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], Mark Reimers[1,3], Michael H. Bachmann[1,4], Daniel Woldring[1,2,*]

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

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

[3]Department of Biomedical Engineering, Michigan State University, 775 Woodlot Dr, East Lansing, MI 48824, USA

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

# Contributed equally

[*] 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. Multiple vaccines are in use, but there are many underserved locations that do not have adequate access to them. Variants may emerge that are highly resistant to existing vaccines, and therefore cheap and readily obtainable therapeutics are needed. Phytochemicals, or plant chemicals, can possibly be such therapeutics. Phytochemicals can be used in a polypharmacological approach, where multiple viral proteins are inhibited and escape mutations are made less likely. Finding the right phytochemicals for viral protein inhibition is challenging, but *in-silico* screening methods can make this a more tractable problem. In this study, we screen a wide range of natural drug products against a comprehensive set of SARS-CoV-2 proteins using a high-resolution computational workflow. This workflow consists of a structure-based virtual screening (SBVS), where an initial phytochemical library was docked against all selected protein structures. Subsequently, ligand-based virtual screening (LBVS) was employed, where chemical features of 34 lead compounds obtained from the SBVS were used to predict 53 lead compounds from a larger phytochemical library via supervised learning. A computational docking validation of the 53 predicted leads obtained from LBVS revealed that 28 of them elicit strong binding interactions with SARS-CoV-2 proteins. Thus, the inclusion of LBVS resulted in a 4-fold increase in the lead discovery rate. Of the total 62 leads, 18 showed promising pharmacokinetic properties in a computational ADME screening. Collectively, this study demonstrates 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 six 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 and severe over time.[3] Thanks to the extraordinarily rapid development of vaccines, transmission of COVID-19 has been attenuated and its symptoms have been greatly reduced in a large fraction of global societies. Although more than 60% of the world is fully vaccinated as of July 2022,[4] large inequities remain. As the virus mutates, however, and modifies its structural and functional components, existing vaccines may become less effective.[5] 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.[6 7] Identifying chemicals with therapeutic potential against SARS-CoV-2 and related viruses would (1) provide additional protection, even for 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 all received at least an EUA (emergency use authorization) from the FDA.[8] However, mono-drug therapy, as the HIV/AIDS epidemic taught us, carries in the risk of rapid development of drug resistance.[9] Already, mutations that confer resistance to Remdesivir[10] and Paxlovid[11] have been discovered, and there is concern that other monotherapies like Molnupiravir may cause drug resistance if they are not used as part of a treatment that targets multiple viral components.[12 13] Moreover, certain Omicron subvariants have documented resistance towards sera from vaccinated individuals, demonstrating resistance against spike protein neutralizing antibodies.[14] 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. [15]

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.[16] Polypharmacology has shown remarkable results for other devastating diseases such as HIV.[17] 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 (GRAS), are an attractive resource for drug development. Although there have been concerns about the transparency of the process by which compounds are added to the GRAS list,[18] 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. 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.[19] Thus, antiviral phytochemicals offer a promising starting point for the screening and discovering of specific drugs that are effective against SARS-CoV-2.

Computational molecular docking has experienced a surge of advances in recent years largely 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 response by the scientific community, thousands of structures of the structural and non-structural protein components of SARS-CoV-2 have been generated and made publicly available.[20] 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)[21 22 23 24] as well as the RNA-dependent RNA polymerase,[25] spike protein,[26] and replicase.[27] Further work describes the potential for phytochemicals to make a positive impact on treating COVID-19 and provides evidence for benefits elicited from flavonoids,[28] polyphenols,[29] and alkaloid drugs.[30]

The advent of machine learning (ML) in drug discovery and development has also facilitated and accelerated predictive processes through the use of Bayesian models,[31] structure-based algebraic topology,[32] convolutional neural networks,[33] and transfer learning.[34] The application of ML has prevailed in various stages including target identification and validation, compound screening and lead discovery, preclinical development, and clinical development.[35]

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 screen several non-structural proteins (NSP1, the NSP3 macrodomain, NSP5, the NSP7-NSP8 complex, 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.[36] The helicase (NSP13) has an essential role in viral replication due to its function in unwinding RNA and DNA.[37] The NSP7-NSP8 heterodimer binds to NSP12 to form the core RNA-dependent RNA polymerase,[38] and the NSP3 macrodomain is responsible for hydrolyzing ADP ribose modifications made by the host cell for regulating the immune response to viral infection.[39] NSP1 is a virulence factor responsible for inhibiting host cell translation by binding to the mRNA channels of the ribosome,[40] and NSP15 is responsible for cleaving viral RNA to evade immune detection.[41] Lastly, NSP10 binds to NSP14 and NSP16 in order to activate their functions,[42] and NSP9 is an important RNA binder.[43] We used the modeling

software Rosetta 3.12 to conduct ligand docking simulations (structure-based virtual screenings or SBVS), to obtain the estimated binding free energies between our phytochemicals and protein structures. We identified lead compounds with high affinity toward individual protein structures by analyzing the distributions of the docking energy scores for each protein.

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 lead compounds from a separate, larger phytochemical library. We used unsupervised learning to cluster the screened anti-viral phytochemicals in our initial library, aiming to extract the chemical features of identified leads. We then identified lead clusters by ranking the fractions of lead phytochemicals within each cluster. Then, we employed supervised learning to classify the un-screened phytochemicals from the larger library into the established clusters. Of the new 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 screening. In our investigation, the use of machine learning significantly sped up the ligand screening process, giving rise to a 4-fold increase in lead compound yield.

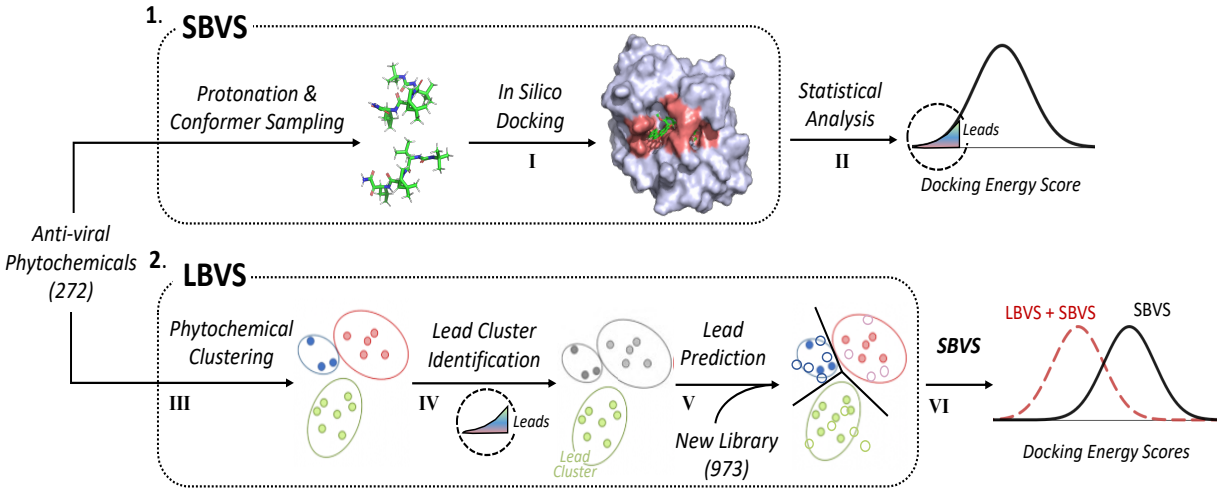

Figure 1. Overview of the structure and ligand-based virtual screening workflow. Numerous SARS-CoV-2 protein structures and 272 anti-viral phytochemicals were first prepared for the Rosetta protein-ligand docking (SBVS). After docking (I), lead phytochemicals were chosen based on the highest performing (lowest docking energy) simulations (II). The initial phytochemical library (272 compounds) was then clustered according to chemical similarity (III). Lead clusters were then identified as the clusters that had the highest proportions of lead phytochemicals (IV). The ligand-based virtual screening then classified 973 phytochemicals from a distinct database into the established clusters (the newly classified phytochemicals are represented by the hollow circles added to the clusters) (V). New phytochemicals classified as belonging to lead clusters were identified and subjected to high-resolution structural docking (VI). The ligand based virtual screening began only once the structural based virtual screening of the 272 initial phytochemicals was completed.

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 initially obtained from the USDA Phytochemical and Ethnobotanical Databases.[19] Three-dimensional structures of 272 of these phytochemicals were downloaded from the ZINC and PubChem databases in SDF format for the initial SBVS. The remaining 71 phytochemicals did not have SDF files that could be found within publicly available databases. OpenBabel,[44] a chemical file conversion and manipulation application, was then used to protonate ligand structures for a pH of 7.4, in order to better simulate *in-vivo* conditions. Ligand conformational space sampling was then performed using the BCL::Conf application.[45] 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.[46] 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.[47]

Protein Preparation for *In Silico* Docking

All SARS-CoV-2 protein structures in the apo form (Table 1, Figure 5B) were obtained from the Protein Data Bank.[20] 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 using both PyMOL[48] and a script within Rosetta named "clean_pdb.py".[47] The spike protein receptor binding domain (RBD) was manually excised from the PDB file 6XM4 using PyMOL. Lastly, the cleaned protein structures were concatenated with the ligand PDB files prior to docking.

Table 1: Protein structures obtained for docking

Protein Names(s)	PDB ID	Crystal Structure Resolution (Å)	Starting Coordinates Sampled in PDB (x,y,z)[a]	Reference
NSP1	7K3N[†]	1.65	(-8.23, 23.87, 42.82)	Ref [49]
NSP3 Macrodomain/ADP Ribose Phosphatase	6WEN[†]	1.35	(18.38, 9.09, 7.12)	Ref [50]
NSP3 Macrodomain/ADP Ribose Phosphatase	6WEY	0.95	(1.96, 17.18, -12.84)	Ref [51]
NSP5/Main Protease	6Y2E	1.75	(-13.67, -26.75, -0.99), (-29.43, -21.15, 26.23), (-13.09, -11.27, 18.19)	Ref [52]
NSP5/Main Protease	7AR5[†]	1.40	(7.82, -3.33, 24.55), (25.22, 5.79, -4.22)	Ref [53]
NSP7-NSP8 complex	6XIP[†*]	1.50	(21.72, 3.23, 6.96)	Ref [54]

NSP7-NSP8 complex	6YHU	2.00	(-26.01, 27.66, 65.30), (-35.96, 36.89, 67.50)	Ref [55]
NSP9	6WXD†	2.00	(56.77, 2.28, 21.60)	Ref [56]
NSP9	6W9Q†*	2.05	(-11.77, -7.87, -4.06), (-6.35, -25.85, -24.59)	Ref [56]
NSP10	6ZCT	2.55	(7.91, 86.52, 19.97)	Ref [57]
NSP13/Helicase	6ZSL†*	1.94	(-16.51, 26.99, -65.81), (-40.15, 37.82, -73.08), (-35.46, 17.58, -71.92), (-17.75, 5.97, -72.65)	Ref [58]
NSP13/Helicase	7NIO†*	2.20	(-25.12, 23.71, -36.28), (-13.77, 45.04, -16.14), (-25.82, -8.97, -38.29), (-47.86, 18.62, -22.20)	Ref [58]
NSP15/Endoribonuclease	6VWW†	2.20	(-78.37, 34.30, -27.01), (-77.69, 17.05, -17.30), (-71.85, 25.17, -8.45), (-65.18, 27.79, -37.56)	Ref [59]
Spike protein (closed state)	6VXX†*	2.80	(212.96, 179.67, 235.36), (237.65, 229.12, 244.34), (176.30, 229.44, 246.03)	Ref [60]
Spike RBD	6XM4b†	2.90	(176.00, 222.46, 148.56), (193.40, 229.39, 109.43)	Ref [61]

[a]These coordinates were obtained via our CASTp protocol described in "Protein-Ligand Binding Site Prediction". They are the centers of 3D regions sampled by the ligands during docking.

[b]The receptor binding domain of the spike protein was manually removed from the structure 6XM4 and then submitted for docking, since no PDB structure for the RBD in the unbound form could be found. Thus, 6XM4 was not docked as a complete structure.

†A limitation of many protein crystal structures is that a large number of them (approximately 70% and increasing as of 2015) have missing or uncertain regions in their PDB files.[62] Often, these unresolved regions are disordered N-/C- termini and highly flexible loops with large B-factors, indicating uncertainty in atom positions.[63] Many of these regions are also added sequences that help stabilize crystal structures.[64] Of our 15 protein structures, 11 (marked with †) had missing or uncertain regions in their PDB files, most of which were under 10 residues in length and were in or near high B-factor regions. Many of these regions were termini. However, we did not sample the entire protein surfaces during docking (as detailed in Protein-Ligand Binding Site Prediction). For 6 of the 11 structures that had uncertain regions, our ligand docking protocol did not sample positions near those uncertain regions.

*Protein structures marked with this asterisk had uncertain areas close to the areas sampled by Rosetta during docking (within 8 Å. Protein structures marked with † but not the asterisk had uncertain regions, but those regions were not close to the areas sampled during the dockings.

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.[65] 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.

Potential binding pocket sampling criteria were established based on the statistics of the binding pockets of protein-ligand complexes obtained from the CASF-2016 dataset. This dataset contains 285 unique, high resolution crystal structures (Figure S1, Table S1, and Table S2).[66] The CASF-2016 dataset was chosen because it had many protein complexes similar in size to the SARS-CoV-2 nonstructural proteins we docked. Our pocket sampling criteria first ranked all pockets for a particular protein structure by volume from largest to smallest, and then compared each pocket to the largest volume pocket. All pockets that had volumes at least 10% of the largest pocket volume were considered potential binding pockets. If any pocket had a volume less than 10% of the largest pocket volume, then their surface areas were compared to the surface area of the previously ranked (larger by volume) 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 less than 10% of the largest pocket 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 as having volumes less than 1000 $Å^3$, 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 in the large pockets 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.[67]

Docking Data Analysis

One thousand models (docking poses) were generated for every sampled binding pocket for each protein-ligand combination we had. Each model supplied data that described 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,[68] 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 compounds' scores for a specific protein were designated as lead candidates against that specific protein.

Rosetta Score Function Selection

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 examined a paper by Smith *et al*. which tested the following Rosetta score functions: RosettaLigand (Pre-talaris2013), Talaris2014, Ref2015, and Betanov16.[69] This study revealed that RosettaLigand outperformed all of the other Rosetta score functions in a scoring test, which measured the ability of a score function to linearly correlate computational binding affinity with experimental binding affinity. RosettaLigand also outperformed all the other Rosetta score functions in a ranking test, which assessed the score functions' abilities to rank experimental binding affinities of different compounds against the same target, and in a docking test, which assessed the score functions' abilities to distinguish actual ligand binding orientation from other poses. These analyses were conducted on the CASF-2016 dataset, a well characterized set of 285 protein-ligand complexes specially curated for the testing of score functions.[66] Overall, the authors determined that RosettaLigand performs well relative to the best score functions available. Thus, we decided to implement RosettaLigand as our score function in this study.

Phytochemical Structure Embedding

To quantitatively cluster and classify molecules, our phytochemicals were converted into numerical form. We used a circular molecular fingerprint method, extended-connectivity fingerprints (ECFPs),[70] to generate molecular descriptors that store the structural information of a given molecule. These descriptors were mapped on a 1024-bit vector, where each bit indicated 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.[70] 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.

Once all identifiers were obtained, the remainders obtained from dividing each identifier by 1024 were computed. These remainders were the vector indices 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 algorithms from Sci-Kit library[71] 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: Agglomerative Hierarchical Clustering with the Ward linkage criterion,[71 72] Spectral Clustering,[71 73] Affinity Propagation,[71 74]and Ordering Points to Identify Cluster Structure (OPTICS).[75 76] Descriptions of each of these clustering methods are available in the Supporting Information in the section titled "Ligand Clustering Methods Tested".

After molecule clustering, we used Shannon entropy (Eq 1) to measure the distribution of phytochemicals among all clusters. A high Shannon entropy indicates that a similar number of phytochemicals have been assigned to each cluster. In this way, identifying a clustering method that yields high Shannon entropy helps to avoid challenges associated with imbalanced classification. [77 78 79]

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

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.

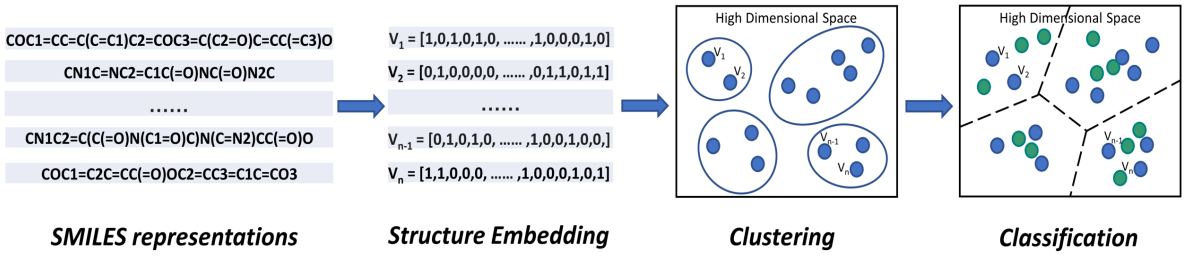

Figure 2. Phytochemical Clustering and Classification Scheme. The ECFP algorithm was used to encode molecule structural information into fixed-length vector representations. The molecule clustering is based on the distance calculations of vector representations of molecules. The un-clustered molecules (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. A classifier was then built to classify new phytochemicals that had not undergone high-resolution docking into the formed clusters using supervised learning. We designated the new phytochemicals classified into lead clusters as predicted 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.[80] In our model, k was chosen to be three, and the distance metric used was Tanimoto.[76] The weights in the weight function for points closer to neighbors were 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.[81] Radial Basis Function (RBF) was used to transform features into 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.[82] LDA is a Gaussian maximum likelihood classification method that assumes each class is under a Gaussian distribution. The estimated 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.[83]

ADME Screening and Protein-Ligand Interaction Analysis

The SwissADME webserver[84] was used on all of the lead compounds in order to test various drug-likeness properties and other important metrics like solubilities, PAINS violations, Brenk alerts, and lead-likeness. PAINS violations detect potentially promiscuous binders, and Brenk alerts identify potentially toxic and/or metabolically unstable moieties in a molecule. Lead-likeness refers to similarities a given compound has to a “lead”, or a starting point for further drug development.[84] The SwissADME drug-likeness category is based on the following 5 rules: Lipinski, Ghose, Veber, Egan, and Muegge. The total number of violations each lead compound had is tabulated in Figure 8A. The lead compounds that had a maximum of 1 cumulative violation in all the screened categories were designated as promising lead compounds.

Interactions between promising lead compounds and the proteins they were leads for were analyzed using PyMOL and UCSF Chimera.[85] Hydrogen bonding interactions were identified using both software, and all of the residues that had hydrogen bonding interactions with the leads are noted in Table 4. Hydrophobic and electrostatic interactions were visualized with UCSF ChimeraX.[86]

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.[65] 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 dataset[66]—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 (Figure 3C, Figure 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. These statistics, among others, were used to establish our pocket sampling criteria mentioned in Methods (Figure S1, Table S1, Table S2). Following this validation, we analyzed each of our 15 SARS-CoV-2 protein structures (Figure 5B) 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 and used these coordinates for the initial ligand placements during high-resolution docking (Figure 3B).

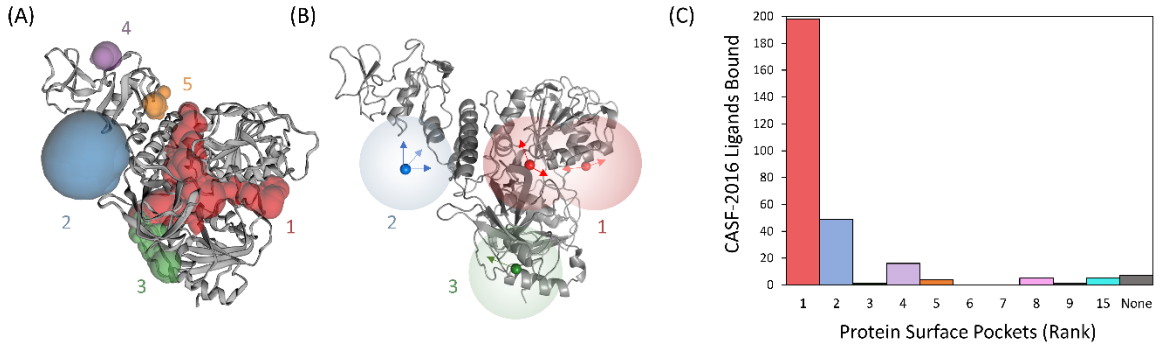

Figure 3. Binding pocket identification and ranking. (A) Concave crevices (pockets) along the protein surface are calculated using CASTp. Distinct pockets are individually colored throughout the SARS-CoV-2 helicase (PDB: 7NIO), shown here. Pockets are numbered according to pocket volume rank. (B) Central 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 $Å^3$. 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.

Each SARS-CoV-2 protein we've chosen in this study plays an essential role in the overall function of the virus. We conducted docking at a maximum of five unique binding pockets for each of these proteins (Figure S1). The bottom panel of Figure S1 shows the pockets

sampled during the docking of each structure. Many of these pockets overlapped with or were proximal to active sites and functional interfaces. For instance, the largest pocket (colored in red) depicted for both NSP3 structures (6WEN and 6WEY) coincides with where the NSP3 macrodomain binds ADP ribose.[87] This is important, since the function of the NSP3 macrodomain is to hydrolyze modified ADP ribose which is produced by the infected cell in order to regulate an immune response.[39] The largest pocket in the main protease structures (6Y2E and 7AR5) is in the location of the active site,[88] and the second largest pocket (colored in blue) in the NSP9 structure 6W9Q contains residue S59, which plays a role in the binding of both RNA and ssDNA.[89] Additionally, the largest pocket sampled in the NSP7-NSP8 structure (6XIP) is in close proximity (~10 Å) to the binding site between the NSP7-NSP8 heterodimer and NSP12,[90 91] and the largest pocket sampled in NSP10 (6ZCT) is approximately 11 Å from the binding site between NSP10 and NSP16.[92] The NSP7-NSP8 heterodimer and NSP10 are both subunits of larger protein complexes, and ligands docked near the binding sites of the subunits could potentially affect the ability of the subunits to assemble into functional oligomers via allosteric modulation.[93]

Clustering and Classification of Phytochemical Ligands. The Ward hierarchical clustering method and the Random Forest method were selected to cluster and classify phytochemicals, respectively. Because the prediction is largely determined by classifying molecules, the classification accuracy rate is a key indicator of the performance of different models. We aimed to generate a balanced distribution of cluster sizes to reduce a learning bias toward the majority class (i.e. the class imbalance problem)[77 78] and promote more efficient learning for the phytochemical classification task. While sensitivity and specificity work well for binary-classification evaluation, they are less suitable for indicating performance within multi-class classification models, such as the models used in our study. Alternatively, calculating Shannon entropy (Eq. 1) reflects how close a distribution is to uniformity and is better suited for multi-class cases.[79] Thus, we used Shannon entropy to measure phytochemical distribution uniformity among our divided classes and then favor the model that yields high Shannon entropy. Model hyperparameters were tuned with different classification methods in order to obtain the best results (Figure 4A). 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 4B). The color of each data point in Figure 4B indicates its cluster. The molecule points colored in black were 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 were RF and KNN, respectively. When the number of clusters was increased, the accuracy rate decreased, while the Shannon entropy increased 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 the number of clusters was increased 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 since a large portion of molecules were 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 values are updated.[71] The overall accuracy of this model was not as good as the accuracy of the Ward hierarchical method. Damping factors in the range [0.59, 0.79) 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 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 user. 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.[75] The KNN and RF classification methods generated a higher accuracy than SVM and LDA. When MinPts was increased 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, this density-based clustering method caused many molecules to be categorized as noise.

Comparing the different methods, we concluded that Ward hierarchical clustering with Random Forest classification produced the best results 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 sizes indicated by its color distribution (Figure 4B). The details of the molecule clustering results are in Table S8 of the Supporting Information.

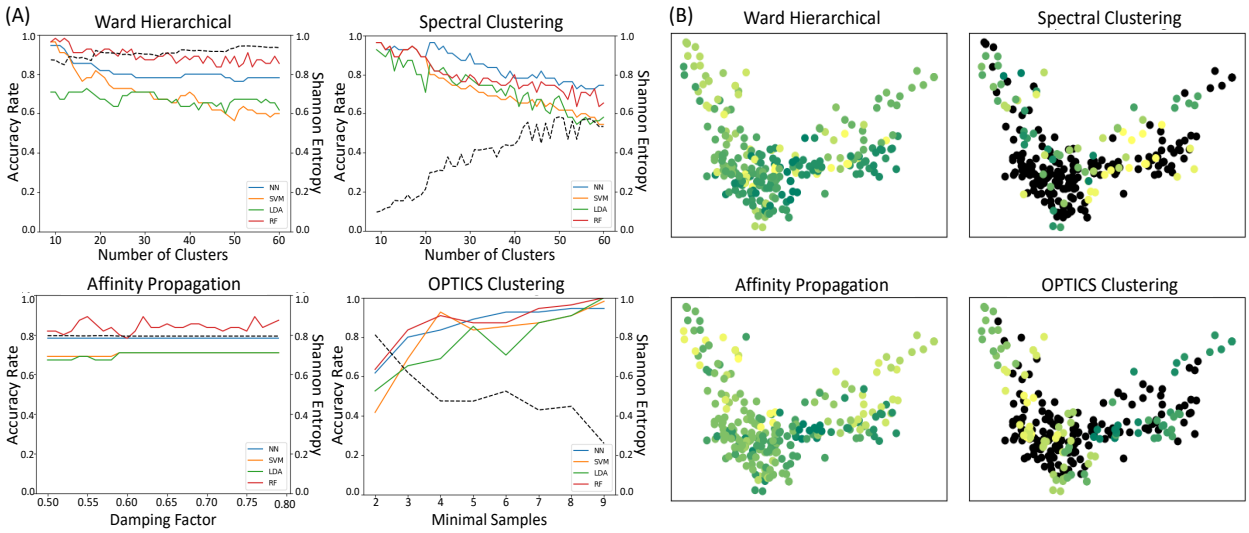

Figure 4. Comparison of phytochemical clustering and classification models. (A) Model hyperparameter tuning combined with classification methods. The left side of the y-axis indicates accuracy (colored solid lines) and the right side of the y-axis indicates Shannon entropy (black dashed line --) (B) 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(s) were unable to group them.

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. We chose to use z-scores because our docking scores had approximately normal distributions, like we hypothesized initially. The z-scores indicate the number of standard deviations from the sample means. In this study, the sample means were the averages of all lowest energy scores for the dockings of the initial 272 anti-viral phytochemicals (in the SBVS) against specific protein structures. In the heatmap of z-scores (Figure 5A 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 equal to or weaker than the average. Using a z-score of -2 as the threshold to identify lead candidates, we identified 34 lead compounds from the initial 272 anti-viral phytochemicals. (Table S3) 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 the helicase (7NIO), main protease (7AR5), and NSP15 (6VWW), and hypericin demonstrates strong binding to the main protease (6Y2E), NSP9 (6WXD), and the spike protein (6VXX). Other molecular docking studies have shown that our main protease leads agathisflavone,[94] amentoflavone,[95] ginkgetin,[96] procyanidin b2,[97] bilobetin,[96] and hypericin[98] are good binders against the main protease. Moreover, we found that other molecular docking studies mention some of the same interacting residues that we observed for our main protease leads hypericin,[99] ginkgetin,[100] bilobetin,[100] and procyanidin b2.[97] For example, Zhu and Xie found that procyanidin b2 forms hydrogen bonding interactions with GLY 143 and GLU 166,[97] and Dey et al. found that bilobetin forms hydrogen bonding interactions with GLU 166 and HIS 163.[100] We observed these same interactions among others.

The dendrogram graph shows the hierarchical orders of formed clusters (Figure 6A right). Clusters 5 and 50 are closely related and have a large dark area in the heatmap. Other noticeable patches of dark areas were observed for clusters such as 5, 36, and 51, which indicate that many of their constituent phytochemicals bind strongly to more than one SARS-CoV-2 protein structure. There is a risk that the multi-target behavior exhibited by some of these compounds may indicate molecular promiscuity. We used the PAINS detector in SwissADME on all of our leads to check for molecular promiscuity and filter out any compounds that performed poorly (Figure 8A); Notably, cluster 5 consists mostly of flavones (Figure 6), which are a subgroup of flavonoids. Many flavonoids have been documented to be promiscuous molecules;[101] however, after we conducted the SwissADME screening, none of the flavones that passed our filtering criteria were found to have any PAINS violations. Hypericin and pseudohypericin (the only compounds in cluster 51) had strong binding affinities to several targets, but they also had PAINS and other violations. Thus, they were not considered among the 18 promising compounds which passed our SwissADME filtering criteria.

The number of lead phytochemicals within each cluster was counted for each protein structure (Figure 6) in order to link cluster specificity to different SARS-CoV-2 structures. We identified the following clusters as lead clusters for our viral proteins (Table 2).

Table 2: Lead clusters and the proteins that they contained leads for.

Cluster Number	Proteins Targeted by Cluster
Cluster 5	NSP1, NSP3, NSP5, NSP7&8, NSP9, NSP13, NSP15, Spike protein, Spike RBD
Cluster 7	NSP10
Cluster 30	NSP1
Cluster 36	NSP3, NSP7&8, NSP13, Spike RBD
Cluster 42	NSP5, Spike RBD
Cluster 49	NSP7&8
Cluster 50	NSP1, NSP7&8, NSP13, NSP15, Spike RBD
Cluster 51	NSP3, NSP5, NSP9, NSP13, NSP15, Spike protein

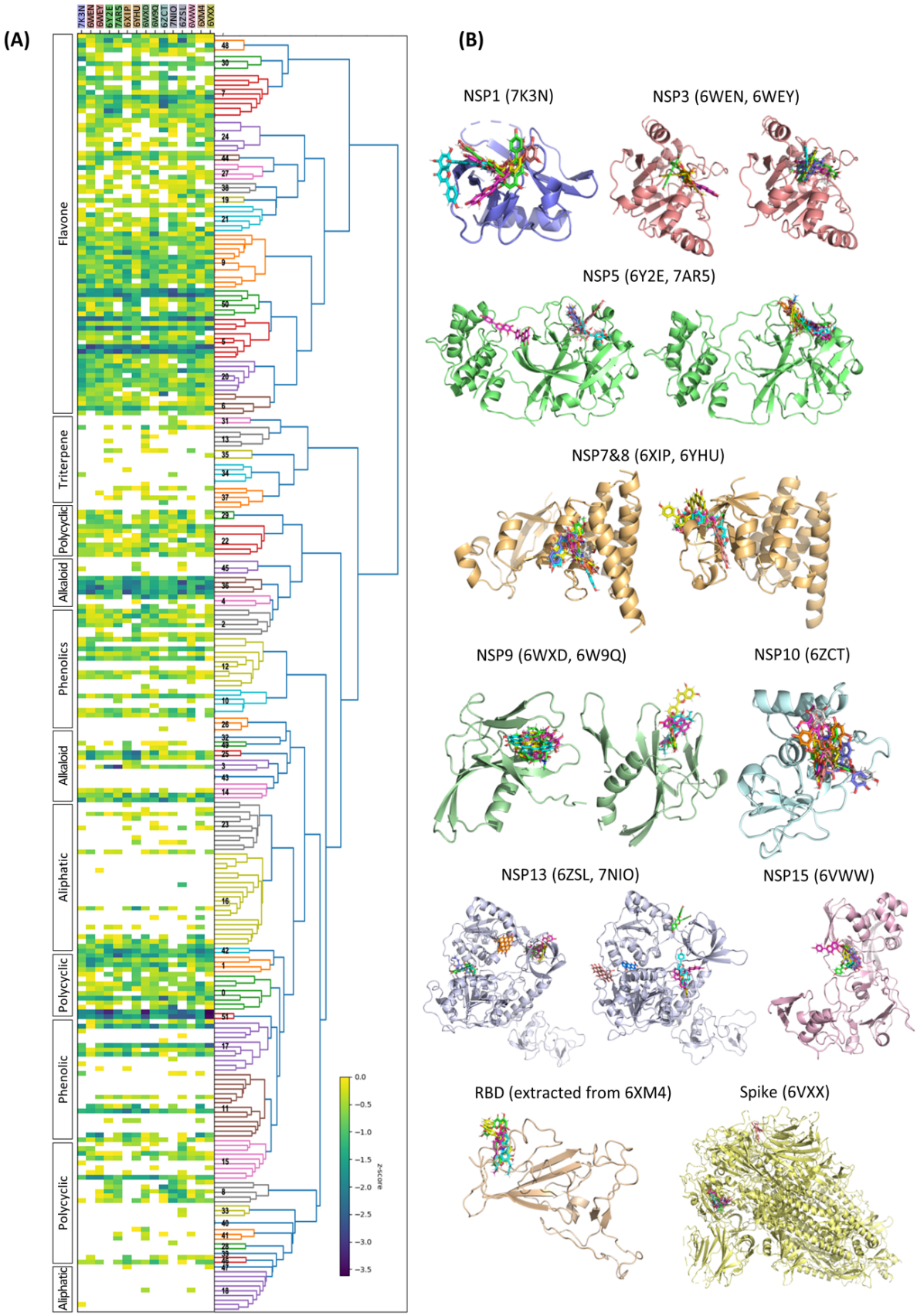
(A)
7K3N
6WEN
6WEY
6Y2E
7AR5
6XIP
6YHU
6WXD
6W9Q
6ZCT
7NIO
6ZSL
6VWW
6XM4
6VXX
Flavone
Triterpene
Polycyclic
Alkaloid
Phenolics
Alkaloid
Aliphatic
Polycyclic
Phenolic
Polycyclic
Aliphatic
z-score
(B)
NSP1 (7K3N)
NSP3 (6WEN, 6WEY)
NSP5 (6Y2E, 7AR5)
NSP7&8 (6XIP, 6YHU)
NSP9 (6WXD, 6W9Q)
NSP10 (6ZCT)
NSP13 (6ZSL, 7NIO)
NSP15 (6VWW)
RBD (extracted from 6XM4)
Spike (6VXX)

Figure 5. (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 S8 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 and can be found by following instructions in the Data and Software Availability section).

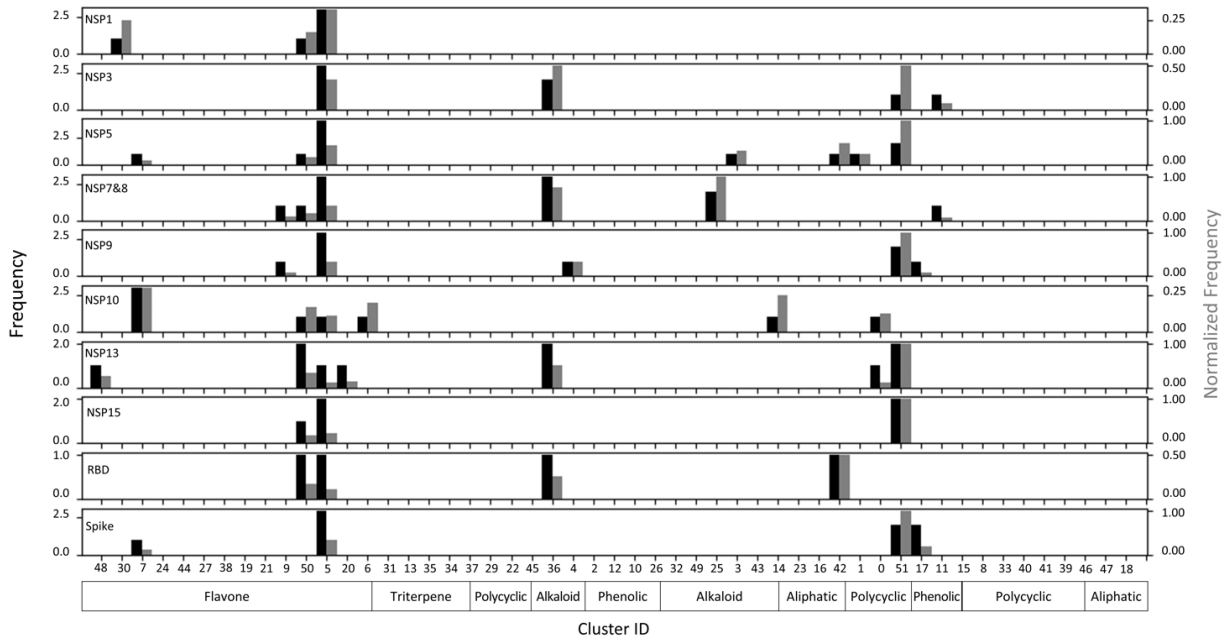

Figure 6. Frequency (black bars) and normalized frequency (gray bars) of identified leads within each molecule cluster. Molecule cluster IDs are given on the x axis, and the approximate chemical classes which most phytochemicals within a cluster belong to are also specified on the x axis.

Evaluation of LBVS Model. The inclusion of our ligand based virtual screening (LBVS) increased the rate of lead identification from 2.18% (SBVS only) to 16.44% (SBVS + LBVS). The 973 new phytochemicals from the larger phytochemical library were classified into 52 formed clusters, and 53 of those compounds were classified into lead clusters (shown in Table 2). 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 the 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 S5). Compared to the z-scores of the initial dockings of the 272 anti-viral phytochemicals, we introduced a negative distributional shift of z-scores (Figure 7A). 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 7B). 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 S4.

A random under-sampling confusion matrix was constructed to measure the performance of our classification (prediction) model (Table S6). 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.

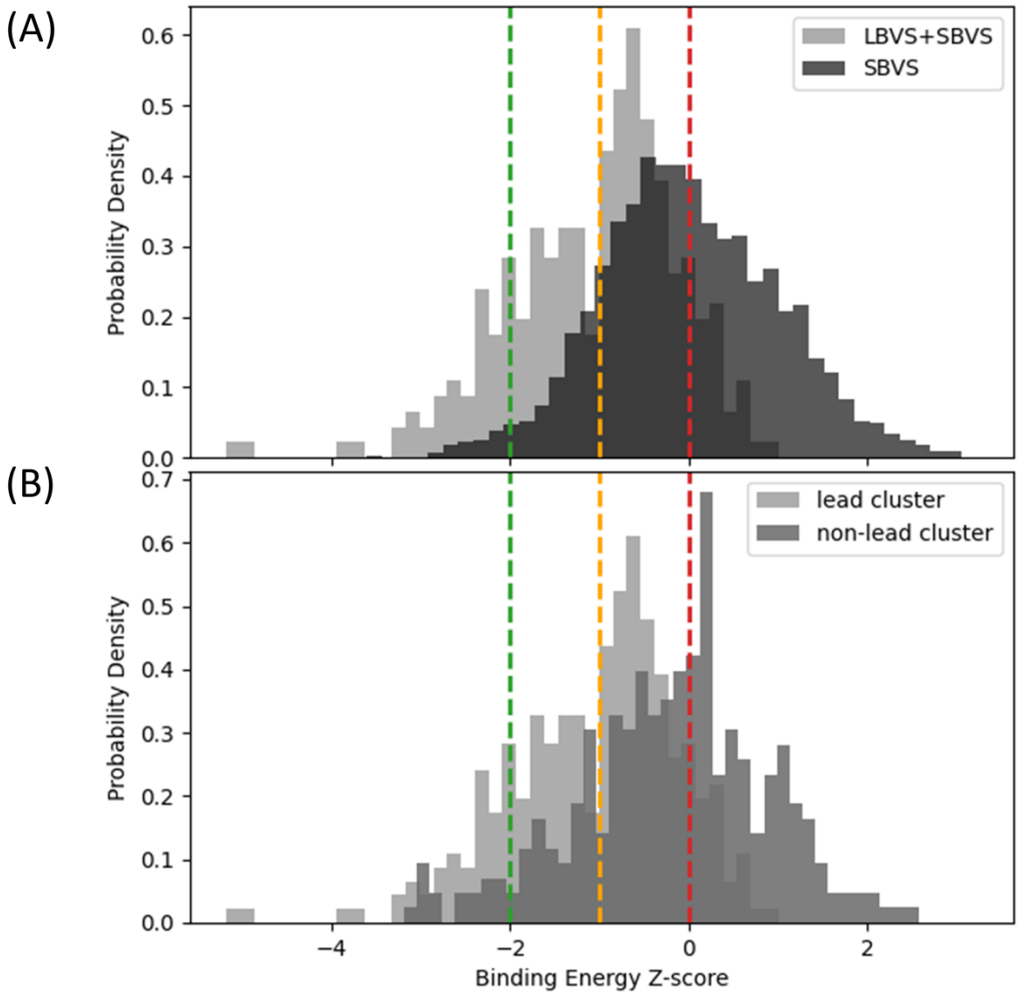

Figure 7. (A) Distribution of docking energy z-scores generated via SBVS alone (dark gray) and SBVS with the inclusion of LBVS (light gray). Overlap is shown in the darkest gray (B) Distribution of docking energy z-scores of phytochemicals classified to lead clusters (light gray) and those classified to non-lead clusters (dark gray) via LBVS. Overlap is shown in the darkest gray. The red, yellow, and green dashed lines label z scores of 0, -1, and -2 respectively.

ADME Screening for all Identified Lead Phytochemicals. We used SwissADME to obtain certain drug property parameters for the 62 lead phytochemicals identified through both the initial SBVS and those identified through LBVS and SBVS combined (Figure 8A).[84] Eighteen compounds (Table 3, Figure S3, Table S9) showed promising results with a maximum of 1 cumulative violation in the following categories: drug-likeness, PAINS, Brenk, and lead-likeness. This threshold was based on the fact that Doravirine, a small molecule drug approved by the FDA in 2018 for the treatment of HIV, had 1 cumulative violation (Table S7).[102]

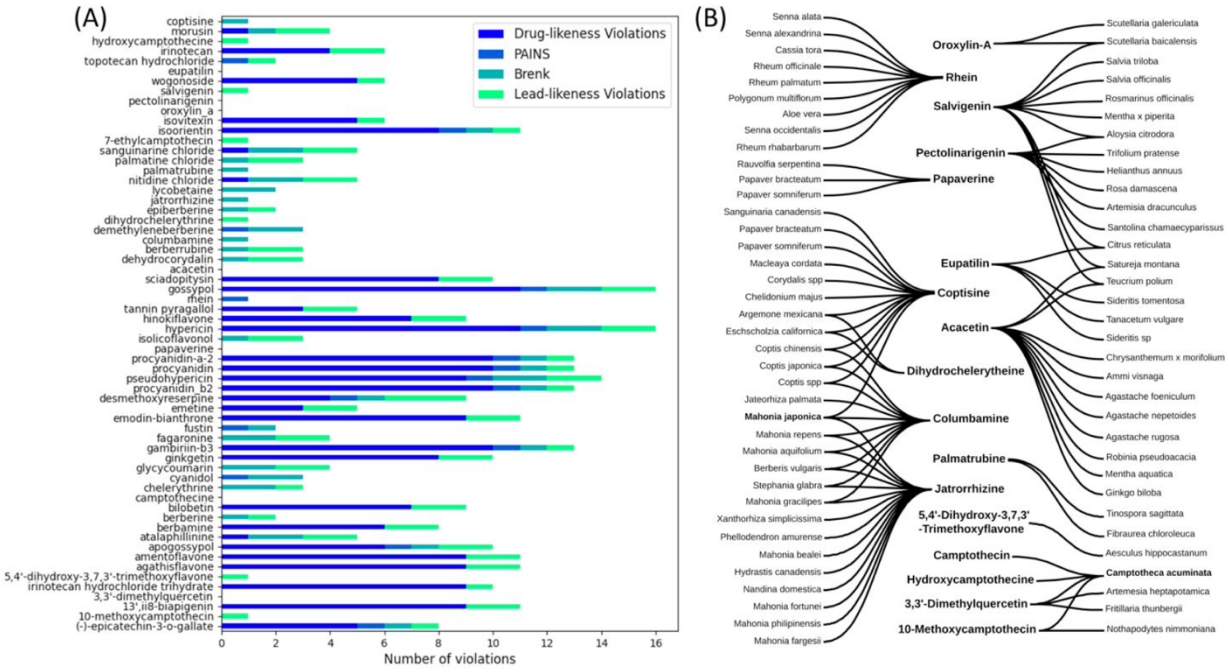

Figure 8. (A) The cumulative violations of each lead molecule in the drug-likeness, PAINS, lead-likeness, and Brenk categories. (B) Plant sources for 17 promising phytochemicals identified through the drug-likeness screening (no plant sources could be found for 7-ethylcamptothecin, so it isn't present in the diagram). 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 3: 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	0.03
	5,4'-dihydroxy-3,7,3'-trimethoxyflavone	9		4	0.02
	Eupatilin	50		4	0.02
	Oroxylin A			4	0.03
	Pectolinarigenin			4	0.03
	Salvigenin			5	0.02
NSP9	3,3'-dimethylquercetin	9	Flavone	4	0.03
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 of the 18 promising phytochemicals like dihydrochelerythrine (alkaloid with antimicrobial and anticancer properties)[103 104] were poorly soluble compared to the other leads despite having either 0 or 1 total violations in all categories. Poor solubility can impact biodistribution and bioavailability, preventing a drug from adequately reaching the tissues where it is most needed.[105] Many of the promising leads were also 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, since they are responsible for the metabolism of many drugs.[84] Additionally, 11 of the 18 promising leads were identified as P-glycoprotein 1 substrates, which may impact their efficacy since that protein is responsible for the excretion of certain compounds out of cells.[106] Of the promising compounds, Rhein stood out since it was classified as either soluble or moderately soluble in all the SwissADME solubility categories, and it was not identified as an inhibitor of any of the aforementioned cytochrome isoforms (no other compounds met this last criterion). Rhein was not identified as a substrate of P-glycoprotein 1, but it did produce one PAINS violation, indicating possible promiscuity in binding. Notably, none of the other promising compounds produced any PAINS violations.

Of the promising leads, rhein and camptothecin were compared with 3 COVID-19 antiviral medicines (Remdesivir, Molnupiravir, and Paxlovid) and the HIV drug Doravirine in another SwissADME screening (Table S7). The comparison indicated that rhein is more soluble, has a higher bioavailability score, has better gastrointestinal (GI) absorption, and has fewer drug likeness violations than Remdesivir (which has 11 drug-likeness violations). Rhein also has greater solubilities than Paxlovid in all of SwissADME's solubility categories. Molnupiravir is more soluble than both camptothecin or rhein in all categories, but its GI absorbance is listed as "low", and it has 3 drug-likeness violations and 1 Brenk alert (indicating possible toxicity and/or metabolic instability). Overall, both rhein and camptothecin are classified as either soluble or moderately soluble in all categories, and they both have fewer total drug-likeness violations, PAINS violations, Brenk alerts, and lead-likeness violations than each of the 3 COVID-19 drugs analyzed (Table S7). This information indicates that rhein and camptothecin may have good biodistribution and comparable pharmacokinetics to well-performing drugs like Doravirine,[107] but future laboratory work is needed to make any definitive analyses. Our assessment is in agreement with reports of the therapeutic potential of rhein[108 109] and camptothecin.[110]

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.* 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.

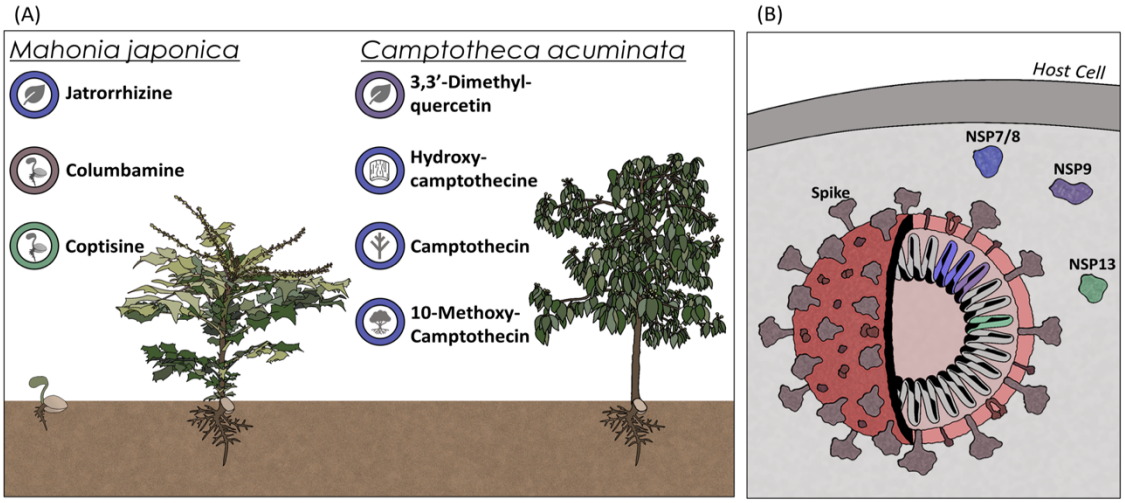

Figure 9. (A) The two plants that contain 3 (*Mahonia japonica*) and 4 (*Camptotheca acuminata*) of the 18 phytochemicals identified as promising in the SwissADME screening. The part of the plant most abundant in a specific phytochemical (leaf, sprout sapling, bark, stem, whole plant) are shown in the icons to the left of the compound names.[111 112 113 114] (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.

Protein-Ligand Interaction Analysis of Promising Leads

All of the 18 compounds that showed promising results in the SwissADME screening were analyzed with PyMOL, UCSF Chimera, and UCSF ChimeraX in order to determine the intermolecular interactions they had with the protein structures they were leads for. Specifically, the best scoring (lowest binding energy) complex for each protein-ligand pair was selected for the analysis.

Table 4: Hydrogen bonding promotes ligand binding but is not required for high affinity interactions. Binding energy (Rosetta energy units; REU) for the 18 promising leads and their respective hydrogen bonding partners are shown.

Compound	Protein(s) targeted	Target PDB ID	Number of h-bonds	H-bonding residues	Docking Score (REU)
Acacetin	NSP7&8	6YHU	2	V66, G64	-14.3
Columbamine	NSP7&8	6YHU	2	Q63	-14.4
	NSP13/Helicase	6ZSL†	2	D463, R165	-18.6
	Spike RBD	6XM4	0	N/A	-16.8
Oroxylin-A	NSP7&8	6YHU	2	V66, G64	-15.4
Salvigenin	NSP7&8	6YHU	1	Q63	-13.8
Pectolinarigenin	NSP7&8	6YHU	2	V66, G64	-14.2
Papaverine	NSP7&8	6XIP†	4	R91	-17.0
Eupatilin	NSP7&8	6YHU	1	V66	-14.6
Coptisine	NSP13/Helicase	6ZSL†	3	S466, R165, K186	-18.5
Dihydrochelerythrine	NSP7&8	6XIP†	3	R91, Q83	-16.8
		6YHU	1	G64	-16.6
	NSP13/Helicase	6ZSL†	2	R165, S465	-18.2
Palmatrubine	NSP7&8	6XIP†	0	N/A	-17.0
		6YHU	2	V66, G64	-15.4
7-ethylcamptothecin	NSP7&8	6XIP†	1	N71	-17.0
		6YHU	0	N/A	-13.4
Camptothecin	NSP7&8	6YHU	2	V167, Q158	-13.5
Hydroxycamptothecin	NSP7&8	6XIP†	4	R106, R91	-16.6
		6YHU	1	G64	-13.7
3,3'-dimethylquercetin	NSP9	6W9Q	4	T67, S59, R39	-18.7
10-methoxycamptothecin	NSP7&8	6YHU	1	Q158	-14.0
Jatrorrhizine	NSP7&8	6YHU	0	N/A	-13.7
Rhein	NSP13/Helicase	7NIO	9	Q401, R564, R440, S286, K285	-18.2
5,4'-dihydroxy-3,7,3'-trimethoxyflavone	NSP7&8	6XIP†	1	R91	-16.6

†These protein structures had regions of structural uncertainty within their PDB structures that were within 8 Å of the pocket bound by a promising lead. Specifically, a partially resolved C-terminus (6XIP) and flexible loop region (6ZSL) are proximal to binding pockets.

The interaction analysis revealed that V66 and G64 were common hydrogen bonding residues present in complexes with 6YHU, and R91 was a common hydrogen bonding residue in complexes with 6XIP. Additionally, R165 was a residue that participated in hydrogen bonding in all of the 6ZSL structures. Although 6XIP and 6YHU were both NSP7-NSP8 structures, they came from different crystal structures of differing global symmetry (oligomer assembly) and resolution, and they had different CASTp-identified potential binding pockets. Oligomerization is known to cause structural changes in protein monomers, potentially affecting pocket structure.[115] Thus, the residues near the ligand positions ended up being different. Interestingly, the best scoring positions of jatrorrhizine against 6YHU, columbamine against the spike RBD, palmatrubine against 6XIP, and 7-ethylcamptothecin against 6YHU lacked any intermolecular hydrogen bonding interactions. This indicates that other binding interactions (e.g., hydrophobic interactions) may play an important role in those complexes. Pi stacking and pi-cation interactions were not observed in any of the complexes analyzed. While a greater number of hydrogen bonds didn't correlate well with a better binding score in Rosetta (R^2 = 0.227), all of the complexes in Table 4 that had binding energy scores below -18 REU had at least 2 hydrogen bonding interactions with neighboring residues. Notably, Rhein had a binding energy score of -18.2 REU and the most hydrogen bonding interactions of any of the promising leads, with a total of 9 interactions (Table 4, Figure 11B). For further analysis, we examined the following compounds from the 18 promising leads: the ligand with the best docking score (3,3'-dimethylquercetin), the ligand with the most hydrogen bonding interactions with its protein target (rhein), and the ligand with the best docking score out of the ligands that had no hydrogen bonding interactions with their protein targets (palmatrubine).

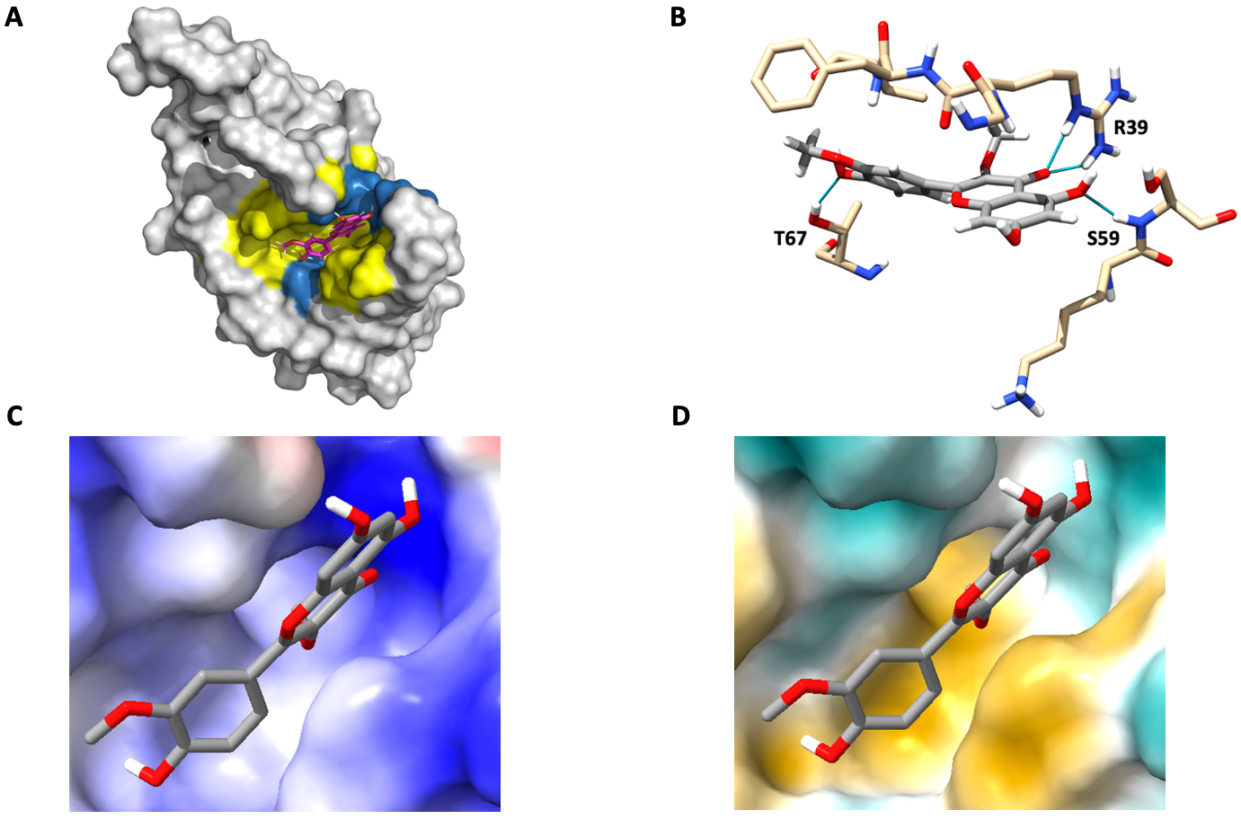

Figure 10. (A) Surface map of the best scoring docked pose of 3,3'-dimethylquercetin against NSP9 (6W9Q). The residues that have hydrogen bonding interactions with 3,3'-dimethylquercetin are shown in blue and the remaining residues within 5 Å of 3,3'-dimethylquercetin are shown in yellow. (B) The 3 residues that have hydrogen bonding interactions with 3,3'-dimethylquercetin are shown and named. Hydrogen bonds are shown as turquoise lines. (C) Surface map showing electrostatic properties of residues neighboring 3,3'-dimethylquercetin. Residues colored in blue have a positive electrostatic potential, with darker blue indicating greater positivity. Red residues have a negative electrostatic potential, with darker red indicating greater negativity. (D) Hydrophilicity surface map of residues neighboring 3,3'-dimethylquercetin, where cyan indicates greater hydrophilicity and yellow indicates greater hydrophobicity. The darker the color, the more pronounced the hydrophilicity or hydrophobicity.

3,3'-dimethylquercetin had the strongest binding affinity out of the 18 promising leads, with a docking score of -18.7 REU, and it had 4 hydrogen bonding interactions with the residues T67, R39, and S59 (Table 4, Figure 10B). 3,3'-dimethylquercetin likely had multiple favorable electrostatic interactions in its binding site, given that most surrounding residues had a positive electrostatic potential as indicated by the blue coloring, and these residues were near ligand oxygens with partial negative charges (Figure 10C). Some residues colored in pale red (indicating a weak negative potential) may have contributed to stabilizing the docked pose since they were near ligand hydrogens in hydroxyl groups that had a partial positive charge. Hydrophobic interactions also may have played a role in stabilizing the docked pose, since many of the surrounding residues had fairly hydrophobic side chains, as indicated by the dark yellow coloring (Figure 10D).

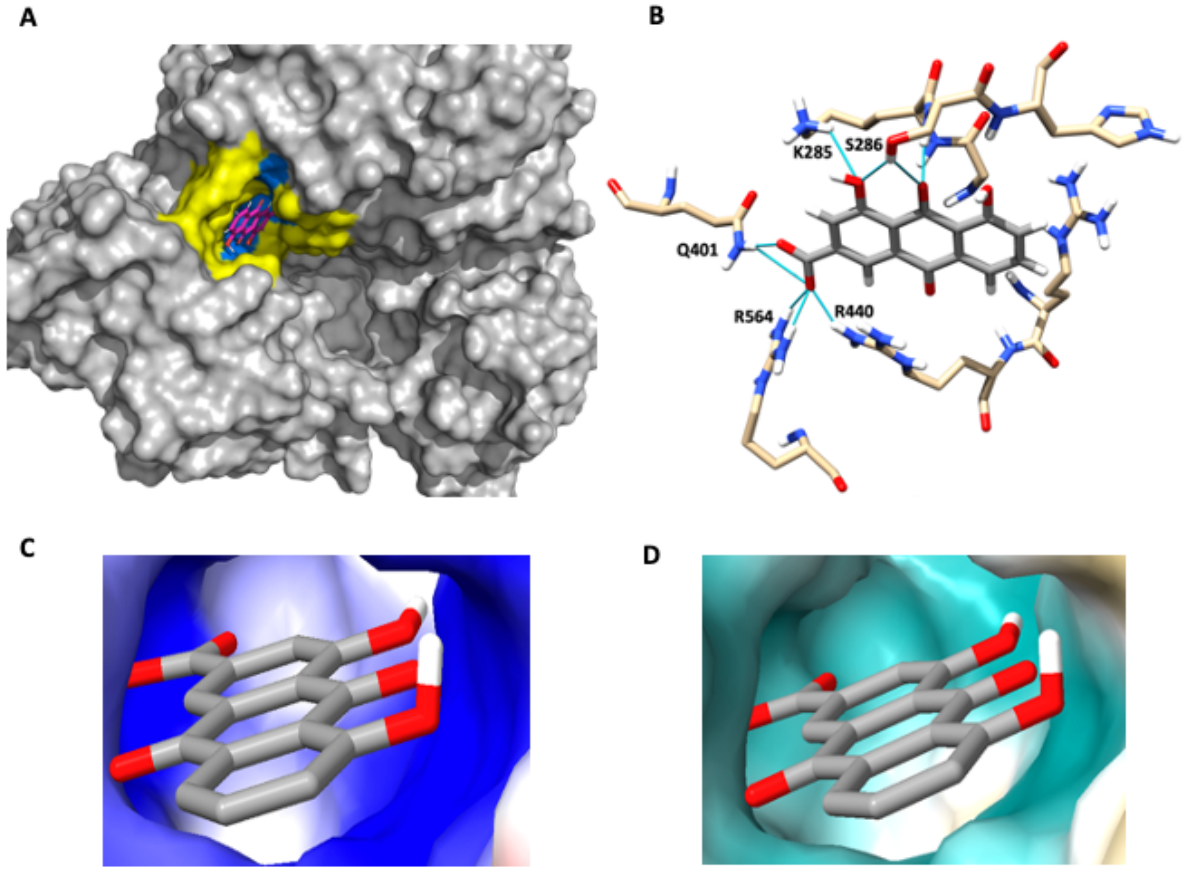

Figure 11. (A) The best scoring docked pose of rhein against the Helicase/NSP13 (7NIO). Rhein is shown in pink, residues with which rhein forms hydrogen bonding interactions are colored in blue, and all other residues within 5 Å of rhein are colored in yellow. (B) The 5 residues that have hydrogen bonding interactions (shown as turquoise lines) with rhein are shown and named. (C) Surface map showing electrostatic properties of residues neighboring rhein. Residues colored in blue have a positive electrostatic potential, with darker blue indicating greater positivity. White residues have a neutral electrostatic potential. (D) Hydrophilicity surface map of residues neighboring rhein, where cyan coloring indicates greater hydrophilicity and yellow coloring indicates greater hydrophobicity.

Rhein had a total of 9 hydrogen bonding interactions with neighboring residues, which was the most out of any of the promising leads. However, rhein was the only promising lead to have a PAINS violation shown by the SwissADME screening, indicating possible promiscuity in binding. A large number of hydrogen bond donors and acceptors in a molecule can lead to greater chances of hydrogen bonding, and this logically seems like it might cause promiscuity. A high density of hydrogen bonding groups in a molecule has been previously shown to indicate promiscuity in rhodanines and thiohydantoins,[116] but it has also indicated selectivity in HIV-1 protease inhibitors when they had more than 7 hydrogen bonding groups.[117] Thus, it is difficult to state anything about the promiscuity of rhein based on its hydrogen bonding properties alone. The electrostatic potential map demonstrates that rhein likely had many favorable electrostatic interactions in its binding site since the surrounding residues had almost exclusively positive potentials, while rhein has 6 oxygens with partial negative charges that were part of carbonyl, hydroxyl, and carboxylate groups (Figure 11C). The hydrophilicity surface map indicates that hydrophobic interactions likely didn't predominate in the docking since most of the neighboring residues were hydrophilic, as indicated by the cyan coloring (Figure 11D). Overall, hydrogen bonding and other electrostatic interactions likely contributed more than hydrophobic interactions to rhein's binding energy of -18.2 REU.

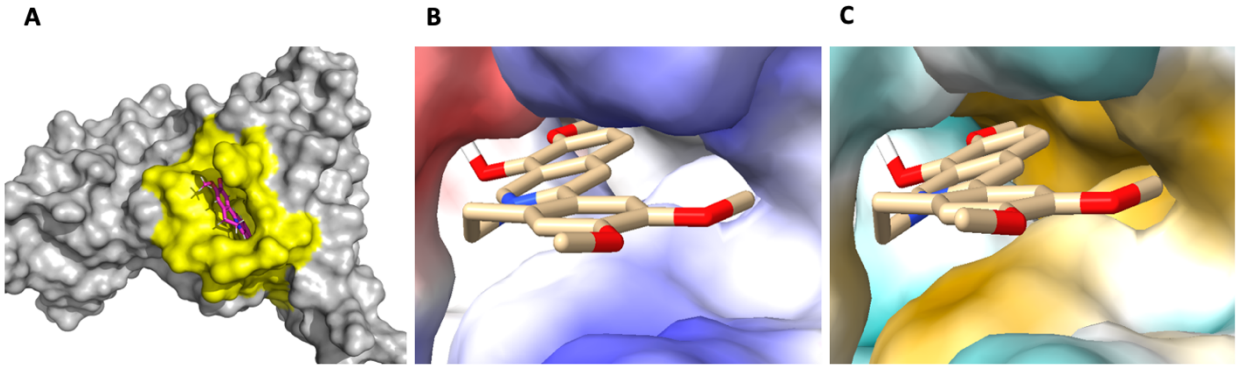

Figure 12: (A) Surface map of the best scoring docked pose of palmatrubine against NSP7&8 (6XIP), showing all residues within 5 Å of palmatrubine in yellow. No hydrogen bonding interactions were observed (B) Surface map showing electrostatic properties of residues neighboring palmatrubine. Residues colored in blue have a positive electrostatic potential, with darker blue indicating greater positivity. Red residues have a negative electrostatic potential, with darker red indicating greater

electronegativity. (C) Hydrophilicity surface map of residues neighboring palmatrubine, where cyan indicates greater hydrophilicity and yellow indicates greater hydrophobicity.

Palmatrubine docked to NSP7&8 (6XIP) had the best docking score out of all the promising leads that didn't have any identified hydrogen bonding interactions. Its docking score of -17 REU was better than 17 of the 25 best docking scores of the 18 promising leads (Table 4). Electrostatic interactions other than hydrogen bonds likely played a role in stabilizing palmatrubine's docked pose, as indicated by the dark blue residues in its vicinity, which could have interacted favorably with the negatively charged oxygens in the hydroxyl and methoxy groups of palmatrubine (Figure 12B). The partial positive charge on palmatrubine's hydroxyl hydrogen may have interacted favorably with the neighboring residues colored in red. Additionally, palmatrubine had a positive charge on its nitrogen (colored in blue), which could have formed a favorable interaction with the residues colored in red. Since palmatrubine was surrounded by many hydrophobic residues, it likely experienced some stabilizing hydrophobic interactions. Many hydrophobic residues were colored a dark yellow, so the interactions they had with palmatrubine could have been stronger that the hydrophobic interactions rhein had with residues in 7NIO (Figure 11D), which were more distant and colored with a paler yellow. Thus, hydrophobic interactions likely contributed more significantly to palmatrubine's docking score than they did to Rhein's docking score.

Overall, rhein and 3,3'-dimethylquercetin stood out from the rest of the promising leads because they both had binding scores below -18 REU and relatively large numbers of hydrogen bonding interactions. However, this discussion demonstrates that compounds like palmatrubine also deserve closer analysis since they can yield promising docking results despite not having any detected hydrogen bonding interactions with their target.

Inhibitory Activity of Certain Leads. Among the 18 promising leads generated by our computational study, a subset of these compounds have been previously reported to have inhibitory activity against COVID-19 based on *in vitro* or *in vivo* assays. Specifically, pectolinarigenin was found to inhibit SARS-CoV-2 replication in Vero cells with an IC_{50} of 12.4 μg/mL.[118] Of the remaining leads compounds that had more than 1 cumulative violation in SwissADME, emetine inhibited SARS-CoV-2 replication in Vero cells with an EC_{50} of 0.147 nM,[119] and hypericin resulted in 84% viral inhibition *in vitro* at a concentration of 10 μM.[98] Additionally, amentoflavone (a lead compound against the SARS-CoV-2 main protease in our study) had an *in vitro* IC_{50} of 8.3 μM against the SARS-CoV main protease,[120] which may be unsurprising given the high structural similarity (RMSD = 0.86Å) and conserved active sites between the viral proteases.[121]

CONCLUSIONS

In this project, we introduce a machine learning-assisted ligand docking workflow to expedite the discovery of lead compounds. We apply this novel approach in the context of efficiently sampling large libraries of naturally abundant phytochemicals to treat SARS-CoV-2 in a polypharmacological manner. Within our workflow, we implement 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 (both structural and non-structural proteins). The initial structure-based virtual screening 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 main purpose of this project is not to make any definitive claims about any compounds screened, but rather to introduce a computational workflow that helps expedite the discovery of 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[118 122] and computational[123] 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 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); protein targets, energy scores, and clusters of lead candidates identified with SBVS (Table S3); descriptions of the 4 clustering methods we tested; protein targets, energy scores, and clusters of lead candidates identified with LBVS (Table S4); percentage of docking energy scores within certain ranges relative to the mean after LBVS was applied (Table S5); random under-sampling confusion matrix used to evaluate LBVS model (Table S6); drug-likeness data comparison between promising leads and the antivirals Doravirine, Paxlovid, Molnupiravir, and Remdesivir (Table S7); docking script used in Rosetta (Figure S2); structures of the 18 lead compounds identified as promising in the SwissADME screening (Figure S3); phytochemicals used in SBVS named, labeled numerically,

and organized into their clusters (Table S8); SwissADME results of the 18 lead compounds identified as promising in the drug-likeness screening (Table S9).

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.

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". 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). Matplotlib was also used for plotting data. R and the statistics module in Python were used for calculating sample means, covariances, correlations, and all other statistical parameters. 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 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/.

Conflicts of Interest

The authors declare no competing interests.

REFERENCES

1. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) and Johns Hopkins University (JHU). (2022).
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. Bernstein, A. S. *et al.* The costs and benefits of primary prevention of zoonotic pandemics. *Science Advances* **8**, (2022).
4. Holder, J. Tracking Coronavirus Vaccinations Around the World. *The New York Times* (2022).
5. 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* **6**, 203 (2021).
6. 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* **27**, 2127–2135 (2021).
7. Liu, L. *et al.* Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).
8. Burki, T. K. The role of antiviral treatment in the COVID-19 pandemic. *The Lancet Respiratory Medicine* **10**, e18 (2022).
9. JOHNSON, V. A. Combination Therapy: More Effective Control of HIV Type 1? *AIDS Research and Human Retroviruses* **10**, 907–912 (1994).
10. Gandhi, S. *et al.* De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an immunocompromised patient: a case report. *Nature Communications* **13**, 1547 (2022).
11. Jochmans $, D. *et al.* The substitutions L50F, E166A and L167F in SARS-CoV-2 3CLpro are selected by a protease inhibitor in vitro and confer resistance to nirmatrelvir. doi:10.1101/2022.06.07.495116.
12. Kozlov, M. Why scientists are racing to develop more COVID antivirals. *Nature* **601**, 496–496 (2022).
13. Sacco, M. D. *et al.* The P132H mutation in the main protease of Omicron SARS-CoV-2 decreases thermal stability without compromising catalysis or small-molecule drug inhibition. *Cell Research* **32**, 498–500 (2022).
14. Ho, D. *et al.* SARS-CoV-2 Omicron BA.2.12.1, BA.4, and BA.5 subvariants evolved to extend antibody evasion. doi:10.21203/rs.3.rs-1696532/v1.
15. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* **4**, 682–690 (2008).

16. Patil, R. *et al.* Computational and network pharmacology analysis of bioflavonoids as possible natural antiviral compounds in COVID-19. *Informatics in Medicine Unlocked* **22**, 100504 (2021).

17. Muhammad, J. *et al.* Network Pharmacology: Exploring the Resources and Methodologies. *Current Topics in Medicinal Chemistry* **18**, 949–964 (2018).

18. Richey Levine, A., Picoraro, J. A., Dorfzaun, S. & LeLeiko, N. S. Emulsifiers and Intestinal Health: An Introduction. *Journal of Pediatric Gastroenterology & Nutrition* **74**, 314–319 (2022).

19. Dr. Duke's Phytochemical and Ethnobotanical Databases. *US Department of Agriculture*.

20. Berman, H. M. *et al. The Protein Data Bank*. *Nucleic Acids Research* vol. 28 http://www.rcsb.org/pdb/status.html (2000).

21. 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**, 942 (2020).

22. 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* **42**, 744–769 (2022).

23. 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**, e0240653 (2020).

24. 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).

25. 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).

26. 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**, 17699 (2020).

27. 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).

28. Ngwa, W. *et al.* Potential of Flavonoid-Inspired Phytomedicines against COVID-19. *Molecules* **25**, 2707 (2020).

29. Wu, Y. *et al.* Polyphenols as Potential Inhibitors of SARS-CoV-2 RNA Dependent RNA Polymerase (RdRp). *Molecules* **26**, 7438 (2021).

30. 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**, 129489 (2021).

31. Gawriljuk, V. O. *et al.* Machine Learning Models Identify Inhibitors of SARS-CoV-2. *Journal of Chemical Information and Modeling* **61**, 4224–4235 (2021).

32. 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).

33. 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**, 104317 (2021).

34. 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).

35. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery* **18**, 463–477 (2019).

36. Yadav, R. *et al.* Role of Structural and Non-Structural Proteins and Therapeutic Targets of SARS-CoV-2 for COVID-19. *Cells* **10**, 821 (2021).

37. Newman, J. A. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature Communications* **12**, 4848 (2021).

38. Mariano, G., Farthing, R. J., Lale-Farjat, S. L. M. & Bergeron, J. R. C. Structural Characterization of SARS-CoV-2: Where We Are, and Where We Need to Be. *Frontiers in Molecular Biosciences* **7**, (2020).

39. Russo, L. C. *et al.* The SARS-CoV-2 Nsp3 macrodomain reverses PARP9/DTX3L-dependent ADP-ribosylation induced by interferon signaling. *Journal of Biological Chemistry* **297**, 101041 (2021).

40. Schubert, K. *et al.* SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nature Structural & Molecular Biology* **27**, 959–966 (2020).

41. Frazier, M. N. *et al.* Characterization of SARS2 Nsp15 nuclease activity reveals it's mad about U. *Nucleic Acids Research* **49**, 10136–10149 (2021).

42. Lin, S. *et al.* Crystal structure of SARS-CoV-2 nsp10 bound to nsp14-ExoN domain reveals an exoribonuclease with both structural and functional integrity. *Nucleic Acids Research* **49**, 5382–5392 (2021).

43. de O. Araújo, J. *et al.* Structural, energetic and lipophilic analysis of SARS-CoV-2 non-structural protein 9 (NSP9). *Scientific Reports* **11**, 23003 (2021).

44. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33 (2011).

45. Kothiwale, S., Mendenhall, J. L. & Meiler, J. BCL::Conf: small molecule conformational sampling using a knowledge based rotamer library. *Journal of Cheminformatics* **7**, 47 (2015).

46. 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).

47. DeLuca, S. How to Prepare Ligands for use in Rosetta. *RosettaCommons.org*.

48. Yuan, S., Chan, H. C. S. & Hu, Z. Using PyMOL as a platform for computational drug design. *WIREs Computational Molecular Science* **7**, (2017).

49. Semper, C., Watanabe, N. & Savchenko, A. Structural characterization of nonstructural protein 1 from SARS-CoV-2. *iScience* **24**, 101903 (2021).

50. Michalska, K. *et al.* Crystal structures of SARS-CoV-2 ADP-ribose phosphatase: from the apo form to ligand complexes. *IUCrJ* **7**, 814–824 (2020).

51. Frick, D. N., Virdi, R. S., Vuksanovic, N., Dahal, N. & Silvaggi, N. R. Molecular Basis for ADP-Ribose Binding to the Mac1 Domain of SARS-CoV-2 nsp3. *Biochemistry* **59**, 2608–2615 (2020).

52. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science (1979)* **368**, 409–412 (2020).

53. Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science (1979)* **372**, 642–646 (2021).

54. Wilamowski, M. *et al.* Transient and stabilized complexes of Nsp7, Nsp8, and Nsp12 in SARS-CoV-2 replication. *Biophysical Journal* **120**, 3152–3165 (2021).

55. Konkolova, E., Klima, M., Nencka, R. & Boura, E. Structural analysis of the putative SARS-CoV-2 primase complex. *Journal of Structural Biology* **211**, 107548 (2020).

56. Littler, D. R., Gully, B. S., Colson, R. N. & Rossjohn, J. Crystal Structure of the SARS-CoV-2 Non-structural Protein 9, Nsp9. *iScience* **23**, 101258 (2020).

57. Rogstam, A. *et al.* Crystal Structure of Non-Structural Protein 10 from Severe Acute Respiratory Syndrome Coronavirus-2. *International Journal of Molecular Sciences* **21**, 7375 (2020).

58. Newman, J. A. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature Communications* **12**, 4848 (2021).

59. Kim, Y. *et al.* Crystal structure of Nsp15 endoribonuclease <scp>NendoU</scp> from <scp>SARS-CoV</scp> -2. *Protein Science* **29**, 1596–1605 (2020).

60. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

61. Zhou, T. *et al.* Cryo-EM Structures of SARS-CoV-2 Spike without and with ACE2 Reveal a pH-Dependent Switch to Mediate Endosomal Positioning of Receptor-Binding Domains. *Cell Host & Microbe* **28**, 867-879.e5 (2020).

62. van Beusekom, B., Joosten, K., Hekkelman, M. L., Joosten, R. P. & Perrakis, A. Homology-based loop modeling yields more complete crystallographic protein structures. *IUCrJ* **5**, 585–594 (2018).

63. Carugo, O. How large B-factors can be in protein crystal structures. *BMC Bioinformatics* **19**, (2018).

64. Carugo, O. Participation of protein sequence termini in crystal contacts. *Protein Science* **20**, 2121–2124 (2011).

65. 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).

66. Su, M. *et al.* Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **59**, 895–913 (2019).

67. Lemmon, G. & Meiler, J. Rosetta Ligand Docking with Flexible XML Protocols. in 143–155 (2012). doi:10.1007/978-1-61779-465-0_10.

68. Kaufmann, K. W. & Meiler, J. Using RosettaLigand for Small Molecule Docking into Comparative Models. *PLoS ONE* **7**, e50769 (2012).

69. Smith, S. T. & Meiler, J. Assessing multiple score functions in Rosetta for drug discovery. *PLOS ONE* **15**, e0240450 (2020).

70. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010).

71. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011).

72. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview, <scp>II</scp>. *WIREs Data Mining and Knowledge Discovery* **7**, (2017).

73. von Luxburg, U. A Tutorial on Spectral Clustering. (2007).

74. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science (1979)* **315**, 972–976 (2007).

75. Ankerst, M., Breunig, M. M., Kriegel, H.-P. & Sander, J. OPTICS. in *Proceedings of the 1999 ACM SIGMOD international conference on Management of data - SIGMOD '99* 49–60 (ACM Press, 1999). doi:10.1145/304182.304187.

76. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, 20 (2015).

77. Li, D.-C., Liu, C.-W. & Hu, S. C. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine* **40**, 509–518 (2010).

78. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* **106**, 249–259 (2018).

79. Delgado, R. & Núñez-González, J. D. Enhancing Confusion Entropy (CEN) for binary and multiclass classification. *PLOS ONE* **14**, (2019).

80. Peterson, L. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).

81. Noble, W. S. What is a support vector machine? *Nature Biotechnology* **24**, 1565–1567 (2006).

82. 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).

83. Izenman, A. J. Linear Discriminant Analysis. in 237–280 (2013). doi:10.1007/978-0-387-78189-1_8.

84. 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**, 42717 (2017).

85. Pettersen, E. F. *et al.* UCSF Chimera: A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

86. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science* **30**, 70–82 (2021).

87. Alhammad, Y. M. O. *et al.* The SARS-CoV-2 Conserved Macrodomain Is a Mono-ADP-Ribosylhydrolase. *Journal of Virology* **95**, (2021).

88. Sardanelli, A. M., Isgrò, C. & Palese, L. L. SARS-CoV-2 Main Protease Active Site Ligands in the Human Metabolome. *Molecules* **26**, 1409 (2021).

89. El-Kamand, S. *et al.* A distinct ssDNA/RNA binding interface in the Nsp9 protein from SARS-CoV 2. *Proteins: Structure, Function, and Bioinformatics* **90**, 176–185 (2022).

90. Kirchdoerfer, R. N. & Ward, A. B. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications* **10**, 2342 (2019).

91. Naydenova, K. *et al.* Structure of the SARS-CoV-2 RNA-dependent RNA polymerase in the presence of favipiravir-RTP. *Proceedings of the National Academy of Sciences* **118**, (2021).

92. Krafcikova, P., Silhan, J., Nencka, R. & Boura, E. Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nature Communications* **11**, 3717 (2020).

93. Quinlan, R. J. & Reinhart, G. D. Effects of Protein-Ligand Associations on the Subunit Interactions of Phosphofructokinase from B. stearothermophilus. *Biochemistry* **45**, 11333–11341 (2006).

94. Rameshkumar, M. R. *et al.* Computational selection of flavonoid compounds as inhibitors against SARS-CoV-2 main protease, RNA-dependent RNA polymerase and spike proteins: A molecular docking study. *Saudi Journal of Biological Sciences* **28**, 448–458 (2021).

95. Imran, M. *et al.* In silico screening, SAR and kinetic studies of naturally occurring flavonoids against SARS CoV-2 main protease. *Arabian Journal of Chemistry* **15**, (2022).

96. Ghosh, R., Chakraborty, A., Biswas, A. & Chowdhuri, S. Computer aided identification of potential SARS CoV-2 main protease inhibitors from diterpenoids and biflavonoids of Torreya nucifera leaves. *Journal of Biomolecular Structure and Dynamics* **40**, 2647–2662 (2020).

97. Zhu, Y. & Xie, D.-Y. Docking Characterization and in vitro Inhibitory Activity of Flavan-3-ols and Dimeric Proanthocyanidins Against the Main Protease Activity of SARS-Cov-2. *Frontiers in Plant Science* **11**, (2020).

98. Matos, A. da R. *et al.* Identification of Hypericin as a Candidate Repurposed Therapeutic Agent for COVID-19 and Its Potential Anti-SARS-CoV-2 Activity. *Frontiers in Microbiology* **13**, (2022).

99. Pitsillou, E. *et al.* Interaction of small molecules with the SARS-CoV-2 main protease in silico and in vitro validation of potential lead compounds using an enzyme-linked immunosorbent assay. *Computational Biology and Chemistry* **89**, 107408 (2020).

100. Dey, D. *et al.* Amentoflavone derivatives significantly act towards the main protease (3CLPRO/MPRO) of SARS-CoV-2: in silico admet profiling, molecular docking, molecular dynamics simulation, network pharmacology. *Molecular Diversity* (2022) doi:10.1007/s11030-022-10459-9.

101. Gilberg, E., Gütschow, M. & Bajorath, J. Promiscuous Ligands from Experimentally Determined Structures, Binding Conformations, and Protein Family-Dependent Interaction Hotspots. *ACS Omega* **4**, 1729–1737 (2019).

102. Deeks, E. D. Doravirine: First Global Approval. *Drugs* **78**, 1643–1650 (2018).

103. Casciaro, B. *et al.* Naturally-Occurring Alkaloids of Plant Origin as Potential Antimicrobials against Antibiotic-Resistant Infections. *Molecules* **25**, 3619 (2020).

104. Okagu, I. U., Ndefo, J. C., Aham, E. C. & Udenigwe, Chibuike. C. Zanthoxylum Species: A Review of Traditional Uses, Phytochemistry and Pharmacology in Relation to Cancer, Infectious Diseases and Sickle Cell Anemia. *Frontiers in Pharmacology* **12**, (2021).

105. Alexis, F., Pridgen, E., Molnar, L. K. & Farokhzad, O. C. Factors Affecting the Clearance and Biodistribution of Polymeric Nanoparticles. *Molecular Pharmaceutics* **5**, 505–515 (2008).

106. Siddiqui, S. *et al.* Virtual screening of phytoconstituents from miracle herb nigella sativa targeting nucleocapsid protein and papain-like protease of SARS-CoV-2 for COVID-19 treatment. *Journal of Biomolecular Structure and Dynamics* **40**, (2022).

107. Afify, M. A. *et al.* Efficacy and safety of doravirine in treatment-naive HIV-1-infected adults: a systematic review and meta-analysis. *Environmental Science and Pollution Research* **28**, 10576–10588 (2021).

108. Zannella, C. *et al.* Regulation of m6A Methylation as a New Therapeutic Option against COVID-19. *Pharmaceuticals* **14**, 1135 (2021).

109. 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).

110. 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**, (2021).

111. 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* 315–316 (1982).

112. Willaman, J. & Li, H. Alkaloid-bearing Plants and Their Contained Alkaloids. *Journal of Pharmaceutical Sciences* (1971).

113. 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.

114. 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).

115. Marcos, E., Crehuet, R. & Bahar, I. Changes in Dynamics upon Oligomerization Regulate Substrate Binding and Allostery in Amino Acid Kinase Family Members. *PLoS Computational Biology* **7**, e1002201 (2011).

116. Mendgen, T., Steuer, C. & Klein, C. D. Privileged Scaffolds or Promiscuous Binders: A Comparative Study on Rhodanines and Related Heterocycles in Medicinal Chemistry. *Journal of Medicinal Chemistry* **55**, 743–753 (2012).

117. Shen, Y., Radhakrishnan, M. L. & Tidor, B. Molecular mechanisms and design principles for promiscuous inhibitors to avoid drug resistance: Lessons learned from HIV-1 protease inhibition. *Proteins: Structure, Function, and Bioinformatics* **83**, 351–372 (2015).

118. Al-Karmalawy, A. A. *et al.* Naturally Available Flavonoid Aglycones as Potential Antiviral Drug Candidates against SARS-CoV-2. *Molecules* **26**, 6559 (2021).

119. Kumar, R. *et al.* Emetine suppresses SARS-CoV-2 replication by inhibiting interaction of viral mRNA with eIF4E. *Antiviral Research* **189**, 105056 (2021).

120. Ryu, Y. B. *et al.* Biflavonoids from Torreya nucifera displaying SARS-CoV 3CLpro inhibition. *Bioorganic & Medicinal Chemistry* **18**, 7940–7947 (2010).

121. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters* **30**, 127377 (2020).

122. 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* **85**, 153401 (2021).

123. 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).