicals to particular protein targets (two or more standard deviations below the mean energy score). The yellow and green cells indicate binding affinities that are only slightly greater than the average, and the white cells indicate binding affinities that are weaker than the average. Using a z-score of -2 as the threshold to identify lead candidates, we identified 34 lead compounds from the 272 anti-viral phytochemicals. (Table S4) Among them, there were several with strong specificity toward a single protein structure. For example, (-)-Epicatechin-3-o-gallate shows a strong binding ability to NSP13 (6ZSL), Gambiriin-b3 and Procyanidin-a-2 show strong binding ability to NSP10 (6ZCT), and Procyanidin B2 shows a strong binding ability to NSP5 (6Y2E). There were also certain phytochemicals that demonstrated a high binding affinity to multiple SARS-CoV-2 viral proteins, i.e., a polypharmacological/multi-target behavior. For example, Agathisflavone demonstrates a high binding affinity to NSP13 (7NIO) and NSP15

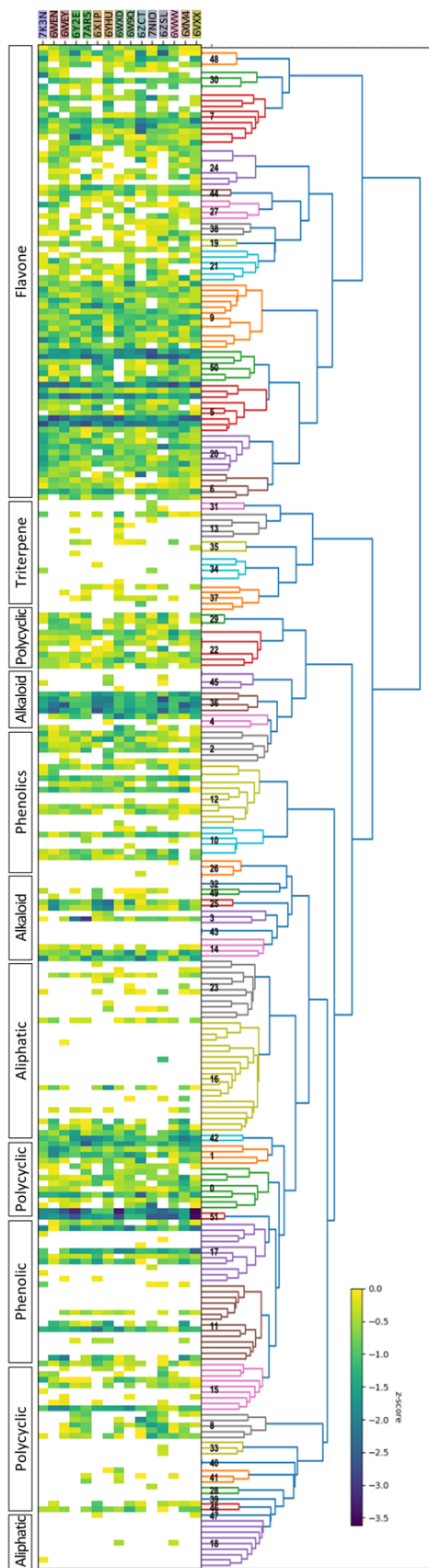
(6VWW) and Hypericin demonstrates strong binding to NSP5 (6Y2E), NSP9 (6WXD), and the Spike protein (6VXX). There is a risk, however, that this multi-target behavior may indicate molecular promiscuity. We used the PAINS detector in SwissADME on all of our leads to check for molecular promiscuity⁵¹; however, in vivo and in vitro work is needed for better analysis on this front.

The dendrogram graph shows the hierarchical orders of formed clusters (Figure 6A right). Closely related clusters 5 and 50 have a large dark area in the heatmap. Other noticeable patches of dark areas were observed for clusters such as 36 and 51, which indicate that many of their constituent phytochemicals bind strongly to more than one SARS-CoV-2 protein structure. The number of lead phytochemicals within each cluster was counted for each protein structure (Figure 7) in order to link cluster specificity to different SARS-CoV-2 structures. We identified the following clusters as lead clusters for our viral proteins:

1. Cluster 5 – NSP1, NSP3, NSP5, NSP7&8, NSP9, NSP13, NSP15, Spike receptor binding domain (RBD), and the Spike protein
2. Cluster 7 – NSP10
3. Cluster 30 – NSP1
4. Cluster 36 – NSP3, NSP7&8, NSP13, and the Spike RBD
5. Cluster 42 – NSP5 and the Spike RBD
6. Cluster 49 – NSP7&8
7. Cluster 50 – NSP1, NSP7&8, NSP13, NSP15, and the Spike RBD
8. Cluster 51 – NSP3, NSP5, NSP9, NSP13, NSP15, and the Spike Protein

The simulated energy z-scores of 16 experimentally validated SARS-CoV-2 main protease inhibitors were used as benchmarks.⁵¹ Sixty-nine percent of these inhibitors were below the mean when docked against the main protease structure 6Y2E, and 63% were below the mean when docked against the main protease structure 7AR5 (Figure S2). We therefore hypothesized that our lead identification threshold (two or more standard deviations below the mean energy score) was fairly rigorous at identifying good binders.

(A)



(B)

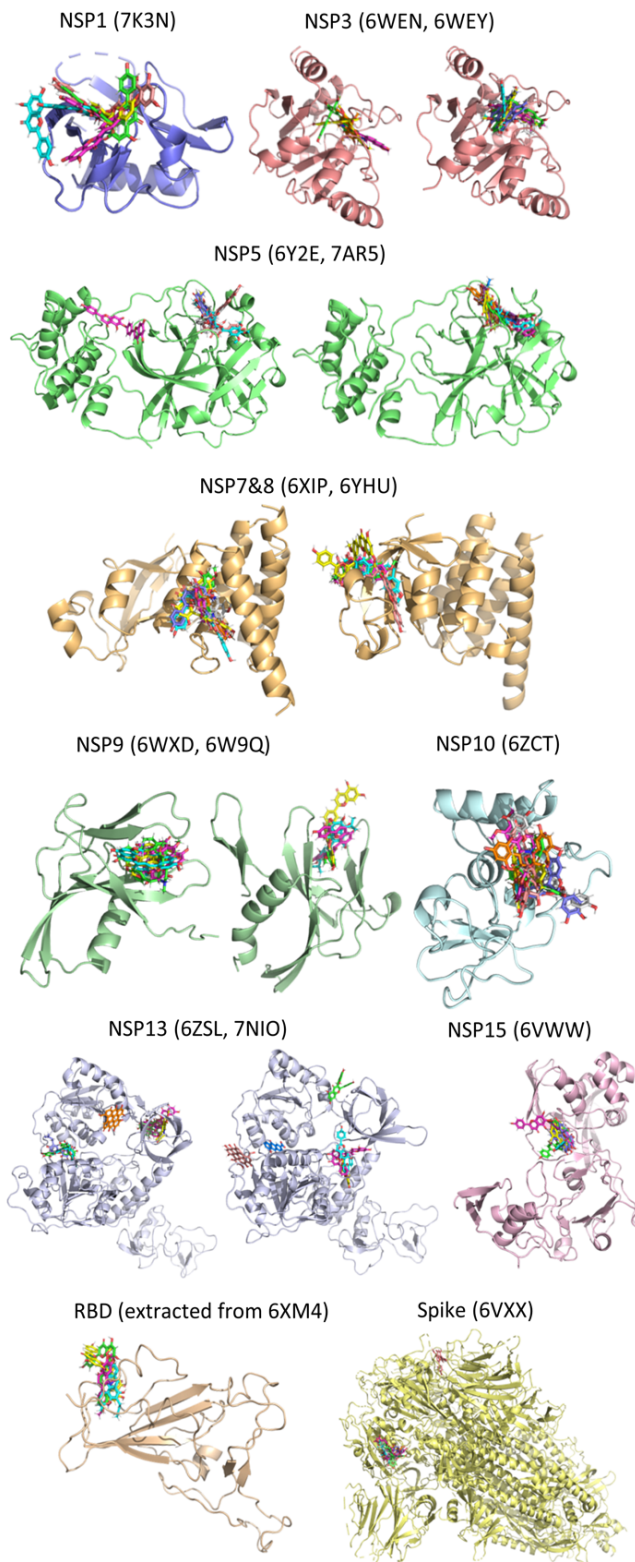


Figure 6. (A) Heatmap of docking energy z-scores of 272 anti-viral phytochemicals initially used in SBVS (left) and the cluster dendrogram with cluster ID labels (right). The phytochemicals are grouped into their clusters, and their names and numerical IDs are given in Table S9 of the Supplemental Information. Phytochemicals are also grouped into approximate chemical categories on the left side of the heatmap. (B) Docked structures for lead candidates (PDBs are available in Supporting Materials).

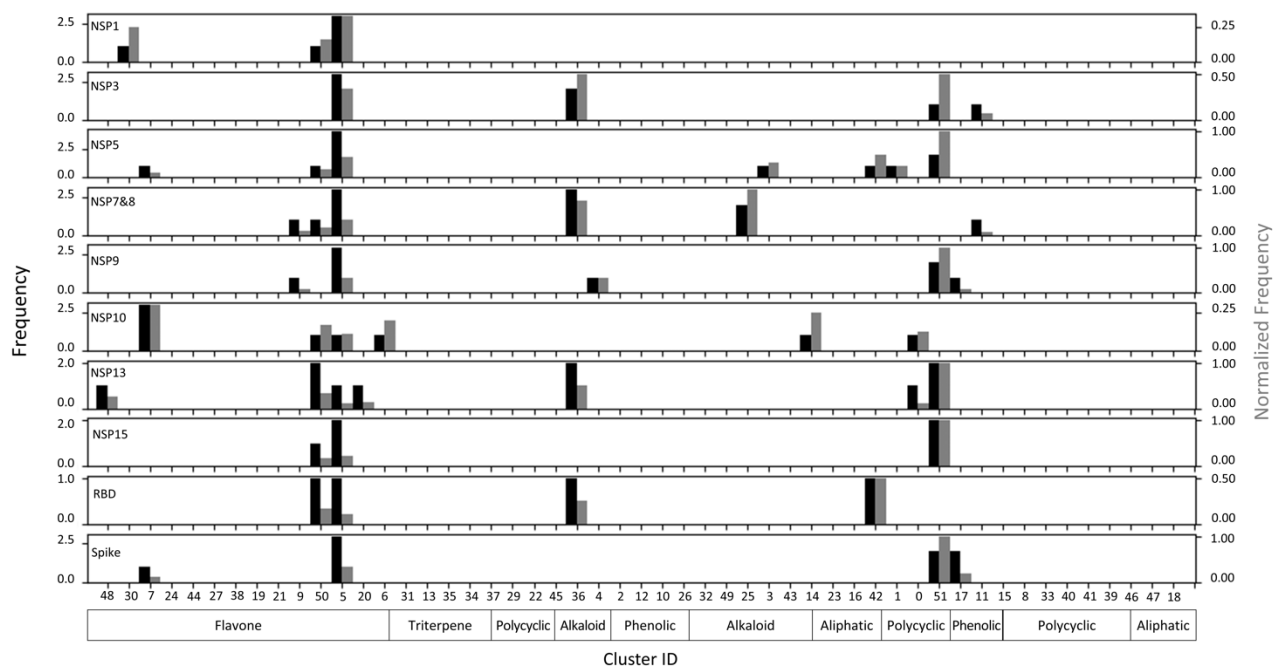


Figure 7. Frequency (black bars) and normalized frequency (gray bars) of identified leads within each molecule cluster. Molecule cluster IDs are given on the x axis, and the approximate chemical classes which most phytochemicals within a cluster belong to are also specified on the x axis.

Evaluation of LBVS Model. The inclusion of our ligand based virtual screen (LBVS) increased the rate of lead identification from 2.18% (SBVS only) to 16.44% (SBVS + LBVS). The 1000 new phytochemicals were classified into 52 formed clusters, and 53 of them were classified into lead clusters. Based on the specificity of clusters, we ran a total 298 docking simulations between these 53 predicted lead phytochemicals and their corresponding protein structures. Among z-scores of 298 dockings, 49 cases (16.5%) were below -2, 214 cases (72.05%) were between -2 and 0, and 34 cases (11.45%) were above 0 (Table S6). Compared to the sample z-scores of the initial dockings of 272 anti-viral phytochemicals, we introduced a negative distributional shift of z-scores (Figure 8A). To further validate the improved predictive power afforded by the ligand based approach, we docked 298 randomly selected phytochemicals that had been classified into non-lead clusters (Figure 8B). A z-test analysis was performed on sample z-scores of the two populations (phytochemicals in lead clusters and those in non-lead clusters). The p-value of 9.41×10^{-24} indicated that the mean difference of these two samples is statistically significant, suggesting molecule clustering and classification methods improved lead

and non-lead class separation by using the extracted chemical features of strong binders to identify others. The additional phytochemicals that we predicted as lead compounds and confirmed by their docking energy scores are available in Table S5.

A random under-sampling confusion matrix was constructed to measure the performance of our classification (prediction) model (Table S7). The matrix was based on protein-ligand pair counting. The recall (true positive rate) of 0.73 and 0.68 were obtained when the energy z-score of -2 and -1 were used to determine actual positive and negative, respectively. This suggested that our model retrieved relevant lead phytochemicals. However, the F1 score of 0.27 and 0.41 suggested that our model could be further improved.

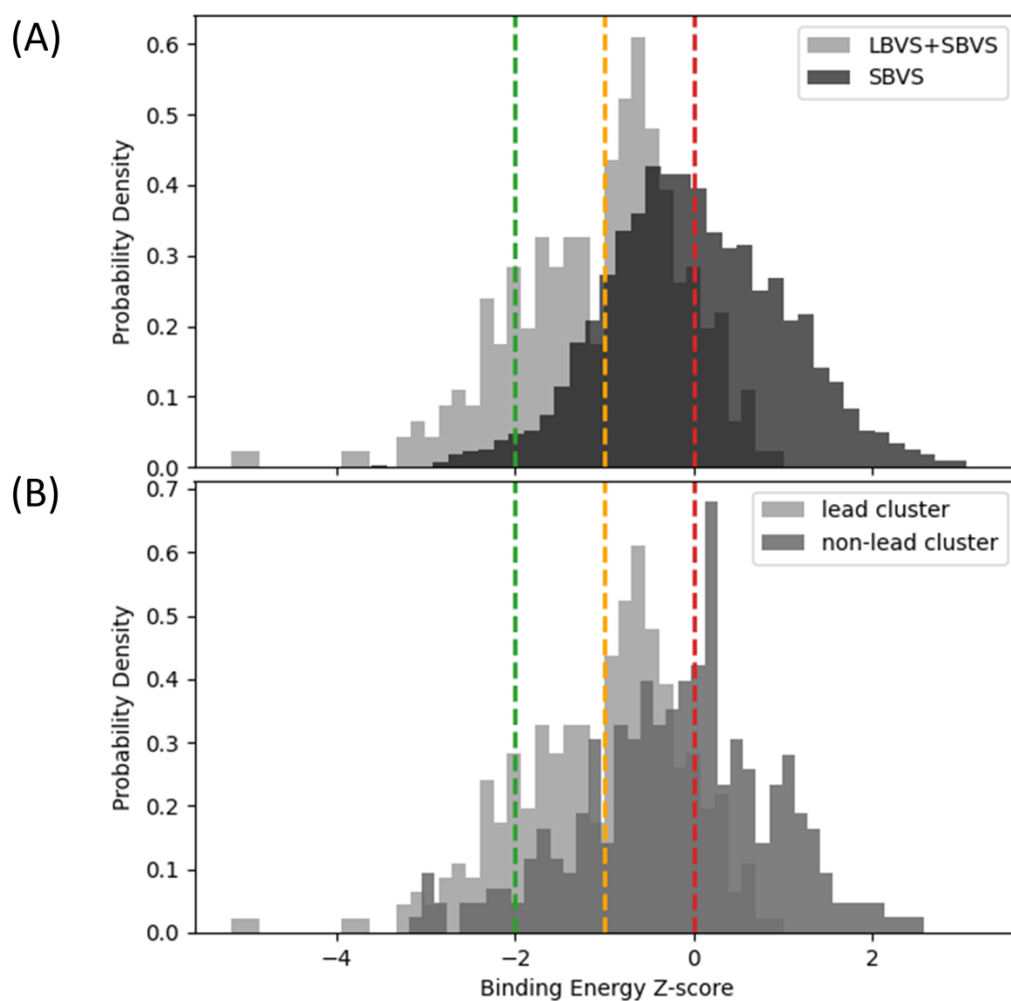


Figure 8. (A) Distribution of docking energy z-scores generated via SBVS alone (dark gray) and SBVS with the inclusion of LBVS (light gray). (B) Distribution of docking energy z-scores of phytochemicals classified to lead clusters (light gray) and those classified to non-lead clusters (dark gray) via LBVS. The red, yellow, and green dashed lines label z scores of 0, -1, and -2 respectively.

Drug-likeness Screening for the 62 Identified Lead Phytochemicals. We used SwissADME to obtain certain drug property parameters for lead phytochemicals identified through both the initial SBVS and those identified through LBVS and SBVS combined (Figure 9A).⁵² Eighteen compounds (Table 1, Table S10) showed promising results with a maximum of 1 violation in all screened categories (drug-likeness, PAINS, Brenk, and lead-likeness). This threshold was based on the fact that Doravirine, a drug approved by the FDA in 2018 for the treatment of HIV, had 1 total violation (Table S8).⁵³ The drug-likeness violations category is based on the following 5 rules: Lipinski, Ghose, Veber, Egan, and Muegge. PAINS alerts detect potentially promiscuous binders and Brenk alerts identify potentially toxic and metabolically unstable molecular moieties. Lead-likeness refers to similarities a given compound has to a “lead”, or a starting point for further drug development.

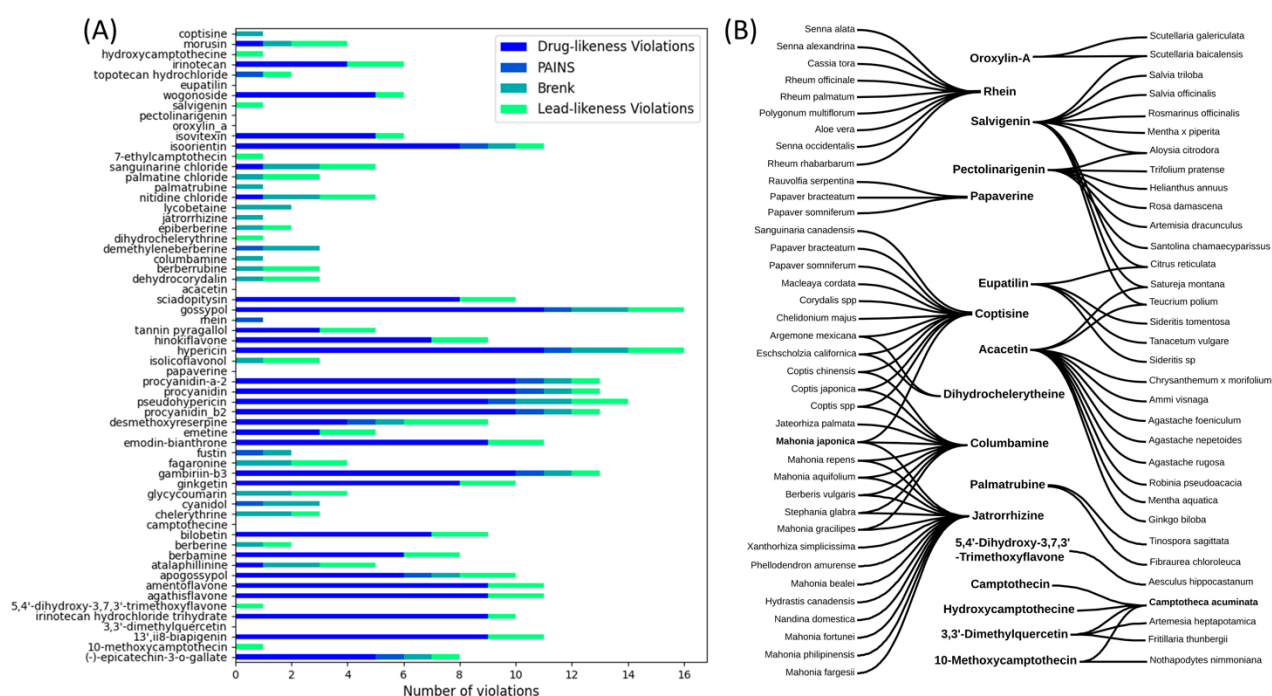


Figure 9. (A) The cumulative violations of each lead molecule in the drug-likeness, PAINS, lead-likeness, and Brenk categories. (B) Plant sources for 17 promising phytochemicals identified through the drug-likeness screening (No plant sources could be found for 7-ethylcamptothecin). Plant names are on the two sides and phytochemicals are in the middle. *Mahonia japonica* and *Camptotheca acuminata* are bolded and contain at least 3 of the promising phytochemicals.

Table 1: Lead phytochemical compounds from LBVS and SBVS with favorable drug-likeness properties targeting structural and non-structural SARS-CoV-2 proteins. Bolded compounds were identified using LBVS rather than SBVS alone. The cytochrome interaction field identifies the number of main P450 cytochrome isoforms (out of 5) that a compound interacts with.

Target	Lead Phytochemicals	Cluster	Category	Cytochrome Interaction	Solubility (mmol/L)
NSP7&8	Columbamine	36	Alkaloid	3	0.04
	Dihydrochelerythrine			5	0.01
	Jatrorrhizine			3	0.04
	Palmatrubine			3	0.04
	Papaverine			5	0.12
	10-methoxycamptothecin	49		4	0.29
	7-ethylcamptothecin			5	0.07
	Camptothecin			3	0.40
	Hydroxycamptothecin			1	0.41
	Acacetin	5		Flavone	4
	5,4'-dihydroxy-3,7,3'-trimethoxyflavone	9	4		0.02
	Eupatilin	50	4		0.02
	Oroxilin A		4		0.03
Pectolinarigenin	4		0.03		
Salvigenin	5		0.02		
NSP9	3,3'-dimethylquercetin		9		Flavone
NSP13	Columbamine	36	Alkaloid	3	0.04
	Coptisine			2	0.04
	Dihydrochelerythrine	5	0.01		
	Rhein	0	Polycyclic	0	0.28
Spike RBD	Columbamine	36	Alkaloid	3	0.04

Some promising phytochemicals like dihydrochelerythrine (alkaloid with antimicrobial and anticancer properties^{54 55}) were poorly soluble compared to the others despite having either 0 or 1 total violations in all categories. Many leads were identified as potential inhibitors of some or all of the five main cytochrome P450 isoforms: CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4. Inhibition of these cytochromes can potentially lead to undesirable drug-drug interactions.⁵² Rhein, however, was not identified as an inhibitor of any of those isoforms. In addition, rhein and camptothecin were compared with 3 COVID-19 anti-viral medicines (Remdesivir, Molnupiravir, and Paxlovid) that are either FDA-authorized or awaiting approval, and they were also compared with Doravirine (Table S8). The comparison indicates that rhein is more soluble, has a higher bioavailability score, has better GI absorption, and has fewer drug likeness violations than Remdesivir (which has 11 drug-likeness violations). Rhein also has greater solubilities than Paxlovid. However, rhein has 1 PAINS alert, which may indicate an undesirable promiscuity. Overall, rhein and camptothecin have few violations and are classified as either soluble or moderately soluble in all categories (like Paxlovid and Doravirine) (Table S8). Our assessment is in agreement with recent reports of the therapeutic potential of rhein^{56 57} and camptothecin.^{58 59}

Lastly, we built a phytochemical-plant network for 17 leads, in order to discover plants that contain more than one lead (Figure 9B) using data from *Dr. Duke's USDA Phytochemical and Ethnobotanical Databases*.¹² The network shows that the plant *Camptotheca acuminata* contains 4 leads, the plant *Mahonia japonica* contains 3 leads, and the rest of the plants have one or two connections to lead phytochemicals.

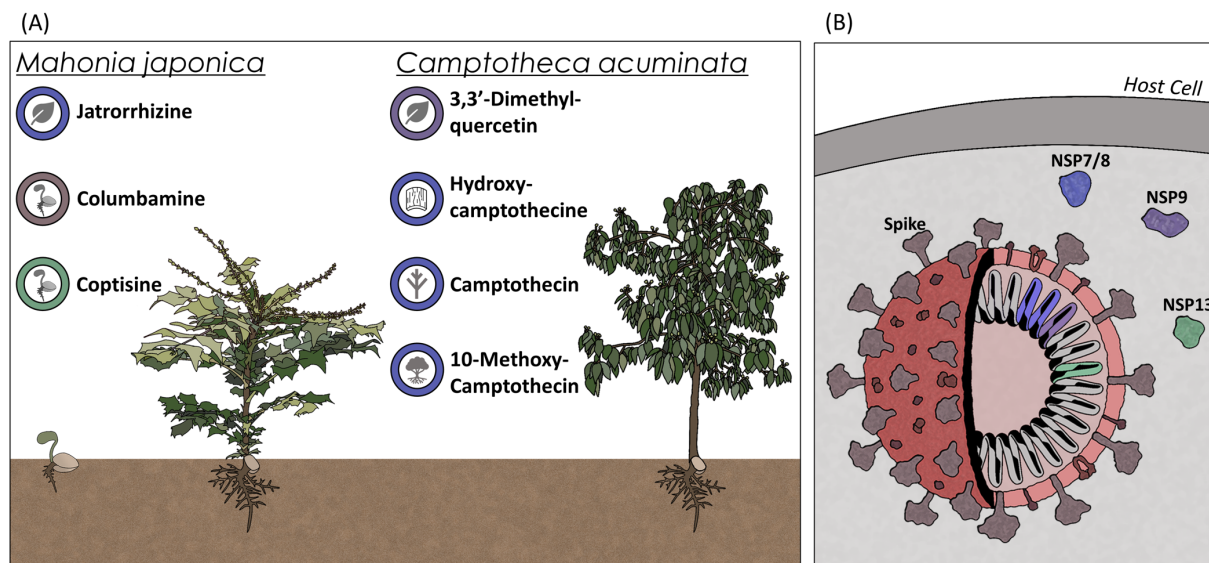


Figure 10. (A) The two plants that contain 3 (*Mahonia japonica*) and 4 (*Camptotheca acuminata*) of the 17 phytochemicals identified first as leads through SBVS or SBVS and LBVS combined, and then shown to have promising pharmacokinetic profiles in the SwissADME screening. The part of the plant most abundant in a specific phytochemical (leaf, sprout sapling, bark, stem, whole plant) are shown in the icons to the left of the compound names.^{60 61 62 63} (B) A SARS-CoV-2 virion and the four labeled viral proteins targeted by the compounds in panel A. Color coded circles in A correspond to the protein targeted by each compound.

CONCLUSIONS

In this project, we produce new evidence in support of a polypharmacological approach for treating SARS-CoV-2 using naturally abundant phytochemicals. We implemented a Rosetta high-resolution protein-ligand docking protocol (SBVS) in combination with ligand clustering via machine learning strategies (LBVS) to identify combinations of promising phytochemical binders against several SARS-CoV-2 proteins (structural and non-structural). The initial structure-based virtual screen identified 34 leads from a library of 272 anti-viral phytochemicals using molecular docking. Ward hierarchical clustering of ligands from the initial screen revealed flavone and alkaloid chemical features to be most predictive of lead compounds. These results informed our ligand-based virtual screen, giving rise to 28 newly identified lead compounds and a 4-fold increase in rate of lead discovery. Applying physicochemical filters on our panel of 62 phytochemical leads, we refined the number of therapeutically promising compounds to 18. Of those, rhein and camptothecin with strong potential binding affinities to NSP13 (7NIO) and NSP7&8 (6YHU), respectively, stood out by showing drug-likeness properties superior to those of Remdesivir, and comparable in many aspects to those of Paxlovid, Doravirine and Molnupiravir.

The purpose of this project is to shine light on potential phytochemicals that could be used in a polypharmacological manner for COVID-19 prevention and treatment. Our analyses are based on high-quality simulation data, statistical inferences, and machine learning predictions. While recent experimental^{64 65} and computational^{66 67} findings corroborate the therapeutic potential of the lead compounds identified here in our work, future *in vivo* and *in vitro* studies are needed to validate ligand function and efficacy. We hope our results and workflow will help to improve the scope of drug discovery efforts and reduce the high failure rate prior to costly lab testing.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available for free on the ACS Publications website at [\(link\)](#). The supplemental files contain the following figures, explanations, and tables in the written order: Histogram showing the number of instances when a protein-ligand complex from CASF-2016 had its ligand bound in a certain CASTp-identified pocket. (Figure S1 top); CASTp-identified binding pockets that were sampled in the docking of each of our protein structures (Figure S1 bottom); explanation of how the CASF-2016 data was used to establish our pocket sampling criteria; data table used to establish pocket volume cutoff criteria (Table S1); data table used to establish pocket surface area cutoff criteria (Table S2); crystal structures used in the score function testing (Table S3); protein targets, energy scores, and clusters of lead candidates identified with SBVS (Table S4); protein targets, energy scores, and clusters of lead candidates identified with LBVS (Table S5); Docking results of 16 known main-protease inhibitors (Figure S2); percentage of docking energy scores within certain ranges relative to the mean after LBVS

was applied (Table S6); random under-sampling confusion matrix (Table S7); drug-likeness data comparison between promising leads and the antivirals Doravirine, Paxlovid, Molnupiravir, and Remdesivir (Table S8); docking script used in Rosetta (Figure S3); phytochemicals used in SBVS named, labeled numerically, and organized into their clusters (Table S9); SwissADME results of the 18 compounds identified as promising in the drug-likeness screening (Table S10).

ACKNOWLEDGEMENTS

We thank Andrew Bruno of the University at Buffalo for the molecular fingerprint algorithm we used (The ECFPs algorithm is available at <https://github.com/ubccr/pinky>), and Dr. Yong-hui Zheng (Department of Microbiology and Molecular Genetics at MSU) for providing the LBVS compound list. We thank the Institute for Cyber Enabled Research (ICER) at Michigan State University for technical help and computational resources. We also sincerely thank Professor Mark Reimers (Department of Biomedical Engineering at MSU) for his critical role in helping to establish the scope of this study and insightful conversations throughout the project.

AUTHOR INFORMATION

Corresponding authors:

Daniel Woldring: woldring@msu.edu

NOTES

Data and Software Availability

For the molecular docking, Rosetta 3.12 was used which can be obtained for free with an academic license (<https://www.rosettacommons.org/software/license-and-download>). Rosetta 3.12 was installed onto a cluster maintained by the Michigan State University Institute for Cyber Enabled Research. Docking jobs on this cluster were submitted using the Slurm workload manager. The CIDs, names, and SMILES of the 272 phytochemicals initially used in SBVS are available in the supporting files in a spreadsheet titled “Ligand_Library_Key_SBVS”. In that spreadsheet, the numerical IDs present on the left side of Figure 6A are connected to the phytochemicals that they represent. The SMILES and names of all the additional compounds screened through LBVS are available in a spreadsheet titled “AdditionalLibraryForLBVS.” The complete SwissADME data for the 62 lead compounds is available in the spreadsheet titled “SwissADMEfinalresults.” The BCL:Conf ligand conformer generator was installed alongside Rosetta 3.12 on the cluster, and it was obtained for free with an academic license from http://www.meilerlab.org/index.php/bclcommons/show/b_apps_id/1. OpenBabel was obtained for free from <http://openbabel.org/wiki/Category:Installation>. Various python scripts were used to generate plots, process docking input files and generate docking jobs on the cluster, and they are all available at (https://github.com/ziruiwang1996/ligand_protein_docking). Other files containing raw docking data, components for the LBVS algorithm, and PDB files of all the lead compounds docked against specific proteins are accessible via a link present in a README.md document located at the GitHub site linked previously. These other folders and files are all inside

a Google Drive folder titled “data,” which is accessed by clicking the link in the README file. Additional score function testing data is available at https://ziruiw.shinyapps.io/score_functions_on_sarscov2/.

REFERENCES

1. Johns Hopkins Coronavirus Resource Center, C. for S. S. and E. (CSSE) at J. H. U. (JHU). COVID-19 Dashboard.
2. Wanga, V. *et al.* Long-Term Symptoms Among Adults Tested for SARS-CoV-2 — United States, January 2020–April 2021. https://www.cdc.gov/mmwr/mmwr_continuingEducation.html (2021).
3. Nations Environment Programme, U. & Livestock Research Institute, I. *A Scientific Assessment with Key Messages for Policy-Makers A Special Volume of UNEP’s Frontiers Report Series PREVENTING THE NEXT PANDEMIC PREVENTING THE NEXT PANDEMIC Zoonotic diseases and how to break the chain of transmission.* (2020).
4. Noh, J. Y., Jeong, H. W. & Shin, E. C. SARS-CoV-2 mutations, vaccines, and immunity: implication of variants of concern. *Signal Transduction and Targeted Therapy* vol. 6 (2021).
5. Pouwels, K. B. *et al.* Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *Nature Medicine* (2021) doi:10.1038/s41591-021-01548-7.
6. Liu, L. *et al.* Striking Antibody Evasion Manifested by the Omicron Variant of SARS-CoV-2. *Nature* (2021) doi:10.1038/s41586-021-04388-0.
7. Burki, T. K. The role of antiviral treatment in the COVID-19 pandemic. *The Lancet Respiratory Medicine* (2022) doi:10.1016/S2213-2600(22)00011-X.
8. Johnson, V. A. *Combination Therapy: More Effective Control of HIV Type 1? AIDS RESEARCH AND HUMAN RETROVIRUSES* vol. 10 (1994).
9. Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nature Chemical Biology* vol. 4 682–690 (2008).
10. Patil, R. *et al.* Computational and network pharmacology analysis of bioflavonoids as possible natural antiviral compounds in COVID-19. *Informatics in Medicine Unlocked* **22**, (2021).
11. Muhammad, J. *et al.* Network Pharmacology: Exploring the Resources and Methodologies. *Current Topics in Medicinal Chemistry* **18**, 949–964 (2018).
12. Dr. Duke’s Phytochemical and Ethnobotanical Databases. *U.S. Department of Agriculture, Agricultural Research Service.*
13. H.M. Berman *et al.* The Protein Data Bank. *Nucleic Acids Research* (2000).

14. Zev, S. *et al.* Benchmarking the Ability of Common Docking Programs to Correctly Reproduce and Score Binding Modes in SARS-CoV-2 Protease Mpro. *Journal of Chemical Information and Modeling* **61**, 2957–2966 (2021).
15. Macip, G. *et al.* Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. *Medicinal Research Reviews* (2021) doi:10.1002/med.21862.
16. Cherrak, S. A., Merzouk, H. & Mokhtari-Soulimane, N. Potential bioactive glycosylated flavonoids as SARS-CoV-2 main protease inhibitors: A molecular docking and simulation studies. *PLoS ONE* **15**, (2020).
17. Stoddard, S. v. *et al.* Optimization rules for SARS-CoV-2 Mpro antivirals: Ensemble docking and exploration of the coronavirus protease active site. *Viruses* **12**, (2020).
18. Ruan, Z. *et al.* SARS-CoV-2 and SARS-CoV: Virtual screening of potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12). *Journal of Medical Virology* **93**, 389–400 (2021).
19. Basu, A., Sarkar, A. & Maulik, U. Molecular docking study of potential phytochemicals and their effects on the complex of SARS-CoV2 spike protein and human ACE2. *Scientific Reports* **10**, (2020).
20. Chandel, V. *et al.* Structure-based drug repurposing for targeting Nsp9 replicase and spike proteins of severe acute respiratory syndrome coronavirus 2. *Journal of Biomolecular Structure and Dynamics* **40**, 249–262 (2022).
21. Ngwa, W. *et al.* Potential of flavonoid-inspired phytomedicines against COVID-19. *Molecules* **25**, (2020).
22. Wu, Y. *et al.* Polyphenols as Potential Inhibitors of SARS-CoV-2 RNA Dependent RNA Polymerase (RdRp). *Molecules* **26**, 7438 (2021).
23. Ghosh, R., Chakraborty, A., Biswas, A. & Chowdhuri, S. Identification of alkaloids from *Justicia adhatoda* as potent SARS CoV-2 main protease inhibitors: An in silico perspective. *Journal of Molecular Structure* **1229**, (2021).
24. Gawriljuk, V. O. *et al.* Machine Learning Models Identify Inhibitors of SARS-CoV-2. *Journal of Chemical Information and Modeling* vol. 61 4224–4235 (2021).
25. Nguyen, D. D., Gao, K., Chen, J., Wang, R. & Wei, G. W. Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning. *Chemical Science* **11**, 12036–12046 (2020).
26. Kumari, M. & Subbarao, N. Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases. *Computers in Biology and Medicine* **132**, (2021).
27. Wang, S., Sun, Q., Xu, Y., Pei, J. & Lai, L. A transferable deep learning approach to fast screen potential antiviral drugs against SARS-CoV-2. *Briefings in Bioinformatics* **22**, (2021).

28. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* vol. 18 463–477 (2019).
29. Yadav, R. *et al.* cells Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19. **10**, 821 (2021).
30. Newman, J. A. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature Communications* **12**, (2021).
31. O’Boyle, N. M. *et al.* Open Babel: An Open chemical toolbox. *Journal of Cheminformatics* **3**, (2011).
32. Kothiwale, S., Mendenhall, J. L. & Meiler, J. BCL::Conf: Small molecule conformational sampling using a knowledge based rotamer library. *Journal of Cheminformatics* **7**, (2015).
33. Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B Structural Science* **58**, 380–388 (2002).
34. Sam DeLuca. How to Prepare Ligands for use in Rosetta. *Rosettacommons.Org*.
35. Tian, W., Chen, C., Lei, X., Zhao, J. & Liang, J. CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Research* **46**, W363–W367 (2018).
36. Su, M. *et al.* Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **59**, 895–913 (2019).
37. Lemmon, G. & Meiler, J. Rosetta ligand docking with flexible XML protocols. *Methods in Molecular Biology* **819**, 143–155 (2012).
38. Pavlovicz, R. E., Park, H. & DiMaio, F. Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking discrimination. *PLoS Computational Biology* **16**, (2020).
39. Smith, S. T. & Meiler, J. Assessing multiple score functions in Rosetta for drug discovery. *PLoS ONE* **15**, (2020).
40. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010).
41. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825--2830 (2011).
42. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 86–97 (2012).
43. von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing* **17**, 395–416 (2007).
44. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).

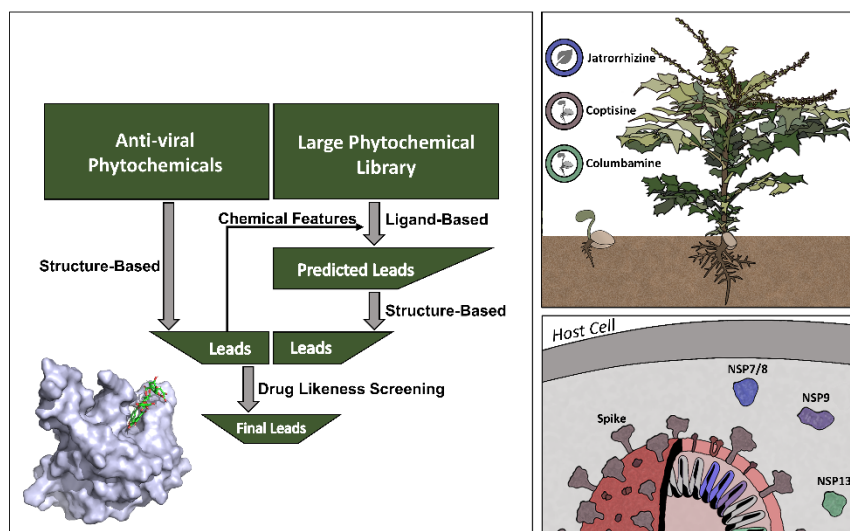
45. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. OPTICS. in 49–60 (Association for Computing Machinery (ACM), 1999). doi:10.1145/304182.304187.
46. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, (2015).
47. Peterson, L. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).
48. Noble, W. S. *What is a support vector machine?* *NATURE BIOTECHNOLOGY* vol. 24 <http://www.nature.com/naturebiotechnology> (2006).
49. Svetnik, V. *et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43**, 1947–1958 (2003).
50. Izenman, A. J. Linear Discriminant Analysis. in *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning* 237–280 (Springer New York, 2013). doi:10.1007/978-0-387-78189-1_8.
51. Li, Z. *et al.* Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs. (2020) doi:10.1073/pnas.2010470117/-/DCSupplemental.
52. Daina, A., Michielin, O. & Zoete, V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports* **7**, (2017).
53. Deeks, E. D. Doravirine: First Global Approval. *Drugs* **78**, 1643–1650 (2018).
54. Casciaro, B. *et al.* Naturally-Occurring Alkaloids of Plant Origin as Potential Antimicrobials against Antibiotic-Resistant Infections. *Molecules* **25**, 3619 (2020).
55. Okagu, I. U., Ndefo, J. C., Aham, E. C. & Udenigwe, Chibuiké. C. Zanthoxylum Species: A Review of Traditional Uses, Phytochemistry and Pharmacology in Relation to Cancer, Infectious Diseases and Sick Cell Anemia. *Frontiers in Pharmacology* **12**, (2021).
56. Zannella, C. *et al.* Regulation of m6A Methylation as a New Therapeutic Option against COVID-19. *Pharmaceuticals* **14**, 1135 (2021).
57. Narkhede, R. R., Pise, A. v., Cheke, R. S. & Shinde, S. D. Recognition of Natural Products as Potential Inhibitors of COVID-19 Main Protease (Mpro): In-Silico Evidences. *Natural Products and Bioprospecting* **10**, 297–306 (2020).
58. Pushparaj, P. N., Abdulkareem, A. A. & Naseer, M. I. Identification of Novel Gene Signatures using Next-Generation Sequencing Data from COVID-19 Infection Models: Focus on Neuro-COVID and Potential Therapeutics. *Frontiers in Pharmacology* **12**, (2021).
59. Mamkulathil Devasia, R., Altaf, M., Fahad Alrefaei, A. & Manoharadas, S. Enhanced production of camptothecin by immobilized callus of *Ophiorrhiza mungos* and a bioinformatic insight into its potential antiviral effect against SARS-CoV-2. *Journal of King Saud University - Science* **33**, 101344 (2021).

60. Ikuta, A. & Itokawa, H. Studies on the Alkaloids from Tissue Culture of *Nandina domestica*. *Plant tissue culture 1982: proceedings, 5th International Congress of Plant Tissue and Cell Culture held at Tokyo and Lake Yamanake, Japan (1982)*. (1982).
61. Willaman, J. J. & Li, H. L. Alkaloid-Bearing Plants and Their Contained Alkaloids, 1957–1968, Supplement of *Lloydia-The Journal of Natural Products*. *Journal of Pharmaceutical Sciences* **60**, 958 (1971).
62. Duke, J. A. Handbook of Phytochemical Constituents of GRAS Herbs and Other Economic Plants. in *Handbook of Phytochemical Constituents of GRAS Herbs and Other Economic Plants* 1–654 (Routledge, 2017). doi:10.1201/9780203752623-1.
63. Zhang, S. *et al.* Cyclane-aminol 10-hydroxycamptothecin analogs as novel DNA topoisomerase I inhibitors induce apoptosis selectively in tumor cells. *Anti-Cancer Drugs* **25**, 614–623 (2014).
64. Al-Karmalawy, A. A. *et al.* Naturally Available Flavonoid Aglycones as Potential Antiviral Drug Candidates against SARS-CoV-2. *Molecules* **26**, 6559 (2021).
65. Ye, M. *et al.* Network pharmacology, molecular docking integrated surface plasmon resonance technology reveals the mechanism of Toujie Quwen Granules against coronavirus disease 2019 pneumonia. *Phytomedicine : international journal of phytotherapy and phytopharmacology* **85**, 153401 (2021).
66. Wang, Z., Zhang, J., Zhan, J. & Gao, H. Screening out anti-inflammatory or anti-viral targets in Xuanfei Baidu Tang through a new technique of reverse finding target. *Bioorganic Chemistry* **116**, 105274 (2021).
67. M, P., Reddy, G. J., Hema, K., Dodoala, S. & Koganti, B. Unravelling high-affinity binding compounds towards transmembrane protease serine 2 enzyme in treating SARS-CoV-2 infection using molecular modelling and docking studies. *European Journal of Pharmacology* **890**, 173688 (2021).

FOR TABLE OF CONTENTS USE ONLY

Phytochemical Drug Discovery for COVID-19 Using High-resolution Computational Docking and Machine Learning Assisted Binder Prediction

Zirui Wang, Theodore Belecciu, Joelle Eaves, Michael H. Bachmann, Daniel Woldring*



Supporting Information

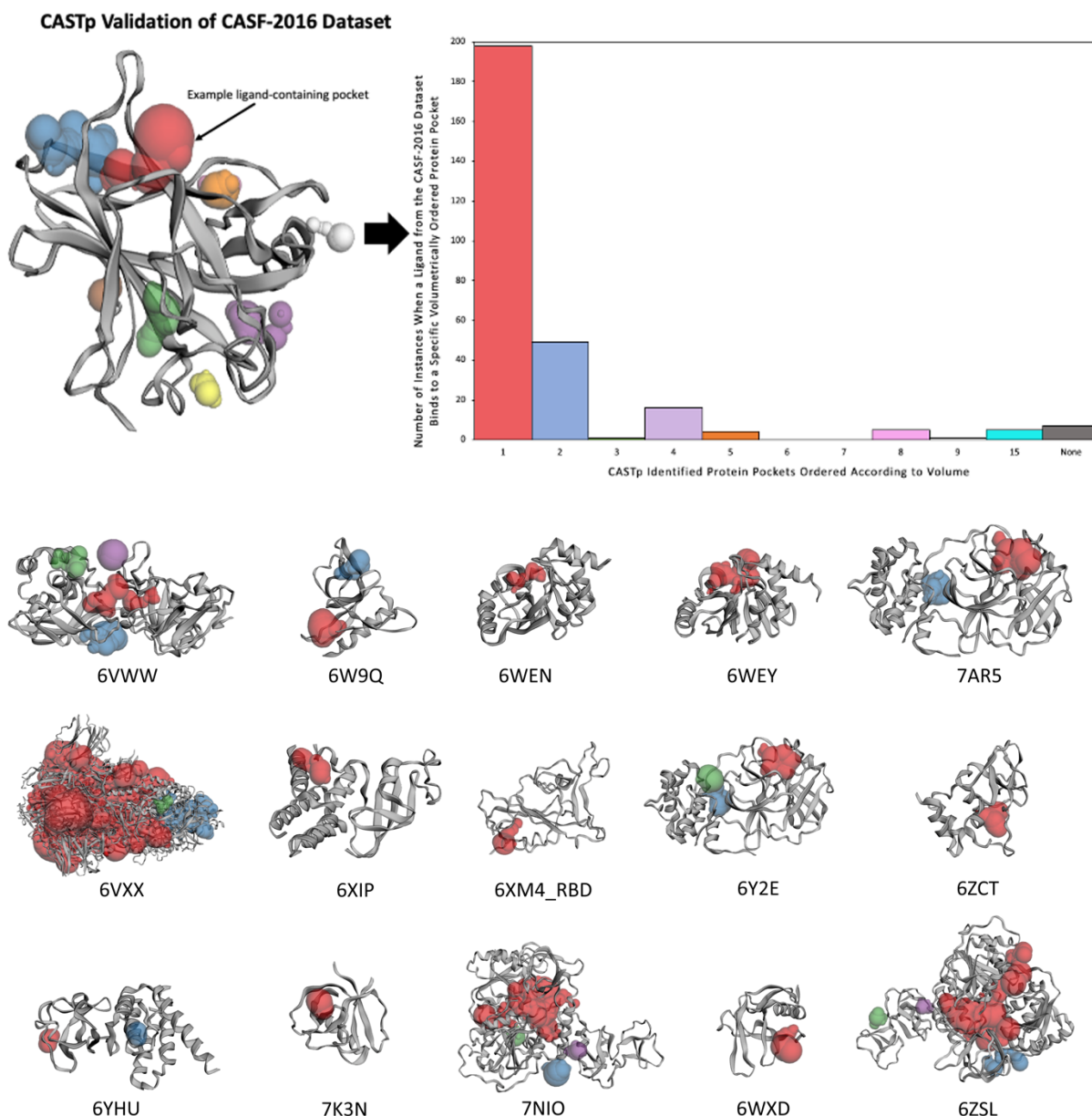


Figure S1. Global Docking Binding Pocket Validation. (top left) Using CASTp, a list of protein pockets were generated and ranked by volume for each structure in the CASF-2016 dataset, which consists of 285 different ligand bound crystal structures (57 different proteins and 285 different main ligands) (Example PDB shown: 4AGQ). (top right) The number of instances where the main ligands from the CASF-2016 dataset bind to a specific volumetrically ordered pocket are tabulated in a histogram. The CASTp-identified binding pockets for the CASF-2016 structures are color-coded according to the following volumetric ordering: red (largest), blue, green, purple, orange, yellow, brown, pink, white (9th largest). Five protein structures had their main ligand located in the 15th largest CASTp-identified binding pockets ranked volumetrically (colored in turquoise). Seven of the structures (2% of the dataset) didn't have their main ligand located in any CASTp-identified pocket (colored in gray). Overall, CASTp is quite successful at predicting the possible ligand binding pockets for the CASF-2016 dataset, since only 7 out of the 285 complexes didn't have their binding pockets identified by it. As shown, the majority of CASF-2016

complexes had their ligand located either in their first or second largest pocket by volume. (Bottom) The pockets identified by CASTp that were used for docking in this study are indicated for all 15 structures and their respective PDB IDs.

Establishment of Binding Pocket Selection Criteria:

In order to sample a reasonable number of binding pockets on our SARS-CoV-2 protein receptors, we analyzed every protein-ligand complex within the CASF-2016 dataset to find what CASTp-identified binding pocket the main ligand was located in. For the CASTp analysis, we prepared protein structures so that they were split into monomers if they were oligomers (in order to be consistent with our docking protocol) and removed any associated waters, ions, and ligands used in the preparation of the crystal structure. Using the pockets' geometric data from CASTp, we checked which volumetrically ordered pocket the main ligand in the crystal structure was bound to. Using surface area and volume data from the bound pockets, we established the cutoff criteria used in selecting the potential binding pockets for the SARS-CoV-2 protein structures we screened in our workflow.

In the CASF-2016 complexes, we observed that most main ligands were bound to the largest pocket by volume in their respective protein structure (Figure S1, top left). However, we also observed that a significant number of ligands were bound to the second largest pocket by volume (Figure S1, top left). We therefore decided to compare the volumes of the second largest pocket to the first largest for the instances when the main ligand was in the second largest. We observed that in such cases the volumes of the second largest pocket was always larger than 10% of the volumes of the largest pockets for their respective structure (Table S1). Therefore, for our SARS-CoV-2 structures, we sampled all pockets with volumes greater than 10% of the largest pocket volume. We also observed that the surface areas of the second largest pockets were frequently larger than the surface areas of the largest pockets by volume (in half of the instances when the main ligand was inside the second largest pocket by volume) (Table S2). Therefore, we decided to also sample pockets whose volumes are less than 10% of the largest pocket volume if they have greater surface areas than a volumetrically larger pocket.

Table S1. Comparison of the volumes of the second largest and the largest pocket in the cases where the second largest pocket contains the ligands (volumes are in CASTp volume units). *represents a group of 5 complexes where only ligands differ

Protein ID*	Volume of the largest Pocket	Volume of the second largest Pocket	$V_{2nd\ largest}/V_{largest}$
1GPK	846.85	245.44	29%
1MQ6	85.41	42.68	50%
1O3F	25.08	13.94	56%
1O5B	61.37	47.39	77%
3B27	424.33	387.34	91%
3NQ9	60.19	52.15	87%
3WTJ	78.31	53.96	69%
4AGQ	51.37	48.64	95%
4TY7	119.38	39.74	33%
3EBP	346.37	288.50	83%

Table S2. Comparison of the surface areas of the second largest and the largest pocket in the cases where the second largest pocket contains the ligands (in CASTp area units). *represents a group of 5 complexes where only ligands differ.

Protein ID*	Surface Area of the largest Pocket	Surface Area of the second largest Pocket	SA _{2nd largest} -SA _{largest}
1GPK	518.80	425.84	-92.96
1MQ6	113.05	90.12	-22.93
1O3F	89.49	43.01	-46.48
1O5B	74.86	104.17	29.31
3B27	202.27	477.13	274.86
3NQ9	76.24	116.85	40.61
3WTJ	141.43	85.46	-55.97
4AGQ	71.43	131.11	59.68
4TY7	151.06	76.50	-74.56
3EBP	583.80	698.79	114.99

Table S3. PDB IDs and ligands of crystal structures used in score function testing.

SARS-CoV-2 NSP5 Structures		SARS-CoV-2 NSP3 Structures	
PDB ID	Ligand	PDB ID	Ligand
7AXM	Pelitinib	5RSE	ZINC336438345
7JRN	Inhibitor GRL0617	5RTE	ZINC13283576
7JYC	Narlaprevir	5S2A	Z1263529624
7D1M	Inhibitor GC376	5RSB	ZINC1601
6XR3	GRL-024-20	5S2K	Z445856640
6WNP	Boceprevir	5RTM	ZINC2005
6XMK	Inhibitor 7J	5RTF	ZINC2047514
6W79	Inhibitor X77	5S2N	Z1787627869
5RF3	Z1741970824	5RVB	ZINC14419577
5RG1	NCL-00024905	5S2T	Z26781964

Table S4. Structure-based VS identified lead candidates for different targets with their corresponding docking free energy score and cluster ID.

Protein	PDB	Energy Score	Phytochemical	Cluster
NSP1	7K3N	-14.03	Agathisflavone	50
		-14.43	Amentoflavone	5
		-13.94	Bilobetin	5
		-13.95	Fustin	30
		-15.29	Hinokiflavone	5
NSP3	6WEN	-20.50	Amentoflavone	5
		-20.06	Chelerythrine	36
		-20.81	Fagaronine	36
		-22.08	Ginkgetin	5
NSP3	6WEY	-18.44	Amentoflavone	5
		-18.78	Bilobetin	5
		-18.88	Chelerythrine	36
		-18.55	Ginkgetin	5
		-19.05	Hypericin	51
		-17.72	Tannin pyragallol	11
		-17.39	Bilobetin	5
-17.17	Desmethoxyreserpine	3		

NSP5	6Y2E	-17.48	Hinokiflavone	5	
		-19.45	Hypericin	51	
		-18.09	Pseudohypericin	51	
		-17.53	Procyanidin B2	7	
	7AR5	-17.11	Agathisflavone	50	
		-16.35	Amentoflavone	5	
		-16.24	Atalaphillinine	42	
		-18.42	Desmethoxyreserpine	3	
		-16.31	Ginkgetin	5	
		-17.43	Glycycomarin	1	
	NSP7&8	6XIP	-17.79	Hinokiflavone	5
			-16.54	Hypericin	51
			-16.64	5,4'-dihydroxy-3,7,3'- trimethoxyflavone	9
			-16.92	Agathisflavone	50
-16.47			Berberine	36	
-17.08			Bilobetin	5	
6YHU		-16.68	Chelerythrine	36	
		-17.92	Fagaronine	36	
		-16.94	Ginkgetin	5	
		-16.98	Papaverine	11	
NSP9	6WXD	-14.02	10-methoxycamptothecin	49	
		-13.67	Berberine	36	
		-13.52	Camptothecin	49	
		-13.45	Ginkgetin	5	
		-14.06	Hinokiflavone	5	
	6W9Q	-17.11	Berberine	4	
		-18.19	Ginkgetin	5	
		-19.83	Hypericin	51	
		-18.37	Pseudohypericin	51	
		-18.65	3,3'-dimethylquercetin	9	
NSP10	6ZCT	-19.10	Apogossypol	17	
		-18.75	Bilobetin	5	
		-19.02	Hinokiflavone	5	
		-15.59	13',ii8-biapigenin	50	
		-15.99	Cyanidol	6	
		-15.86	Emetine	14	
	7NIO	-16.32	Emodin-bianthrone	0	
		-15.42	Gambirinin-b3	7	
		-15.75	Hinokiflavone	5	
		-16.14	Procyanidin	7	
NSP13	6ZSL	-16.05	Procyanidin-a-2	7	
		-18.72	13',ii8-biapigenin	50	
		-19.06	Agathisflavone	50	
		-17.51	Amentoflavone	5	
		-17.49	Hypericin	51	
		-18.61	Pseudohypericin	51	
	7NIO	-18.18	Rhein	0	
		-19.06	(-)-Epicatechin-3-o-gallate	48	
		-18.90	Agathisflavone	50	
		-18.56	Amentoflavone	5	
6ZSL	-19.63	Berberine	36		
	-18.72	Fagaronine	36		
	-18.86	Hypericin	51		
	-18.21	Isolicoflavonol	20		
	-19.05	Pseudohypericin	51		
	-20.01	Agathisflavone	50		
-20.40	Amentoflavone	5			

NSP15	6VWW	-18.67	Hinokiflavone	5
		-18.28	Hypericin	51
		-19.87	Pseudohypericin	51
Spike RBD	6XM4	-15.99	Agathisflavone	50
		-16.08	Atalaphillinine	42
		-16.90	Chelerythrine	36
		-16.92	Ginkgetin	5
Full Spike Closed State	6VXX	-15.53	Amentoflavone	5
		-14.85	Apogossypol	17
		-14.64	Bilobetin	5
		-14.35	Gambirinin-b3	7
		-15.55	Ginkgetin	5
		-15.10	Gossypol	17
		-17.31	Hypericin	51
-17.09	Pseudohypericin	51		

Table S5. Ligand-based VS identified lead candidates for different targets with their corresponding docking free energy score and cluster ID.

Protein	PDB	Energy Score	Phytochemical	Classified Cluster
NSP3	6WEN	-19.64	Palmatine chloride	36
		-18.62	Nitidine chloride	36
	6WEY	-17.97	Sanguinarine chloride	36
		-19.75	Sciadopitysin	5
NSP5	6Y2E	-16.95	Morusin	42
		-18.33	Sciadopitysin	5
		-17.00	7-Ethylcamptothecin	49
		-16.80	Dihydrochelerythrine	36
	6XIP	-17.31	Epiberberine	36
		-16.64	HydroxyCamptothecine	49
		-16.64	Irinotecan	49
		-16.51	Irinotecan hydrochloride trihydrate	49
		-16.41	Palmatine chloride	36
		-16.98	Palmatrubine	36
		-17.69	Sciadopitysin	5
		-13.41	7-Ethylcamptothecin	49
		-14.33	Acacetin	5
		-18.50	Berberrubine	36
		-14.39	Columbamine	36
		-15.63	Dehydrocorydalin	36
		-14.01	Demethyleneberberine	36
		-16.57	Dihydrochelerythrine	36
		-14.60	Epiberberine	36
		-14.58	Eupatilin	50
-13.66	HydroxyCamptothecine	49		
NSP7&8	6YHU	-15.30	Irinotecan	49
		-15.00	Isoorientin	50
		-14.59	Isovitexin	50
		-13.69	Jatrorrhizine	36
		-15.49	Lycobetaine	36
		-18.94	Nitidine chloride	36
		-15.42	Oroxylin A	50
		-13.90	Palmatine chloride	36
		-15.43	Palmatrubine	36

		-14.17	Pectolarigenin	50
		-13.77	Salvigenin	50
		-16.40	Sanguinarine chloride	36
		-14.77	Sciadopitysin	5
		-14.64	Topotecan hydrochloride	49
		-13.43	Wogonoside	50
NSP9	6WXD	-17.79	Sciadopitysin	5
NSP13	6ZSL	-18.64	Columbamine	36
		-18.49	Coptisine	36
		-18.24	Dihydrochelerythrine	36
		-16.23	Columbamine	36
		-15.77	Nitidine chloride	36
Spike RBD	6XM4	-17.30	Palmatine chloride	36
		-16.46	Sanguinarine chloride	36
		-16.49	Sciadopitysin	5

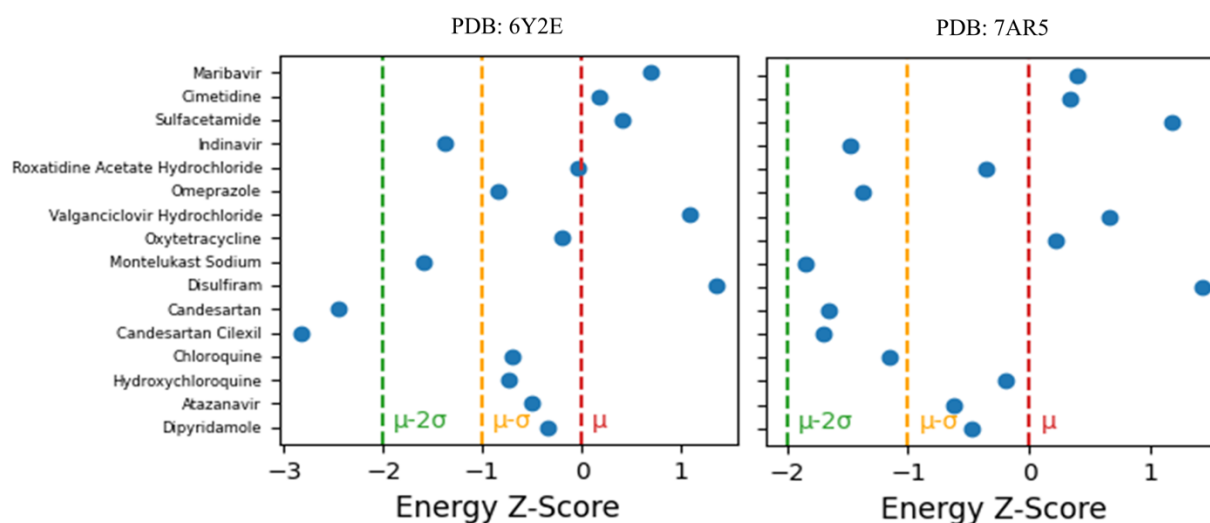


Figure S2. (A) Benchmarking thresholds using docking data of 16 known main protease inhibitors for main protease structures 6Y2E and 7AR5. The compounds were docked through our main docking protocol, and their energy scores were compared to those the mean energy score of the 272 initially docked phytochemicals. 69% of them were below the mean in the case of 6Y2E, and 63% were below the mean in the case of 7AR5.

Table S6. The percentage of docking energy scores in each specified range for predicted lead phytochemicals after LBVS.

Protein	PDB	< $\mu-2\sigma$	$\mu \sim \mu-2\sigma$	> μ
NSP1	7K3N	0%	82.14%	17.86%
NSP3	6WEN	5.88%	94.12%	0%
	6WEY	17.65%	82.35%	0%
NSP5	6Y2E	28.57%	71.43%	0%
	7AR5	0%	71.43%	28.57%
NSP7&8	6XIP	25.71%	71.43%	2.86%

	6YHU	62.5%	37.5%	0%
NSP9	6WXD	16.67%	83.33%	0%
	6W9Q	0%	100%	0%
NSP10	6ZCT	0%	100%	0%
NSP13	7NIO	0%	82.35%	17.65%
	6ZSL	9.38%	84.37%	6.25%
NSP15	6VWW	0%	50%	50%
RBD	6XM4	13.89%	66.67%	19.44%
Spike	6VXX	0%	100%	0%
Total		16.50%	72.05%	11.45%

Table S7. Random Under-sampling Confusion Matrix. When the energy score of at least 2 standard deviations below the average (or energy z-score less than -2) was used to determine the actual positive and negative, a precision of 0.16, recall of 0.73, fall-out of 0.47, and F-score of 0.27 were obtained. When an energy z-score less than -1 was used to determine the actual positive and negative, a precision of 0.45, recall of 0.68, fall-out of 0.4, and F-score of 0.41 were obtained.

	Threshold z-score = -2		Threshold z-score = -1	
	Predicted Positive	Predicted Negative	Predicted Positive	Predicted Negative
Actual Positive	TP=49 (8.22%)	FN=18 (3.02%)	TP=133 (22.32%)	FN=62 (10.40%)
Actual Negative	FP=249 (41.78%)	TN=280 (46.98%)	FP=165 (27.68%)	TN=236 (39.60%)

Table S8. SwissADME data for the comparison of camptothecin and rhein to 3 COVID-19 drugs and 1 HIV drug (Doravirine).

Molecule	Camptothecin	Rhein	Remdesivir (Gilead)	Molnupiravir (Merck)	Paxlovid (Pfizer)	Doravirine
MR	95.31	68.81	150.43	76.02	125.68	95.36
TPSA	81.42	114.73	213.36	143.14	131.4	105.7
iLOGP	2.49	1.28	3.52	0.17	3.01	2.69
XLOGP3	1.74	2.23	1.91	-1.34	2.17	2.09
WLOGP	1.82	0.24	2.21	-1.65	1.6	3.81
MLOGP	1.64	0.29	0.18	-1.15	0.41	1.89
Silicos-IT Log P	3.29	1.96	-0.05	-1.82	2.25	3.27
Consensus Log P	2.2	1.2	1.56	-1.16	1.89	2.75
ESOL Log S	-3.49	-3.36	-4.12	-0.83	-3.58	-3.9
ESOL Solubility (mol/l)	3.27E-04	4.39E-04	7.59E-05	1.46E-01	2.64E-04	1.26E-04
ESOL Class	Soluble	Soluble	Moderately soluble	Very soluble	Soluble	Soluble
Ali Log S	-3.07	-4.27	-6.01	-1.17	-4.56	-3.94
Ali Solubility (mol/l)	8.58E-04	5.31E-05	9.69E-07	6.81E-02	2.74E-05	1.15E-04
Ali Class	Soluble	Moderately Soluble	Poorly soluble	Very soluble	Moderately soluble	Soluble
Silicos-IT LogSw	-5.83	-3.46	-4.77	0.12	-3.94	-5.71

Silicos-IT Solubility (mol/l)	1.49E-06	3.44E-04	1.71E-05	1.32	1.14E-04	1.95E-06
Silicos-IT class	Moderately soluble	Soluble	Moderately soluble	Soluble	Soluble	Moderately soluble
GI absorption	High	High	Low	Low	High	High
BBB permeant	No	No	No	No	No	No
Pgp substrate	Yes	No	Yes	No	Yes	No
CYP1A2 inhibitor	Yes	No	No	No	No	No
CYP2C19 inhibitor	No	No	No	No	No	No
CYP2C9 inhibitor	Yes	No	No	No	No	Yes
CYP2D6 inhibitor	No	No	No	No	No	No
CYP3A4 inhibitor	Yes	No	Yes	No	Yes	No
log K _p (cm/s)	-7.19	-6.44	-8.62	-9.26	-7.81	-7.41
Total Drug-likeness violations	0	0	11	3	2	0
PAINS alerts	0	1	0	0	0	0
Brenk alerts	0	0	1	1	0	0
Lead-likeness violations	0	0	2	0	2	1
Synthetic Accessibility	3.84	2.55	6.33	4.49	4.82	3.28
Bioavailability Score	0.55	0.56	0.17	0.55	0.55	0.55

Docking Script RosettaLigand Score Function

```

<ROSETTASCRIPTS>
  <SCOREFXNS>
    <ScoreFunction name="ligand_soft_rep" weights="ligand_soft_rep">
      <Reweight scoretype="fa_elec" weight="0.42"/>
      <Reweight scoretype="hbond_bb_sc" weight="1.3"/>
      <Reweight scoretype="hbond_sc" weight="1.3"/>
      <Reweight scoretype="rama" weight="0.2"/>
    </ScoreFunction>
    <ScoreFunction name="hard_rep" weights="ligand">
      <Reweight scoretype="fa_intra_rep" weight="0.004"/>
      <Reweight scoretype="fa_elec" weight="0.42" />
      <Reweight scoretype="hbond_bb_sc" weight="1.3"/>
      <Reweight scoretype="hbond_sc" weight="1.3"/>
      <Reweight scoretype="rama" weight="0.2"/>
    </ScoreFunction>
  </SCOREFXNS>
  <LIGAND_AREAS>
    <LigandArea name="inhibitor_dock_sc" chain="X" cutoff="6.0" add_nbr_radius="true"
all_atom_mode="true"/>
    <LigandArea name="inhibitor_final_sc" chain="X" cutoff="6.0" add_nbr_radius="true"
all_atom_mode="true"/>
    <LigandArea name="inhibitor_final_bb" chain="X" cutoff="7.0" add_nbr_radius="false"
all_atom_mode="true" Calpha_restraints="0.3"/>
  </LIGAND_AREAS>
  <INTERFACE_BUILDERS>
    <InterfaceBuilder name="side_chain_for_docking" ligand_areas="inhibitor_dock_sc"/>
    <InterfaceBuilder name="side_chain_for_final" ligand_areas="inhibitor_final_sc"/>
    <InterfaceBuilder name="backbone" ligand_areas="inhibitor_final_bb" extension_window="3"/>
  </INTERFACE_BUILDERS>
  <MOVEMAP_BUILDERS>
    <MoveMapBuilder name="docking" sc_interface="side_chain_for_docking" minimize_water="false"/>
    <MoveMapBuilder name="final" sc_interface="side_chain_for_final" bb_interface="backbone"
minimize_water="false"/>
  </MOVEMAP_BUILDERS>
  <SCORINGGRIDS ligand_chain="X" width="30">
    <ClassicGrid grid_name="classic" weight="1.0"/>
  </SCORINGGRIDS>
</TASKOPERATIONS>

```

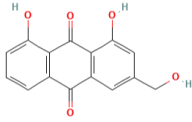
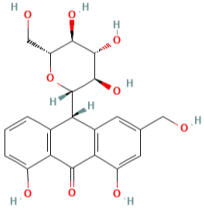
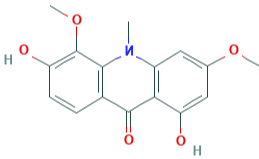
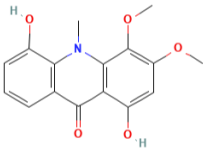
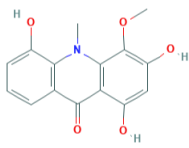
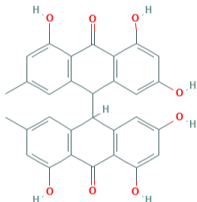
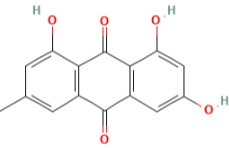
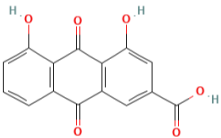
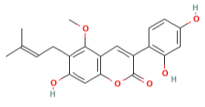
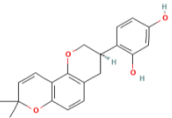
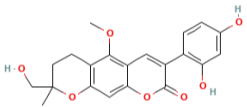
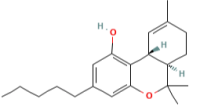
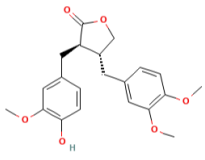
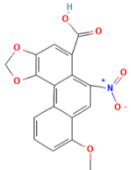
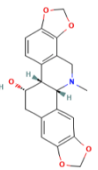
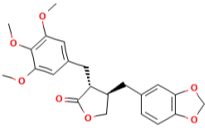

```

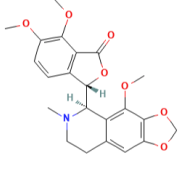
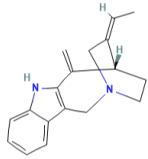
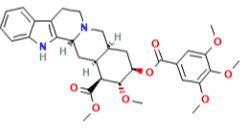
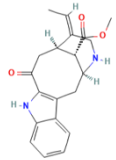
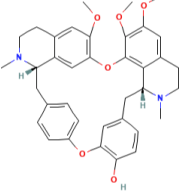
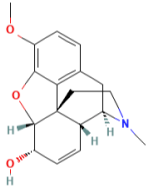
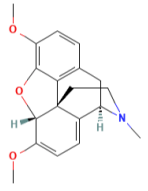
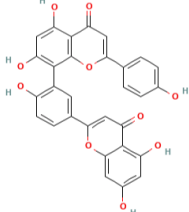
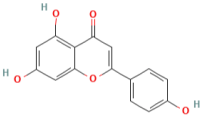
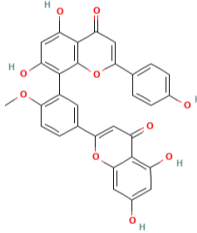
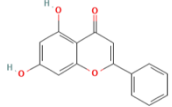
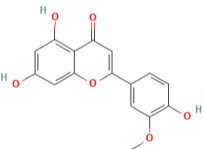
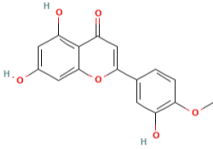
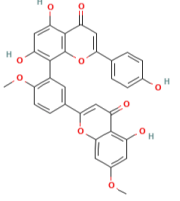
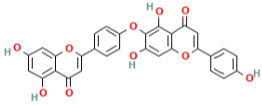
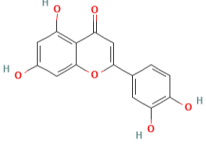
</TASKOPERATIONS>
<MOVERS>
  <StartFrom name="start_from" chain="X"> </StartFrom>
  <Transform name="transform" chain="X" box_size="14" move_distance="0.2" angle="20" cycles="1000"
repeats="1" temperature="5" initial_perturb="5.0"/>
  <HighResDocker name="high_res_docker" cycles="6" repack_every_Nth="3" scorefxn="ligand_soft_rep"
movemap_builder="docking"/>
  <FinalMinimizer name="final" scorefxn="hard_rep" movemap_builder="final"/>
  <InterfaceScoreCalculator name="add_scores" chains="X" scorefxn="hard_rep"/>
</MOVERS>
<PROTOCOLS>
  <Add mover_name="start_from"/>
  <Add mover_name="transform"/>
  <Add mover_name="high_res_docker"/>
  <Add mover_name="final"/>
  <Add mover_name="add_scores"/>
</PROTOCOLS>
</ROSETTASCRIPTS>

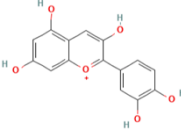
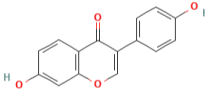
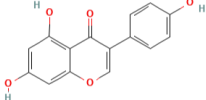
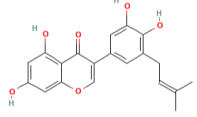
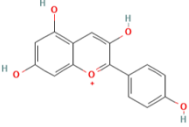
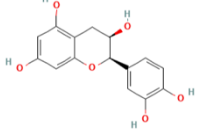
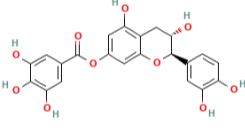
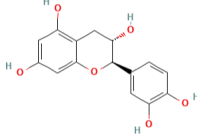
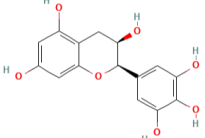
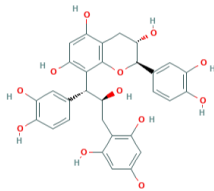
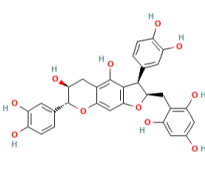
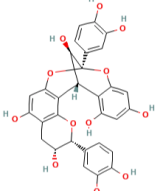
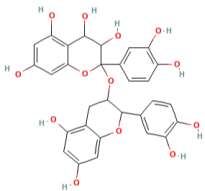
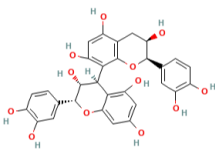
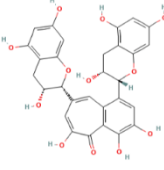
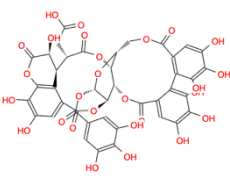
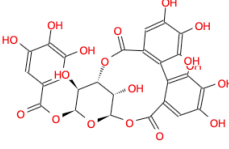
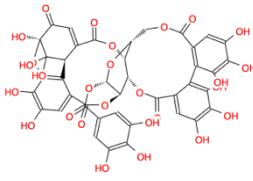
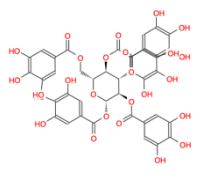
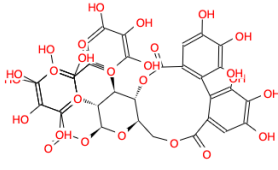
```

Figure S3: Docking Script used with RosettaLigand (pre-talaris 2013) for ligand docking runs.

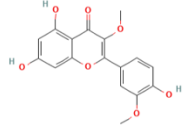
Table S9 Clustering of the 272 anti-viral phytochemicals used in SBVS. Compound names and ligand IDs used in the supplementary spreadsheet titled “Ligand_Library_Key_SBVS” are given. Cluster numbers are given in the left column. Cluster numbers here correspond to the cluster numbers on the dendrogram in Figure 6A.

0	 ALOE-EMODIN (27)	 ALOIN (28)	 CITPRESSINE-1 (95)	 CITRUSININE-1 (97)	
	 CITRUSININE-II (98)	 EMODIN-BIANTHRONE (128)	 EMODIN (136)	 RHEIN (234)	
1	 GLYCYCOUMARIN (154)	 GLABRIDIN (158)	 LICOPYRANOCUMARIN (185)	 TETRAHYDROCANNABINOL (255)	
2	 ARCTIGENIN (41)	 ARISTOLCHIC-ACID (42)	 CHELIDONINE (82)	 DIHYDROANHYDROPODORZIZOL (118)	 LYCORINE (194)

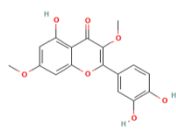
	 <p>NARCOTINE (203)</p>
3	 <p>APPARICINE (39)</p>  <p>DESMETHOXYRESERPINE (122)</p>  <p>PERIVINE (213)</p>
4	 <p>BERBAMINE (55)</p>  <p>CODEINE (99)</p>  <p>THEBAINE (257)</p>
5	 <p>AMENTOFLAVONE (33)</p>  <p>APIGENIN (37)</p>  <p>BILOBETIN (65)</p>  <p>CHRYSLIN (87)</p>  <p>CHRYSOERIOL (88)</p>  <p>DIOSMETIN (120)</p>  <p>GINKGETIN (151)</p>  <p>HINOKIFLAVONE (163)</p>  <p>LUTEOLIN (190)</p>

6	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;">  <p>CYANIDOL (104)</p> </div> <div style="text-align: center;">  <p>DAIDZEIN (109)</p> </div> <div style="text-align: center;">  <p>GENISTEIN (149)</p> </div> <div style="text-align: center;">  <p>GLYCYRRHISOFLAVONE (156)</p> </div> </div> <div style="text-align: center; margin-top: 20px;">  <p>PELARGONIDIN (211)</p> </div>
7	<div style="display: grid; grid-template-columns: repeat(4, 1fr); gap: 10px;"> <div style="text-align: center;">  <p>(-)-EPICATECHIN (4)</p> </div> <div style="text-align: center;">  <p>CATECHIN-7-O-GALLATE (76)</p> </div> <div style="text-align: center;">  <p>CATECHIN (77)</p> </div> <div style="text-align: center;">  <p>EPIGALLOCATECHIN (137)</p> </div> <div style="text-align: center;">  <p>GAMBIRININ-A1 (147)</p> </div> <div style="text-align: center;">  <p>GAMBIRININ-B3 (148)</p> </div> <div style="text-align: center;">  <p>PROCYANIDIN-A-2 (219)</p> </div> <div style="text-align: center;">  <p>PROCYANIDIN (220)</p> </div> <div style="text-align: center;">  <p>PROCYANIDIN B2 (227)</p> </div> <div style="text-align: center;">  <p>THEAFLAVIN (256)</p> </div> </div>
8	<div style="display: grid; grid-template-columns: repeat(4, 1fr); gap: 10px;"> <div style="text-align: center;">  <p>CHEBULAGIC-ACID (80)</p> </div> <div style="text-align: center;">  <p>TANNIN (270)</p> </div> <div style="text-align: center;">  <p>GERANIIN (272)</p> </div> <div style="text-align: center;">  <p>PENTAGALLOYL GLUCOSE (273)</p> </div> <div style="text-align: center; margin-top: 20px;">  <p>EUGENIIN (274)</p> </div> </div>

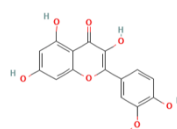
9



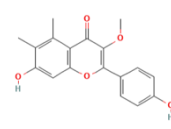
3,3'-DIMETHYLQUERCETIN (11)



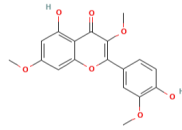
3,7'-DIMETHYLQUERCETIN (12)



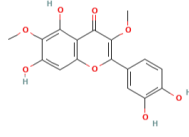
3-METHYLQUERCETIN (13)



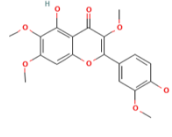
4',7-DIHYDROXY-3-METHOXY-5,6-DIMETHYLFLAVONE (15)



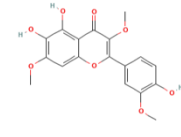
5,4'-DIHYDROXY-3,7,3'-TRIMETHOXYFLAVONE (17)



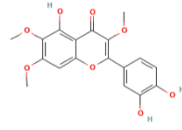
AXILLARIN (50)



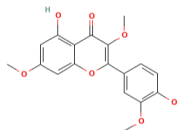
CHRYSOSPLENETIN (89)



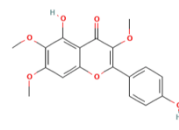
CHRYSOSPLENOL-C (90)



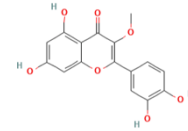
CHRYSOSPLENOL-D (91)



PACHYPODOL (209)

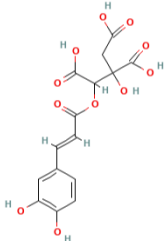


PENDULETIN (212)

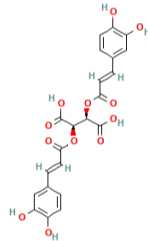


QUERCETIN-3-O-METHYL-ETHER (230)

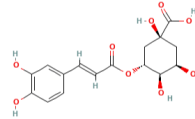
10



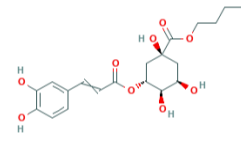
2-O-CAFFEYOYL-(+)-ALLOHYDROXYCITRIC-ACID (10)



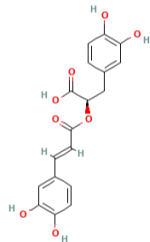
CHICORIC-ACID (83)



CHLOROGENIC ACID (85)

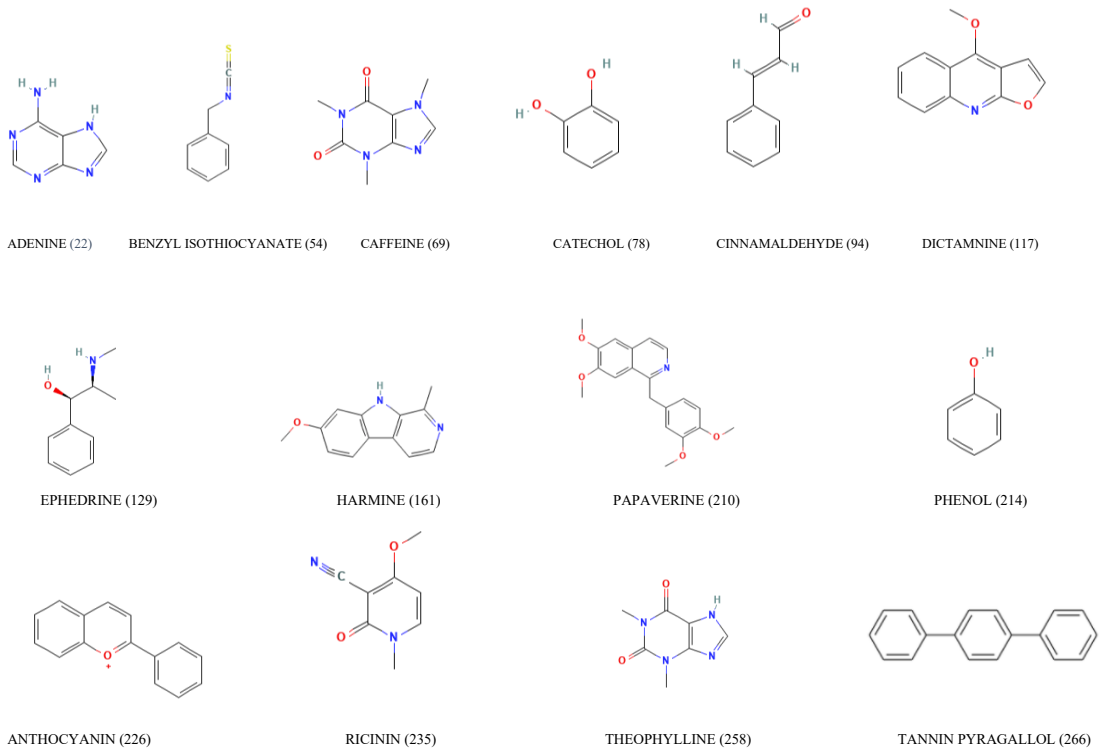


CHLOROGENIC ACID BUTYL ESTER (86)

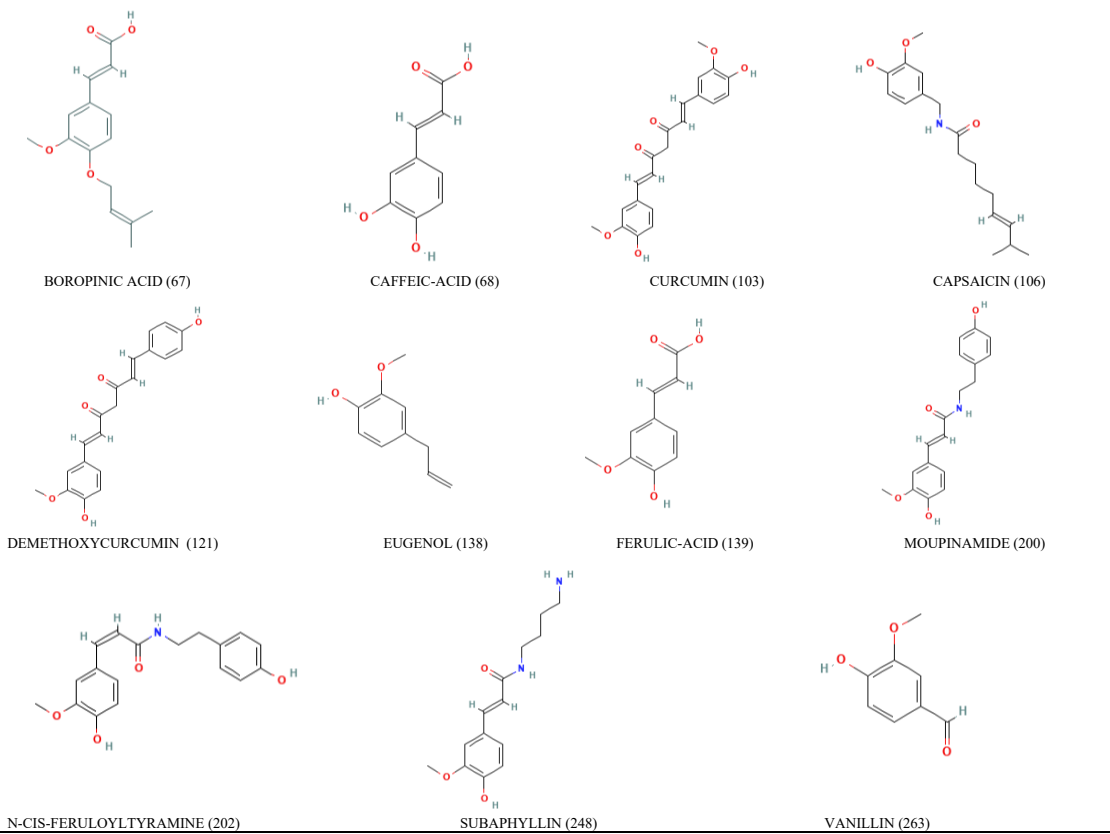


ROSMARINIC ACID (236)

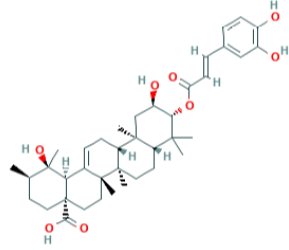
11



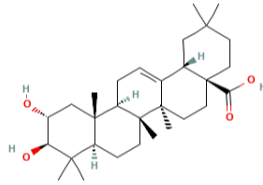
12



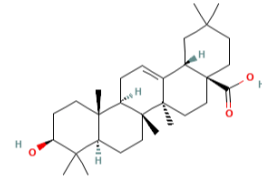
13



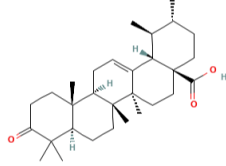
3-O-TRANS-CAFFEOYLTORMENTIC ACID (14)



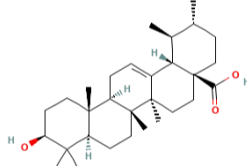
MASLINIC ACID (196)



OLEANOLIC ACID (206)

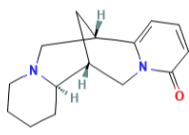


URSONIC ACID (261)

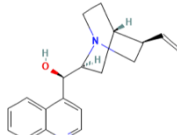


URSOLIC ACID (262)

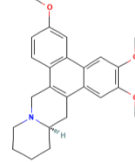
14



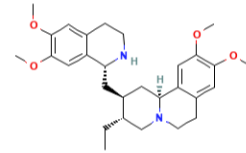
ANAGRINE (34)



CINCHONIDINE (93)

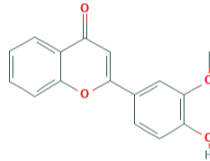


CRYPTOPEURINE (102)

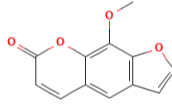


EMETINE (127)

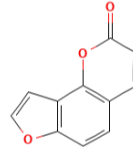
15



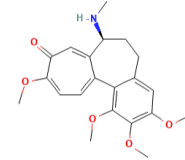
4-HYDROXY-3-METHOXYFLAVONE (16)



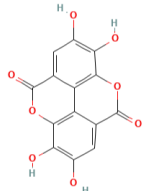
8-METHOXY PSORALEN (21)



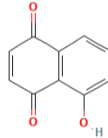
ANGELICIN (35)



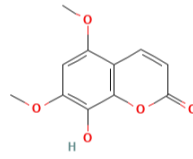
COLCHAMINE (100)



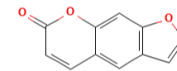
ELLAGIC ACID (126)



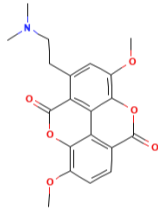
JUGLONE (179)



LEPTODACTYLONE (191)

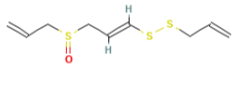


PSORALEN (225)

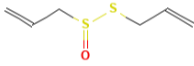


TASPINE (253)

16



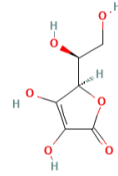
AJOENE (24)



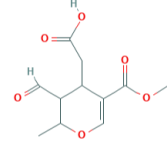
ALLICIN (25)



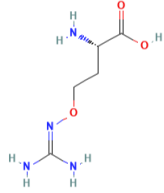
ALLYL ALCOHOL (26)



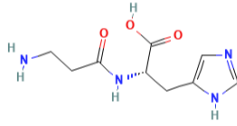
ASCORBIC ACID (45)



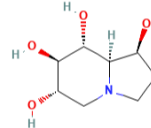
CALCIUM ELENOLATE (70)



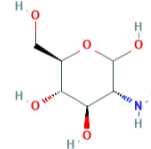
CANAVANINE (72)



CARNOSINE (73)



CASTANOSPERMINE (74)



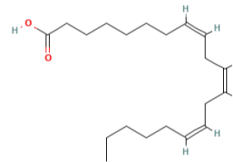
D-GLUCOSAMINE (108)



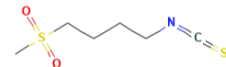
DIALLYL DISULFIDE (115)



DIALLYL TRISULFIDE (116)



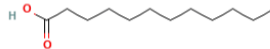
DIHOMO GAMMA LINOLEIC ACID (123)



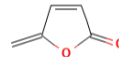
ERYSOLIN (134)



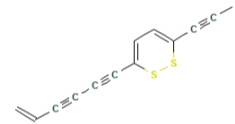
FORMALDEHYDE (141)



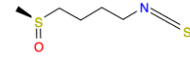
LAURIC ACID (183)



PROTOANEMONIN (222)



THIARRUBRINE-A (259)



SULPHORAPHAN (267)

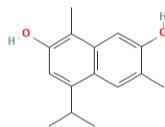


NONACOSANE (268)

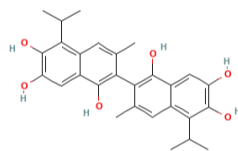


OCTACOSANOL (269)

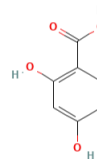
17



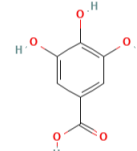
2,7-DIHYDROXYCADALENE (9)



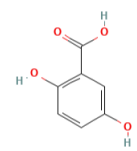
APOSSYTOPOL (38)



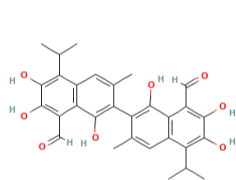
BETA-RESERICYLIC ACID (58)



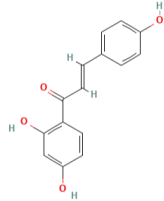
GALLIC ACID (146)



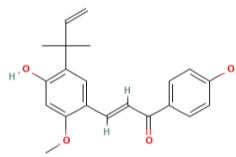
GENTISIC ACID (150)



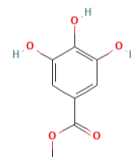
GOSSYPOL (157)



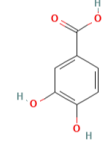
ISOLIQUIRITIGENIN (176)



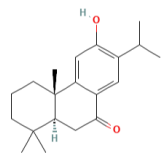
LICOCHALCONE-A (184)



METHYL GALLATE (197)

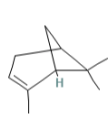


PROTocatechuic ACID (223)

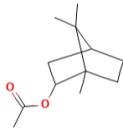


SUGIOL (252)

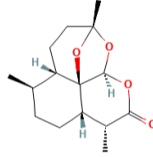
18



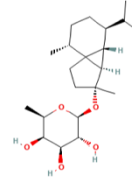
ALPHA-PINENE (31)



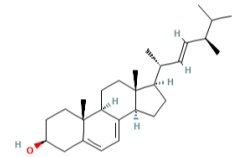
BORNYL ACETATE (66)



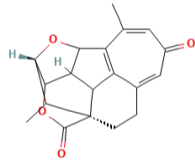
DEOXYARTEMISININ (113)



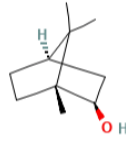
EPICUBEBOL GLYCOSIDE (130)



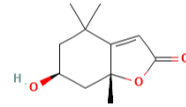
ERGOSTEROL (133)



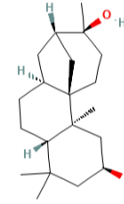
HAINANOLIDE (160)



ISOBORNEOL (173)

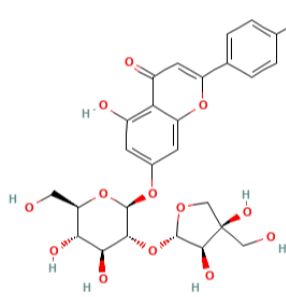


LOLIOLIDE (193)

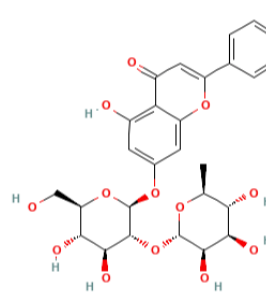


STEMODIN (245)

19

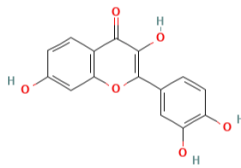


APIIN (51)

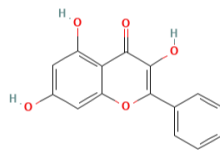


RHOIFOLIN (238)

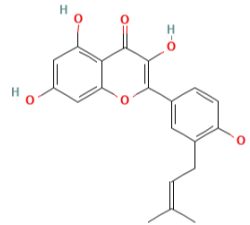
20



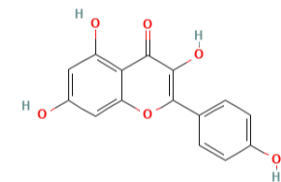
FISETIN (140)



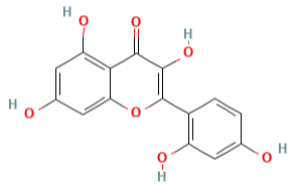
GALANGIN (145)



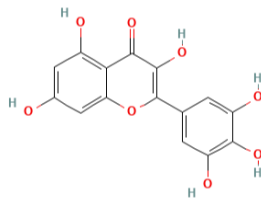
ISOLICOFLAVONOL (175)



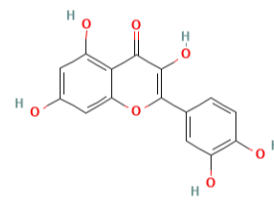
KAEMPFEROL (181)



MORIN (198)

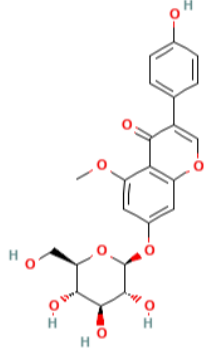


MYRICETIN (199)

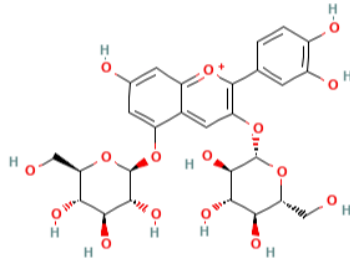


QUERCETIN (231)

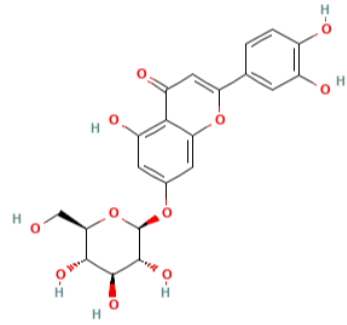
21



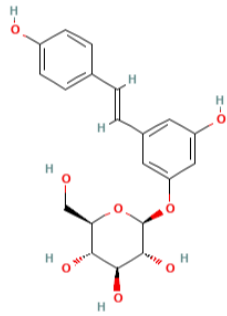
5-O-METHYLGENISTIN (19)



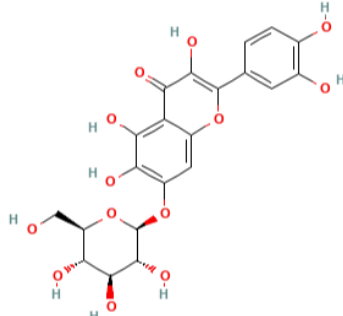
CYANIN (105)



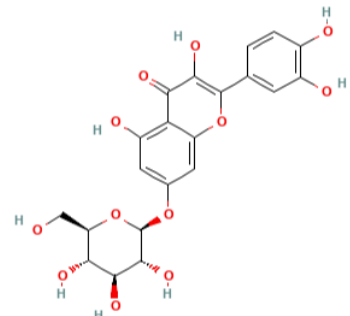
LUTEOLIN-7- GLUCOSIDE (189)



POLYDATIN (217)

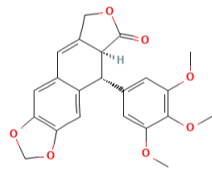


QUERCETAGITIN (229)

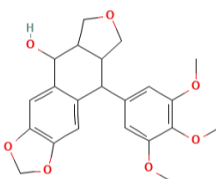


QUERCIMERITRIN (232)

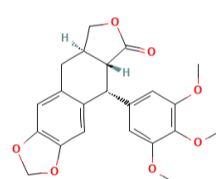
22



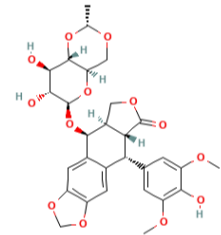
ALPHA-APOCROPODOPHYLLOTOXIN (29)



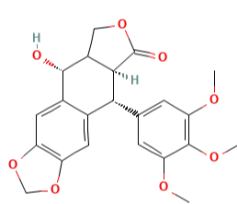
ANHYDROPODOPHYLLOL (36)



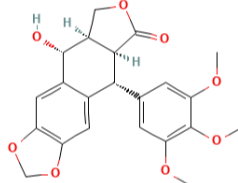
DEOXYPODOPHYLLOTOXIN (114)



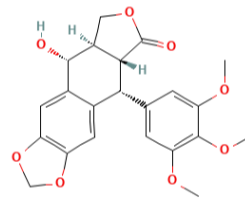
ETOPOSIDE (135)



LIGNANS (186)

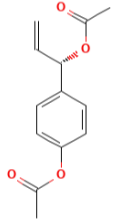


PICROPODOPHYLLOTOXIN (215)

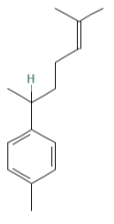


PODOPHYLLOTOXIN (216)

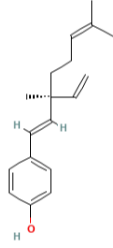
23



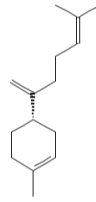
(1'S)-1'-ACETOXYCHAVICOL ACETATE (6)



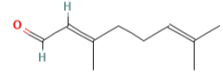
AR-CURCUMENE (40)



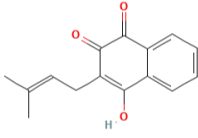
BAKUCHIOL (53)



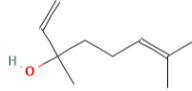
BETA_BISABOLENE (61)



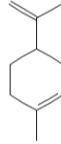
CITRAL (96)



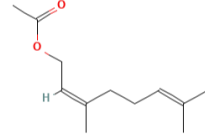
LAPACHOL (182)



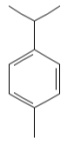
LINALOOL (187)



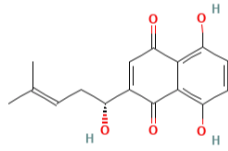
LIMONENE (192)



NERYL ACETATE (205)

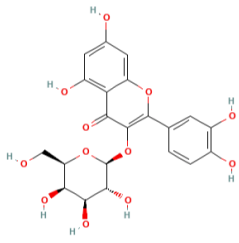


P CYMENE (208)

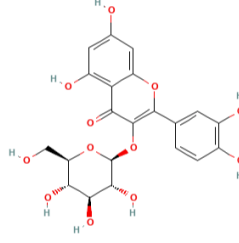


SHIKONIN (250)

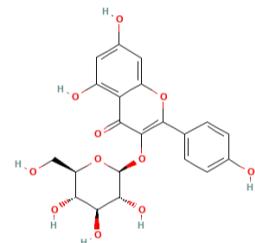
24



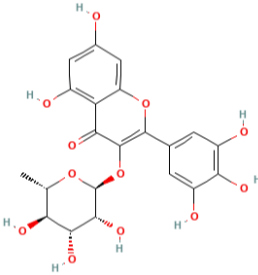
HYPERIN (169)



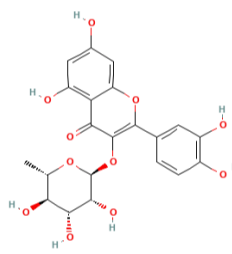
ISOQUERCETIN (177)



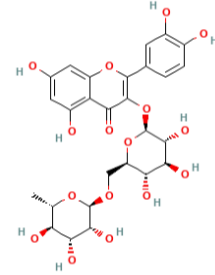
KAEMPFEROL-3-O-GLUCOSIDE (180)



MYRICITRIN (201)

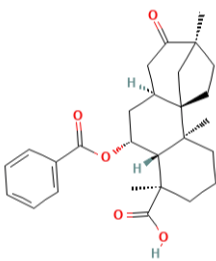


QUERCITRIN (233)

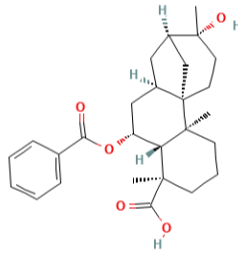


RUTIN (237)

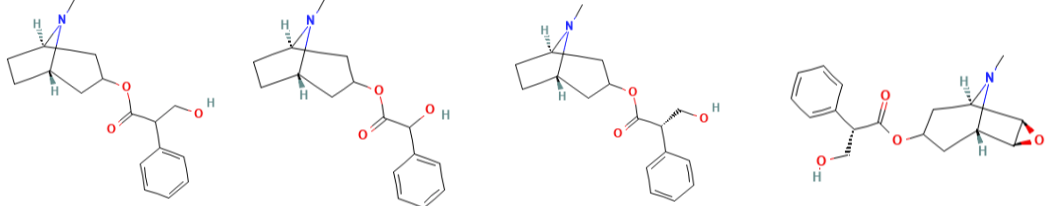
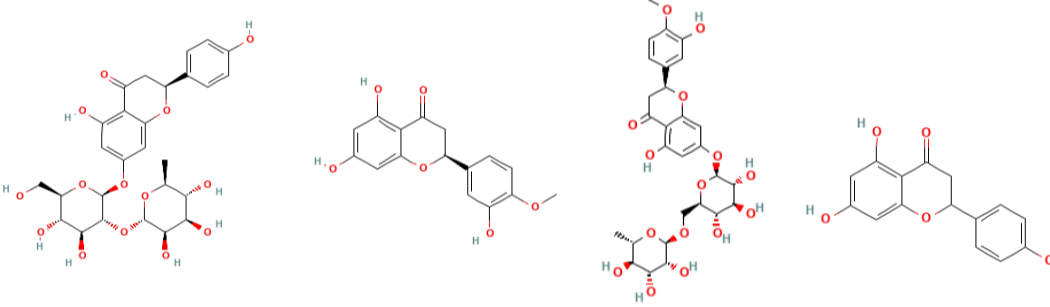
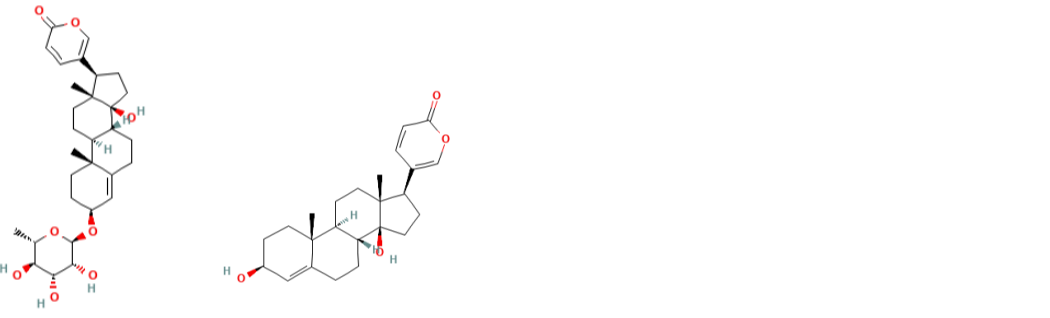
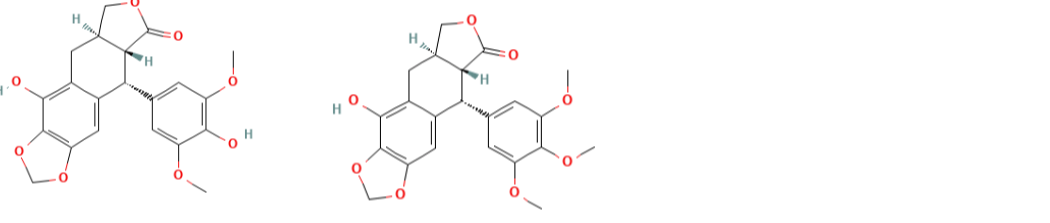
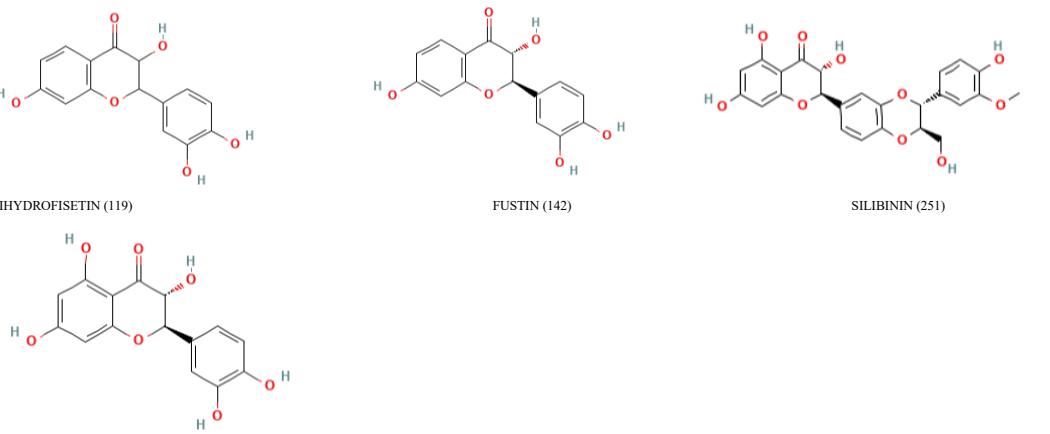
25

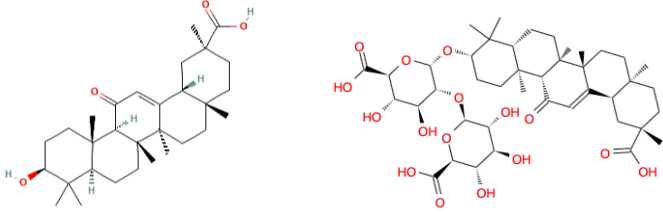
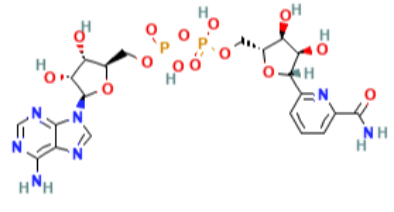
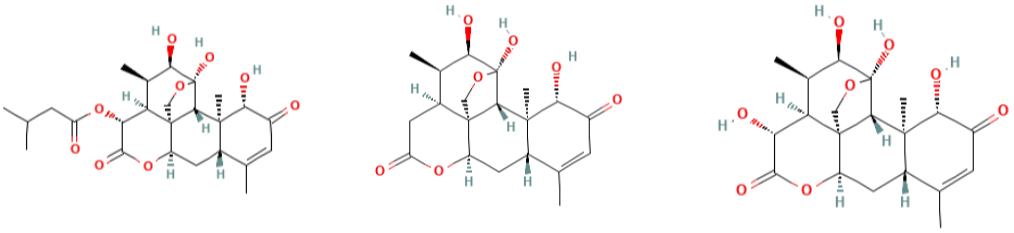
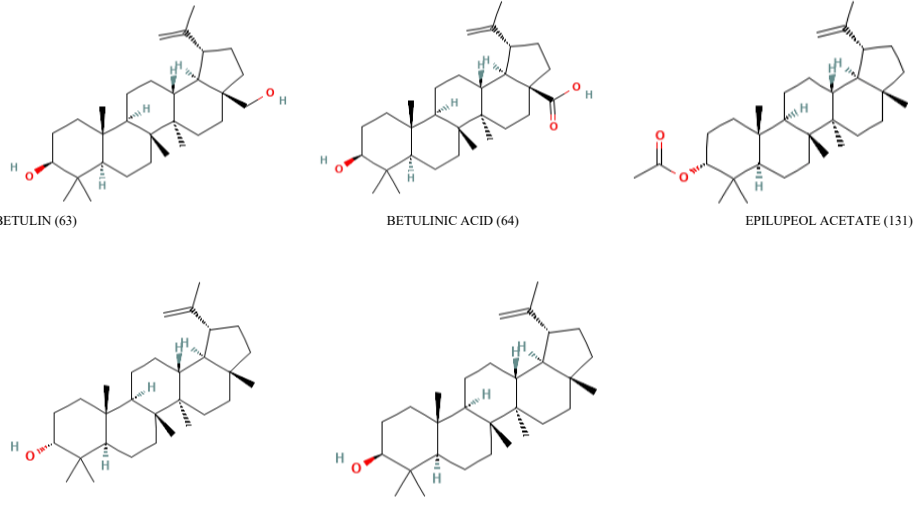
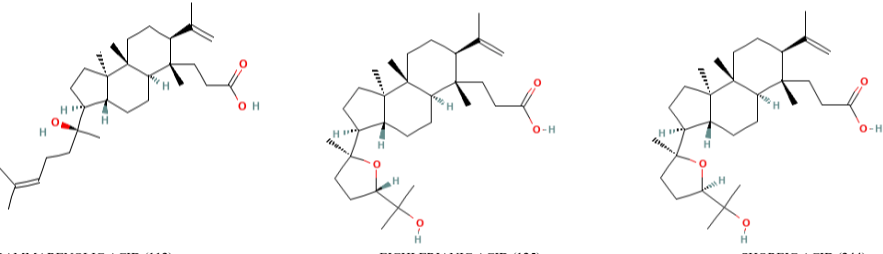


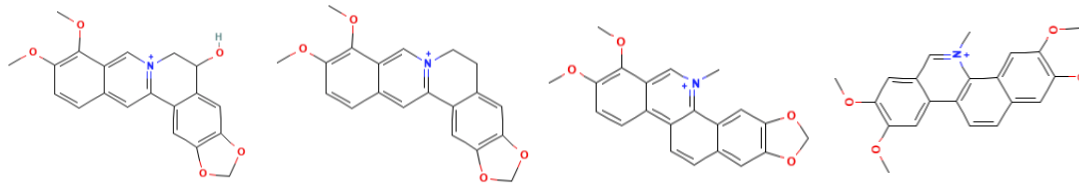
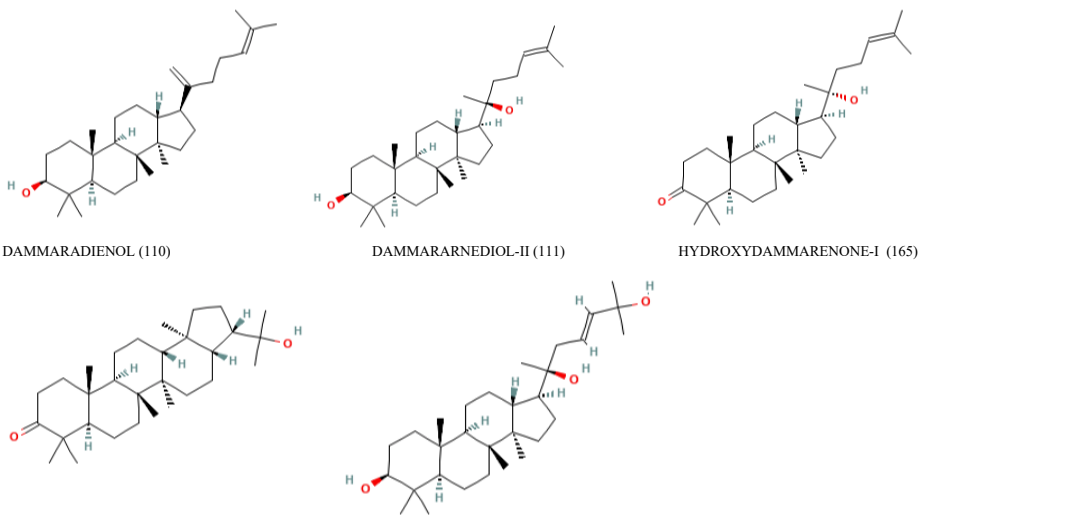
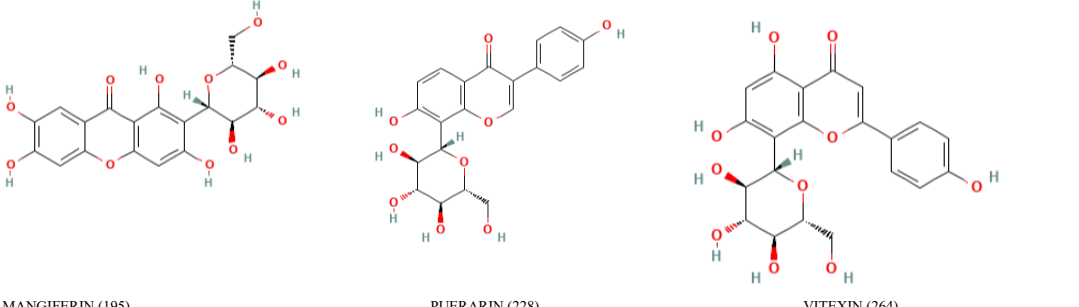

SCOPADULCIC-ACID-B (240)

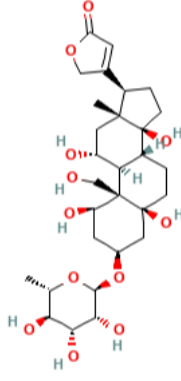
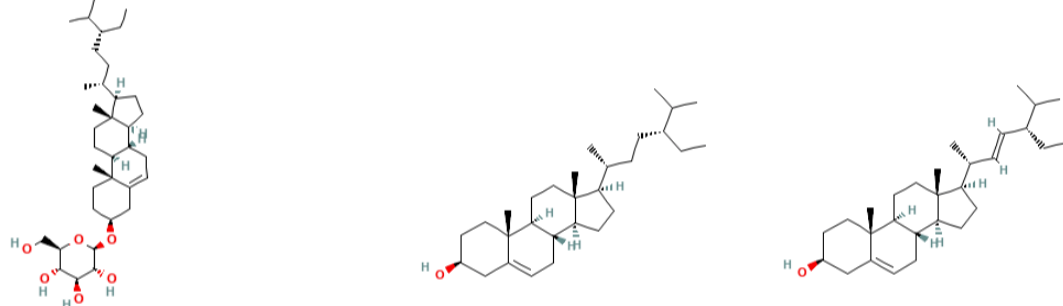
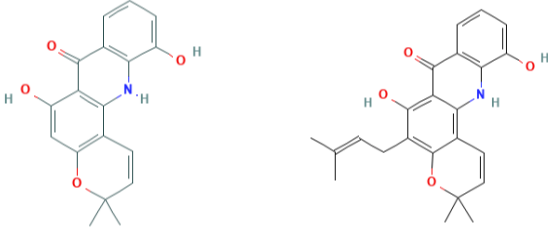
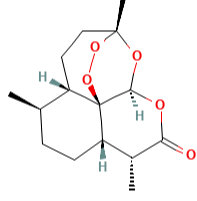
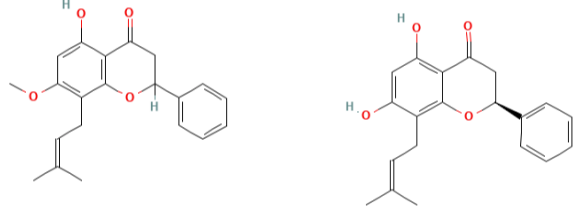


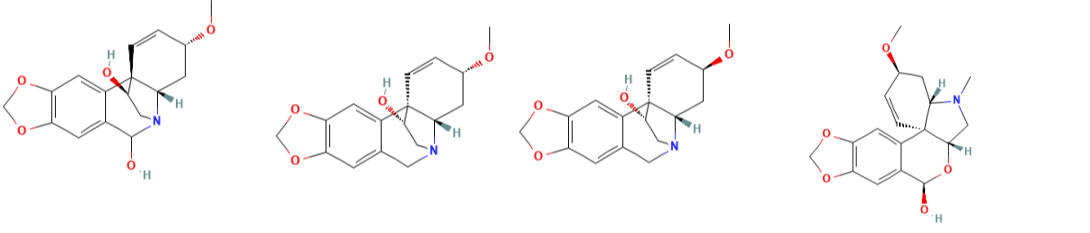
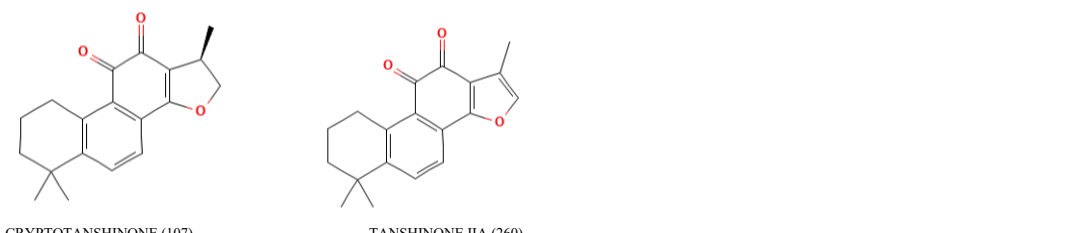

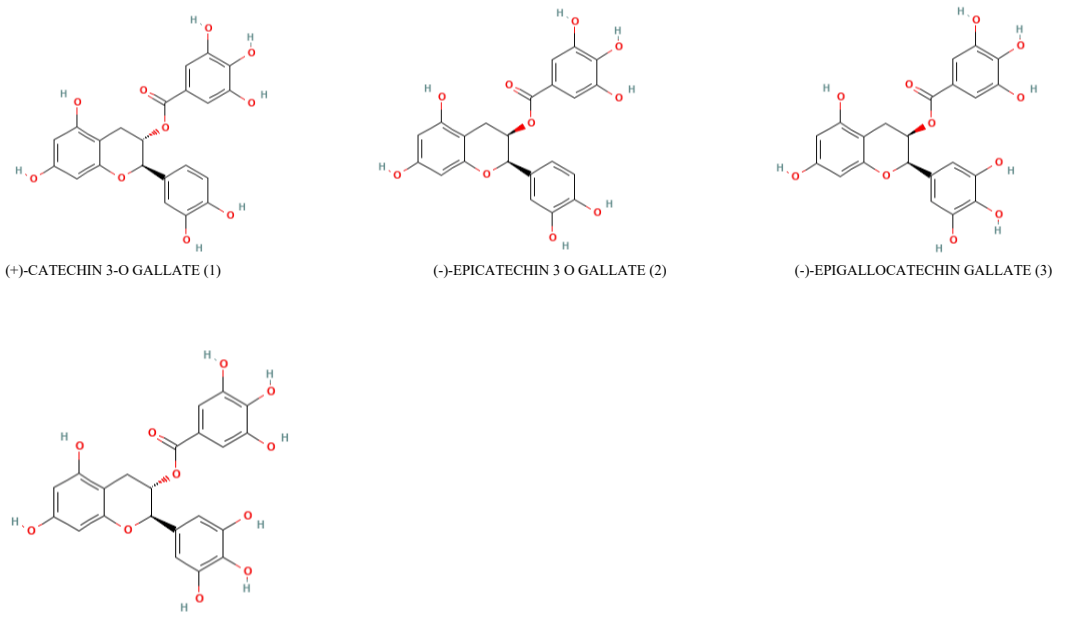
SCOPADULIN (241)

26	 <p>ATROPINE (48) HOMATROPINE (164) HYOSCYAMINE (167) SCOPOLAMINE (242)</p>
27	 <p>AURANTININ (49) HESPERITIN (162) HESPERIDIN (171) NARINGENIN (204)</p>
28	 <p>PROSCILLARIDIN-A (221) SCILLARENIN (239)</p>
29	 <p>ALPHA PELTATINE (32) BETA PELTATINE (62)</p>
30	 <p>DIHYDROFISSETIN (119) FUSTIN (142) SILIBININ (251)</p> <p>TAXIFOLIN (254)</p>

31	 <p>GLYCYRRHETIC ACID (155) GLYCYRRHIZIN (271)</p>
32	 <p>FENUGREEKINE (144)</p>
33	 <p>CASTELANONE (75) CHAPARRINONE (79) GLAUCARUBOLONE (153)</p>
34	 <p>BETULIN (63) BETULINIC ACID (64) EPILUPEOL ACETATE (131)</p> <p>EPILUPEOL (132) LUPEOL (188)</p>
35	 <p>DAMMARENIC ACID (112) EICHLERIANIC ACID (125) SHOREIC ACID (244)</p>

36	 <p>BERBERASTINE (56) BERBERINE (57) CHELERYTHRINE (81) FAGARONINE (143)</p>
37	 <p>DAMMARADIENOL (110) DAMMARARNEDIOL-II (111) HYDROXYDAMMARENONE-I (165)</p> <p>HYDROXYHOPANONE (166) ISOFOQUIEROL (174)</p>
38	 <p>MANGIFERIN (195) PUERARIN (228) VITEXIN (264)</p>
39	 <p>GITOXIN (275)</p>

40	 <p>OUABAIN (207)</p>
41	 <p>BETA SITOSTEROL 3-O-BETA D GLUCOPYRANOSIDE (59) BETA SITOSTEROL (60) STIGMASTEROL (246)</p>
42	 <p>ATALAPHILLIDINE (46) ATALAPHILLININE (47)</p>
43	 <p>ARTEANNUAN (43)</p>
44	 <p>7-O-METHYL GLABRANINE (20) GLABRANIN (152)</p>

45	 <p>6-HYDROXYCRINAMINE (18) CRINAMINE (101) HAEMANTHAMINE (159) PRETAZETTINE (218)</p>
46	 <p>CRYPTOTANSHINONE (107) TANSHINONE IIA (260)</p>
47	 <p>WITHAFERIN A (265)</p>
48	 <p>(+)-CATECHIN 3-O GALLATE (1) (-)-EPICATECHIN 3 O GALLATE (2) (-)-EPIGALLOCATECHIN GALLATE (3)</p> <p>(-)-GALLOCATECHIN 3-O GALLATE (5)</p>

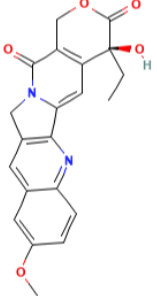
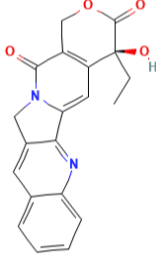
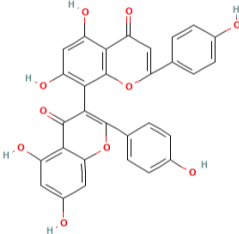
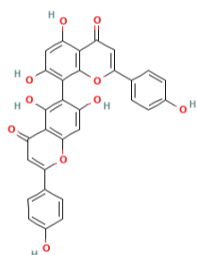
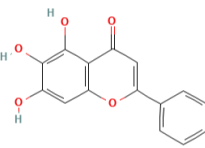
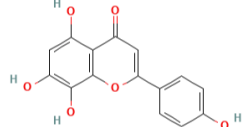
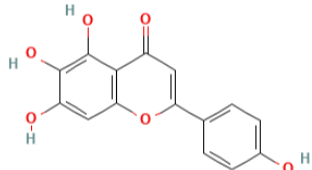
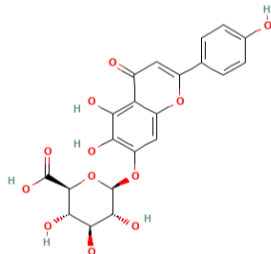
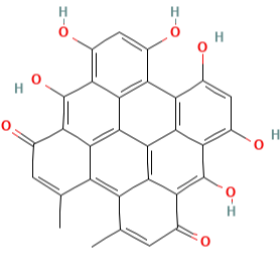
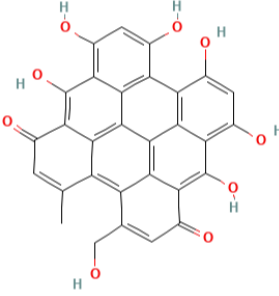
49	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>10-METHOXYCAMPTOTHECIN (7)</p> </div> <div style="text-align: center;">  <p>CAMPTOTHECIN (71)</p> </div> </div>
50	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>13',118-BIAPIGENIN (8)</p> </div> <div style="text-align: center;">  <p>AGATHISFLAVONE (23)</p> </div> <div style="text-align: center;">  <p>BAICALEIN (52)</p> </div> <div style="text-align: center;">  <p>ISOSCUTELLAREIN (178)</p> </div> </div> <div style="display: flex; justify-content: space-around; align-items: center; margin-top: 20px;"> <div style="text-align: center;">  <p>SCUTELLAREIN (243)</p> </div> <div style="text-align: center;">  <p>SCUTELLARIN (249)</p> </div> </div>
51	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>HYPERICIN (168)</p> </div> <div style="text-align: center;">  <p>PSEUDOHYPERICIN (224)</p> </div> </div>

Table S10. Drug-likeness screening results and docking results of the 18 promising leads. Bolded compounds were identified using LBVS and SBVS. Molecule design could be applied for further property improvements. * The three solubility categories are ESOL, Ali, and Silicos IT. ** Out of 5 main P450 cytochromes: CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4.

Compounds	Average solubility* (mol/L)	Number of cytochromes inhibited**	Target	PDB	Docking E-Score	Cluster
10-methoxycamptothecin	2.9E-04	4	NSP7&8	6YHU	-14.0	49
3,3'-dimethylquercetin	2.8E-05	4	NSP9	6W9Q	-18.7	9
5,4'-dihydroxy-3,7,3'-trimethoxyflavone	1.5E-05	4	NSP7&8	6XIP	-16.6	9
7-ethylcamptothecin	7.1E-05	5	NSP7&8	6XIP	-17.0	49
				6YHU	-13.4	
Acacetin	3.3E-05	4	NSP7&8	6YHU	-14.3	5
Camptothecin	4.0E-04	3	NSP7&8	6YHU	-13.5	49
			NSP7&8	6YHU	-14.4	
Columbamine	3.7E-05	3	NSP13	6ZSL	-18.6	36
			Spike RBD	6XM4	-16.2	
Coptisine	4.3E-05	2	NSP13	6ZSL	-18.5	36
			NSP7&8	6XIP	-16.8	
Dihydrochelerythrine	8.6E-06	5	NSP7&8	6YHU	-16.6	36
			NSP13	6ZSL	-18.2	
Eupatilin	2.0E-05	4	NSP7&8	6YHU	-14.6	50
Hydroxycamptothecin	4.1E-04	1	NSP7&8	6XIP	-16.6	49
				6YHU	-13.7	
Jatrorrhizine	3.7E-05	3	NSP7&8	6YHU	-13.7	36
Oroxilin A	2.7E-05	4	NSP7&8	6YHU	-15.4	50
Palmatrubine	3.7E-05	3	NSP7&8	6XIP	-17.0	36
				6YHU	-15.4	
Papaverine	1.2E-04	5	NSP7&8	6XIP	-17.0	36
Pectolarigenin	2.8E-05	4	NSP7&8	6YHU	-14.2	50
Rhein	2.8E-04	0	NSP13	7NIO	-18.2	0
Salvigenin	1.7E-05	5	NSP7&8	6YHU	-13.8	50