Virtual Screening on Indonesian Herbal Compounds as COVID-19 Supportive Therapy: Machine Learning and Pharmacophore Modelling Approaches

Linda Erlina[1,2], Rafika Indah Paramita[1,2*], Wisnu Ananta Kusuma[3,4*], Fadilah Fadilah[1,2], Aryo Tedjo[1,2], Irandi Putra Pratomo[2,5], Nabila Sekar Ramadhanti[3], Ahmad Kamal Nasution[3], Fadhlal Khaliq Surado[3], Aries Fitriawan[3], Khaerunissa Anbar Istiadi[2], Arry Yanuar[6]

1. Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia. Jalan Salemba Raya number 4, DKI Jakarta - 10430, Indonesia
2. Bioinformatics Core Facilities - IMERI, Faculty of Medicine, Universitas Indonesia. Jalan Salemba Raya number 6, DKI Jakarta - 10430, Indonesia
3. Department of Computer Science, Faculty of Mathematics and Natural Science, IPB University. Jalan Meranti Wing 20 level 5 Kampus IPB, Bogor, West Java - 16680, Indonesia
4. Tropical Biopharmaca Research Center, Institute of Research and Community Empowerment, IPB University. Jalan Taman Kencana number 3, Bogor, West Java-16128, Indonesia
5. Department of Pulmonology and Respiratory Medicine, Faculty of Medicine, Universitas Indonesia – Universitas Indonesia Hospital, Depok, West Java – 16424, Indonesia
6. Biomedical Computational and Drug Design Laboratory, Faculty of Pharmacy, Universitas Indonesia. Kampus Baru UI Depok, West Java – 16424, Indonesia

*Correspondence authors: Rafika Indah Paramita (rafikaindah@ui.ac.id) and Wisnu Ananta Kusuma (ananta@apps.ipb.ac.id)

Abstract

Background: The latest development of COVID-19 spread in Indonesia has reached 311,176 cases, with 11,374 patients died, updated on October 6, 2020. Unfortunately, these numbers continue to overgrow, and no drug has yet been approved for effective treatment. This study aims to determine the potential candidate compounds in Indonesian herbal medicine as a COVID-19 supportive therapy using a machine learning and pharmacophore modelling approach.

Methods: For the machine learning approach, we used three classification methods that have different ways in decision making, such as Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Random Forest (RF). Moreover, for the pharmacophore modelling approach, we performed a structure-based method on the 3D structure of the main protease SARS-CoV-2 (3CLPro) and used the SARS, MERS, and SARS-CoV-2 repurposing drugs known from the literature as data sets on the ligand-based method. Finally, we used molecular docking to analyse the interactions between the 3CLpro protein (main protease) and 14 hit compounds from the Indonesian Herbal Database (HerbalDB) and Lopinavir as a positive control.

Results: The machine learning approach with SVM, RF, and MLP methods and pharmacophore modelling approach were used for screening in herbal compounds obtained from HerbalDB. Based on the screening on HerbalDB using these two prediction approaches, we got 14 hit compounds. We then performed molecular docking to determine the interaction of these compounds with the main protease SARS-CoV-2 as an inhibiting agent. From the molecular docking analysis, it was found that six potential compounds as the main proteases of the SARS-CoV-2 inhibitor, i.e. Hesperidin, Kaempferol-3,4'-di-O-methyl ether (Ermanin); Myricetin-3-glucoside, Peonidine 3-(4'-arabinosylglucoside); Quercetin 3-(2G-rhamnosylrutinoside); and Rhamnetin 3-mannosyl-(1-2)-alloside.

Conclusions: We used layered virtual screening with machine learning and pharmacophore modelling approaches that could provide more objective and optimal virtual screening and avoid subjective decision making on research results. Herbal compounds from various plants have potential as antiviral candidates for SARS-CoV-2. Based on our research and literature study, one of Indonesia's potential commodity crops is *Psidium guajava* (guava), and people can use it directly as a preventive effort.

Keywords: COVID-19, Machine Learning, Pharmacophore Modelling, Molecular Docking, Indonesian Herbal Compounds, 3CLPro, SARS-CoV-2

Background

The new coronavirus, called SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), was first identified in Wuhan, China, in December 2019 [1]. SARS-CoV-2 belongs to the Coronaviridae family, a single-stranded RNA virus (+ ssRNA) that is widespread among humans and other mammals, causing a wide range of infections from common cold symptoms to fatal illnesses, such as severe respiratory syndrome [2,3]. The latest development of COVID-19 spread in Indonesia has reached 311,176 cases, with 11,374 patients died, updated on October 6, 2020 (data taken fromhttps://www.worldometers.info/coronavirus/country/indonesia/). Unfortunately, the infected numbers continue to overgrow, and there are no drugs approved yet as an effective treatment. Therefore, the need to discover and develop drugs to treat the Coronavirus Disease 2019 (COVID-19) is urgent.

There are two categories of anti-coronavirus therapy depending on the target, one act on the human immune system or human cells, and the other one is on the coronavirus itself. In terms of the human immune system, the innate immune system response plays an essential role in controlling the replication and infection of coronavirus and to enhance the immune response [4]. Blocking the signalling pathways of human cells required for virus replication may exhibit a specific antiviral effect. The therapies that work on the coronavirus itself include preventing the synthesis of viral RNA by acting on the genetic material of the virus, inhibiting virus replication through acting on critical enzymes of the virus, and blocking the virus from binding to human cell receptors or inhibiting the viral assembly process by working on several structural proteins [5].

Exploring new medicines for emerging and rapidly spreading diseases such as SARS-CoV-2 could be carried out through drug repurposing strategy to bypass the pre-clinical steps that usually require laborious works and resources [6]. Drug repurposing is conducted by finding new efficacy of registered drug compounds. Drug repurposing can be typically performed by analysing the interaction of compounds drugs with proteins related to the diseases (Drug-Target Interaction or DTI), then predicting new DTIs in which the interactions are previously unknown [7,8]. Drug repurposing is commonly done on conventional medicines. However, in Indonesia, in which people are more familiar with using herbal to care their health in daily life, we also need to consider developing agents from herbal, which could be utilised by the people.

To support the drug repurposing strategy and help reduce the time and cost of laboratory experiments, we used a virtual screening as one method of computer-aided drug design [9]. The virtual screening process usually works by identifying the ability of structures to bind each other, for instance, the drug compound and its protein targets. Virtual screening is usually based on compound similarity or database docking [10]. However, cheminformatic studies found that computer science approaches, such as pharmacophore analysis [9] and some machine learning techniques are useful in identifying the interaction between drug and its protein targets [10–12],

Fitriawan *et al*. [10] has developed a deep learning classification model for *Nicotinamide Adenine Dinucleotide* (NAD) protein target problem and used PubChem fingerprints as a feature. Meanwhile, Dhanda *et al*. [11] used a combination of hybrid fingerprint models to develop a Support Vector Machine (SVM) prediction for drugs compound. The research from Liu [12] used different approaches for combining the classifier, called ensemble machine learning. Johnson and Maggiora [13] analysed chemical compounds similarity that compounds with similar structures tend to have similar properties. Utilising this concept, adding a machine learning method could improve the performance in finding the drug compounds.

In this study, we research to find potential candidate compounds in Indonesian plants as anti-SARS-CoV-2 with the primary objective for prevention and possible for curation by using big data analysis and machine learning and compared with the pharmacophore modelling approach. The compounds and proteins, as the overlapping results between the machine learning approach and pharmacophore modelling approach, were validated using molecular docking. The results of this study produced several potential compound candidates that could be used as preventive purposes because the candidate plants (especially commodity crops) could be easily used directly by the community.

Methods

In this study, we combined two approaches of screening, by machine learning and pharmacophore modelling. The compounds that overlap from two approaches were further analysed using molecular docking. The graphical method in this study is represented in Fig. 1.

Machine Learning Approach

There are four steps in DTI prediction using machine learning approach. This process started with a literature study to collect drugs and protein targets interaction from public research notes and public domain database as the training dataset. The chemical structure features and genomic sequence features were then extracted from drugs and protein targets collected in the previous step. Next, the training datasets were tuned to get the hyperparameters which used later to generate the models. The last step was to utilise the predictive models to make predictions for herbal compounds data set. The machine learning approach was conducted on Intel (R) Xeon (R) Silver 4110 CPU @ 2.10GHz; 65.58 GB memory. All data and source codes of the machine learning approach used in this research can be accessed at https://github.com/TropBRC-BioinfoLab/virtual-screening-covid19.

Data Acquisition

The original datasets used in this study that consisted of drugs and protein targets were obtained from Li and Clercq [6] and Wu et al. [5] in 2020. There are 81 virus-based drugs (Additional File 1), 17 human-based drugs (Additional File 2), 15 host-based proteins and eight virus-based proteins (Table 1). Wu et al. [5] systematically analysed proteins encoded by the SARS-CoV-2 gene, compared them to the target proteins from other coronaviruses, and predicted their structure using homology modelling. Also, Li and Clercq [6] investigated the potential for reusing antiviral agents based on the therapeutic experience with two infections caused by other coronaviruses. The antiviral drugs' potential in [6] and [5] were determined by a significant binding affinity score on drug-target interaction. To extend the exploration of drug-target interactions, we input protein targets and drugs into SuperTarget web resources [13]. The outputs of SuperTarget were not only the interactions between drugs and protein targets but also the new protein targets and new drugs (Table 2) that were not previously mentioned in [5] and [6]. The total number of data obtained from literature and SuperTarget is 119 drugs, 335 protein targets, and 685 interactions (Additional file 3). Moreover, the total possible interaction that might exist is 119 drugs*335 targets = 39,865 interactions. Thus, the total dataset is 39,865 samples that consist of 685 samples with positive interactions and 39,180 samples with unknown interactions (negative).

As described before, this study aims to find the potential compound in Indonesian plants as anti-SARS-CoV-2 with the primary objective of prevention. Thus, we collected 400 Indonesia herbal compounds obtained from HerbalDB [13] as a testing dataset. This dataset has no label. Our proposed model would predict the labels positive or negative.

Drug-Target Representation

In DTI prediction, the input data required the numerical representations of compounds and proteins on the classification model. The compound descriptors are the Simplified Molecular-Input Line-Entry System (SMILES). By using SMILES, the fingerprint of a chemical structure can be obtained to represent compounds effectively. Fingerprint (FP) is the encoding of a compound into a Boolean FP vector representing the existence of a substructure within the compound's molecule. PubChem [14] issued 881 structural keys. This structural key is used as a compound similarity measure for similar compounds searching on their website http://pubchem.ncbi.nlm.nih.gov.

PubChem fingerprint was chosen because it contained the ability to explain more characteristics of a compound. Moreover, the PubChem fingerprint consisted of 881 0/1 features. It meant this characterization only needs one bit of storage for every feature in a compound while used other kinds of features that at least using float might need up to 32 bit for one feature. This small size of fingerprint helps to accelerate the machine learning process. PubChem fingerprint uses a substructure key-based on the 2D structure of a compound that is also used for similarity search [15], the same as the purpose of this paper, which finds a similar herbal compound from existing compound-protein interaction. Another research about database fingerprint (DFP), which includes PubChem fingerprint, stated that DFP is enough for compound data sets representation [16].

The simplest of protein descriptors is amino acid composition. There are 20 components, each of which is represented using a single letter code. However, the weakness of amino acid composition descriptors is that the same amino acid composition may correspond to diverse sequences as sequence order is lost [17]. The Dipeptide Composition (DC) can cover the sequence order information. Thus, this study used Dipeptide Composition (DC) as a protein descriptor. Dipeptides are combinations of 2 amino acid components (such as AA, AR, AN, AD, AC). DC converts protein sequences into 400 features. DC can be defined by (1).

$$X_{dep(i)} = \frac{n_{dep(i)}}{N} \quad (1)$$

Where *dep(i)* is the *i*-th dipeptide of 400 dipeptides, $X_{dep(i)}$ represents the ratio of occurrences of *dep(i)*, $n_{dep(i)}$ is the number of occurrences of *dep(i)*, and *N* is the sum of occurrences of all dipeptides.

This study used DC as a protein descriptor. The reason for using DC is that it is easily extracted from protein sequences, consists of 400 features that cover characteristics of a protein, and can obtain good performance in the problem of classification or prediction [18]. Ong et al. [19] comparatively evaluated the effectiveness of the protein descriptor-sets using the same machine learning method and parameter optimization algorithm and examined whether the combination of descriptor improved the predictive performance. In this study, Ong et al. [19] used six individual descriptor-sets, including DC and four combination sets. The results show that all descriptors used in the study generally obtain good and similar performance. Moreover, the use of combination descriptor-sets only gives slightly better prediction than the use of individual descriptor-sets.

In this research, PubChem fingerprint and dipeptide descriptor were used as the drug compound features, and the protein target features, respectively. PubChem fingerprint was acquired using PubChemPy library in Python, while the dipeptide descriptor was calculated using *protr* package in R. Each record consists of 881 compound fingerprints and 400 protein dipeptide descriptors total features to represent the DTI samples is 1281 features.

Machine learning methods

We used three machine learning methods that have different ways of deciding to build a model for classifying objects into the appropriate class in the binary classification problem. The SVM makes a decision based on hyperplane [20] (Fig. 2). The hyperplane is obtained by minimizing the maximum distance of hyperplane and support vector (margins) with a minimum error that can be calculated based on the following equation:

$$\min P(w,b) = \frac{1}{2}\|w\|^2 + \varepsilon \tag{2}$$

with w as weight vector, b as bias score, and ε as a minimum error from the calculation [56].

To avoid misclassification of each training sample, the Regularization parameter (C parameter) is introduced to optimize the margin. The Eq. 2 can be improved as following [21] :

$$\begin{gathered} min_{w,b,\varepsilon}\frac{1}{2}w^T w + C\sum_{i=1}^{l}\varepsilon_i, \\ \text{subject to } y_i(w^T\phi(x_i) + b) \geq 1 - \varepsilon_i, \\ \varepsilon_i \geq 0, i = 1, \ldots, l, \end{gathered} \tag{3}$$

where $\phi(x_i)$ maps x_i into a higher dimensional space and C > 0 is the regularization parameter. The problem on Eq. 3 considers a high dimensional data. Due to the possible high dimensionality of the vector variable w, solve the following dual problem. [21] Using the primal-dual relationship, the optimal w is

$$w = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i), \tag{4}$$

and the decision function is

$$sgn(w^T\phi(x) + b) = sgn(\sum_{i=1}^{l} y_i \alpha_i \ K(x_i, x) + b), \tag{5}$$

The fitting model can be obtained by tuning the regularization parameter and the parameter of the kernel (K) used in training. In this research, we used the RBF kernel, which is defined as [22] :

$$K(x_i, x) = \exp\left(-\gamma\|x_i - x\|^2\right), \tag{6}$$

Where x is the data, i is the dimension and γ is a free parameter. The γ parameter from Eq. 6 and regularization parameter C from Eq. 3 need to be decided before training the data. To find the best score for the parameter (C, γ), we used the grid search with 5-cross validation [23].

The second machine learning method used in this study was Random Forest (RF). RF was a bagging-type ensemble of uncorrelated decision trees that trains several trees in parallel and used voting or the majority decision of the trees as the final decision [24]. Random forest (RF) constructed a large number of decision trees based on averaging random selection of predictor variables. When constructing the trees, whenever a split is considered, a random selection of *m* predictors was selected as a subset of split candidates from the complete set of predictors.

The fitting model can be obtained by tuning hyperparameters. The important hyperparameters included the number of subsamples of the original features used to build each decision tree (*mtry*) and the weight assigned to each class. We used grid search with 5-cross-validation when conducting tuning hyperparameter optimization to ensure that the random forest was exposed to all the statistical distributions in the training dataset.

The third machine learning method used in this research was Multi-Layer Perceptron (MLP). MLP works based on an artificial neural network [25]. In the MLP, the input was first transformed using a non-linear transformation. The input nodes in the input layer provided information from the outside to the network. The hidden layer nodes performed computations and transfer information from the input nodes to the output nodes. An MLP can have one or more hidden layers. In this research, we used two hidden layers. Lastly, the output nodes were responsible for computation and transferring information from the network to the outside. The optimal model can be obtained by tuning hyperparameters such as hidden layer size, activation function, optimizer, and class weight. The fitting model can be obtained by minimizing error or loss function. By using these different methods, it was expected that the more optimal screening results could be obtained than those of using only one method. Each model had a different range of results. Thus, we could reduce the number of the potential compound candidate by analysing the overlap of prediction results from those three machine learning methods.

Building prediction model

The first step of this building model was normalization. First of all, the drug compounds and target data that had already combined was normalized and split. The training dataset consisted of 685 samples with interactions (positive class) and 39,180 samples with those unknown interactions (negative class). Thus, this dataset was actually unbalanced with the ratio between positive and negative dataset of 1:57. The random oversampling with the replacement was applied to the 685 positive datasets to obtain 10,578 samples of positive data.

Moreover, the random under-sampling was applied to 39,180 negative datasets to reduce this dataset to 30% of the total negative dataset, to 11,754 samples of negative data. Thus, we had a total of 22,332 samples. This sampling was done five times to get five random datasets. Thus, we had five datasets that each of them consisted of 22,332 samples. Next, we randomly have chosen 70% of the total samples as a training dataset and 30 % of them as validation dataset. In the feature space, we had five matrixes of 15,632 x 1281 as the training set and 6,700 x 1281 as a validation set.

One of these five datasets was tuned for Multilayer Perceptron (MLP), Random Forest (RF), and Support Vector Machine (SVM) using grid search technique with 5-fold cross-validation implemented using the grid search function from the scikit-learn package in Python [26]. Grid search then saved the best parameters tuned based on the result of AUC from the cross-validation used inside the function. Four other models were built with the hyperparameter tuned from the first model. Next, the resulted models of each method were validated using the validation data set. The performance results, including accuracy, precision, recall, f-measure, and Area Under Curve (AUC) were calculated. Fig. 3 shows the schema of our approach.

Predicting Indonesia herbal compound

The prediction of Indonesia herbal compound was conducted using five models of each method (MLP, RF, SVM). We use Indonesia herbal compounds collected from HerbalDB database [27]. These herbal compounds had no label. Thus we predicted their interaction with the protein target using the validated model proposed in this study. For each method, the prediction result was obtained from the average probability score of the five prediction models. The herbal compounds that were predicted to have interactions with the protein target at least by two methods or have the average probability score ≥ 0.5 would be used for further analysis (Fig. 4).

Pharmacophore Modelling

Pharmacophore is defined by the interaction patterns of bioactive molecules with their target represented by a three-dimensional (3D) abstract feature arrangement that determines the types of interaction rather than specific functional groups. These types of interaction can, for example, include the formation of hydrogen bonds, charged interactions, metal interactions, or hydrophobic (H) and aromatic (AR) contacts [28]. Pharmacophore models can be generated using two different approaches depending on the input data used for model construction. In the structure-based approach, the interaction pattern of a molecule and its targets are directly extracted from the target ligand complex that is determined experimentally [29]. In the case of ligand-based modelling, the three-dimensional (3D) structures of two or more known active molecules are aligned, and common pharmacophore features shared among these training set molecules are identified. In the ligand-based approach, all the general chemical features of the pharmacophores should be considered essential, whereas in the structure-based approach it can be considered whether the chemical features of a molecule are directly involved in the ligand-binding or not [30].

In the pharmacophore modelling approach, we used two methods, namely Structure-Based Drug Design (SBDD) and Ligand-Based Drug Design (LBDD) using LigandScout 4.3 software [31] (One month-free trial). Based on a comparative analysis of 8 pharmacophore tools such as Catalyst, MOE, Pharmer, Unity, POT, LigandScout, Pharao, Phase, we have analysed the compound library enrichment. The analysis of algorithm combinations shows that LigandScout is capable of improving the enrichment of other algorithms. In particular, LigandScout seem to be complementary as there is an improvement of both enrichment factors if it used in a consecutive screening pipeline [32].

Pharmacophore modelling methods was conducted on macOS Mojave version 10.14.6; 2,3GHZ Intel Core i9 Processor; and 16 GB 2400 MHz DDR4 memory. For structure-based methods, we used the 3D structure of SARS-CoV-2 main protease, which could be downloaded from Protein Data Bank (PDB) with ID code 6LU7 [33]. We chose the pharmacophore sites of the native ligand and identified the pharmacophore features. Next, LigandScout performed screening of medicinal plant compounds from HerbalDB based on the similarity of pharmacophore of the native ligand of SARS-CoV-2 main protease.

Moreover, for the LBDD method, we collected 45 known SARS, MERS, and SARS-CoV-2 repurposing drug therapies from literature and used them as data sets (Additional file 4**)**. The molecules were downloaded from PubChem or prepared using MarvinSketch [34] and saved in .sdf format. The molecules were then separated into training and test set, 15 molecules as a training set and 30 molecules as a test set using sklearn.model_selection.train_test_split method in Python. For the pharmacophore modelling validation we adapt the methods from Wolber and Langer, 2015 and Seidel T, 2017 [31] [35]. To validate the pharmacophore model results, we did a validation process using decoy molecules that were generated using DUDE (www.dude.docking.org). We used DUDE to create decoy because DUD-E is one of the largest databases publicly available that offers the possibility to assess Virtual Screening programs efficiency in discriminating ligands from inactive compounds [36]. We prepared a library database in Screening Perspective for active (test set), and decoy set molecules and saved the library in .ldb format. Then, we screened active compounds (test set) and decoys based on each 10-models of 3D pharmacophore. When the screening process was finished, a hit list of molecules, that matches the pharmacophore, was shown in the Library View. Parameters of validation (ROC, AUC and EF) were calculated to choose the best model [37–39]. Pharmacophore models that gave the best score of validation parameters were used for virtual screening against Indonesian medicinal plant compounds database (HerbalDB) [27].

Molecular Docking

In the molecular docking step, we used hit compounds candidates yielded by two approaches, machine learning and pharmacophore modelling, and used macromolecules of SARS-CoV-2 main protease (PDB ID: 6LU7). Molecular docking was conducted on macOS Mojave version 10.14.6; 2,3GHZ Intel Core i9 Processor; and 16 GB 2400 MHz DDR4 memory. To validate the molecular docking, we did the redocking process of 6LU7 native ligand in AutoDock4 [40] software. The docking parameters used in this step is Lamarckian genetic algorithm [41] with default docking parameter; binding site coordinates x=-9.732, y=11.403, and z=68.925; grid box size 40 x 56 x 40. Autodock uses a Lamarckian Genetic Algorithm (LGA), which introduces local search based on the traditional genetic algorithm, making it more efficient in figuring out the optimal docking [42]. With these parameters, we got the RMSD value of native ligand as < 2 Å [43] and then applied these parameters for other ligands. Docking results were carried out based on scoring and posing functions. Docking interactions were clustered to decide the Gibbs energy (ΔG) and optimum docking energy conformation and ligand-residue interaction were considered as the fine-docked pose.

Results

Machine Learning

In the training phase, we conducted the hyperparameter tuning for each method (RF, MLP, and SVM) to get the optimal prediction model. The hyper-parameter used for each method (MLP, RF, and, SVM) can be seen in Additional file 5. The performance prediction model calculated using validation dataset was showed in Table 3. The accuracy and f-measure of the model of each method were high, around 98%, respectively. It meant that random over-sampling and random under-sampling could perform well.

Next, the models that had been optimized and validated used to predict 400 Indonesia herbal compound collected from HerbalDB. [17] Table 4 showed some predicted results of herbal compounds that target 3CLPro, PLPro, and RdRp. The remaining prediction results can be seen in Additional file 6. These candidate compounds have the potentiality to be compared to the pharmacophore modelling results. Some potential compounds resulted from a

machine learning approach and pharmacophore modelling approach were further analysed using molecular docking.

Pharmacophore Modelling

Structure-based Drug Design (SBDD) Methods

For SBDD methods, we analysed 3CLpro (main protease) protein in its 3D structure (PDB ID: 6LU7) using LigandScout software. The complex of main protease-ligand and its pharmacophore features are shown in Fig. 5a and 5b, respectively. Based on that pharmacophore feature, we screened herbal compounds from the HerbalDB database. From this screening, we got eight hit compounds consists of **Kaempferol 3,4'-di-O-methyl ether (Ermanin)**; 4-Methylpentyl glucosinolate; 6-alpha-Hydroxyadoxoside; Laurotetanine; Orientanol E; 5-Methoxy-8-O-beta-D-glucosyloxypsoralen; **Rhamnetin 3-mannosyl-(1-2)-alloside**; and 5,7,3',4'-Tetrahydroxyflavanone 7-alpha-Larabinofuranosyl-(1-6)-glucoside.

Ligand-based Drug Design (LBDD) Methods

From LBDD analysis, we got ten pharmacophore models, and then we validated to get the best pharmacophore model using decoy compounds. The validation parameters were $AUC_{100\%}$ and $EF_{1\%}$, the pharmacophore feature of the best pharmacophore model and its validation parameters are shown in Figure 6. The best pharmacophore model is model 4 with hit rate 27.17% (520 hits from total 1914 compounds (30 actives and 1,884 decoys)), $AUC_{100\%}$ 0.77 and $EF_{1\%}$ 13.4 and it has five pharmacophore features consists of three hydrogen bond acceptor (HBA) and two hydrogen bond donor (HBD). From the best pharmacophore model that was generated in the previous step, then we screened against herbal compounds from HerbalDB and got top 30 hit compounds. The top 30 of hit compounds are shown in Table 5.

Molecular Docking

Before we started to dock the hit compounds to 3CLpro protein, we have redocked the native ligand to 3CLpro binding site, to confirm the suitability of the docking algorithm for virtual screening. The RMSD of re-docking of 6LU7 native ligand was 0.34 Å, respect to the co-crystallized one. Although neither an effective antiviral drug nor a vaccine against COVID-19 is currently available, several reports suggested that HIV-1 protease inhibitors, such as Lopinavir, have the potential as SARS-CoV-2 protease inhibitor [6]. In an attempt to have reference values (positive control), we decided to consider Lopinavir as comparative standards for the molecular docking. Based on machine learning, structure-based and ligand-based pharmacophore results, we got 14 hit compounds that overlap from machine learning and pharmacophore modelling approach. Then we used molecular docking to analyse the interaction between 3CLpro (main protease) protein in its 3D structure (PDB ID: 6LU7) with 14 hit compounds and used Lopinavir as a positive control (Table 6).

From the molecular docking results, the tested compounds showed various binding energies (ΔG). Compounds that have binding energy close to Lopinavir (positive control) are **Hesperidin, Kaempferol-3,4'-di-O-methyl ether (Ermanin); Myricetin-3-glucoside, Peonidine 3-(4'-arabinosylglucoside); Quercetin 3-(2G-rhamnosylrutinoside); and Rhamnetin 3-mannosyl-(1-2)-alloside**. Hesperidin shows the lowest binding energy (-8.72 kcal/mol) and close to Lopinavir binding energy (-9.41 kcal/mol). As seen in Fig. 7, Lopinavir has hydrogen bond with Glu166, which Glu166 is an essential residue for keeping the S1 pocket in the right shape and the enzyme in the active conformation [44]. Hesperidin, Kaempferol-3,4'-di-O-methyl ether (Ermanin), Quercetin 3-(2G-rhamnosylrutinoside), Peonidine 3-(4'-arabinosylglucoside), Quercetin 3-(2G-rhamnosylrutinoside), and Rhamnetin 3-mannosyl-(1-2)-alloside have the hydrogen bond with Glu166 residue as well. Lopinavir also has binding interaction with the catalytic dyad (Cys-145 and His-41) of SARS-CoV-2, as well as other six compounds. The catalytic dyad is functionally essential residues (Cys-145 and His-41) that displayed stable behaviour [45].

Discussion

The SARS-CoV-2 virus is still emerging around the world. The number of infected people continues to overgrow, and still no definitive therapy that has been approved for effective treatment. Finding broad-spectrum inhibitors that could reduce the effects of coronavirus infection in human remains a challenging research focus. Given the time-consuming nature of developing and registering antiviral drugs, drug repurposing is one of shortcut to cure the disease. For most of these drugs that have been prepared, they have sufficient experience and dosage, and their safety and ADME situation are well known.

Despite continuing the research of conventional medicine, Indonesia that has mega biodiversity has potential herbal compounds as SARS-CoV-2 inhibitor for an alternative. In order to get the potential herbal compounds by

using a computational approach, we have to be careful about the research methods. We should not use our self-preferential in certain herbal, because it would lead to a subjective decision on the research results. Especially when the computational approach only uses molecular docking method. While molecular docking is a powerful tool for pharmaceutical research after decades of development, there is a limitation of docking accuracy due to relatively simple scoring functions.

Additionally, entropic factors are generally not captured well by scoring based on a single structure. As a result, structure-based ligand screening by docking often generates a large number of false positive hits [46]. To minimize the false positive hits by conducted research with molecular docking only, we tried to use two different approaches in generating the prediction model before we did the virtual screening on HerbalDB compounds. In this study, we used machine learning and pharmacophore modelling methods that are complementary to each other to generate more accurate prediction model.

The machine learning approach was used to perform big data analysis DTI dataset that collected from literature and public domain database. This approach used pharmacological features obtained by integrating both the chemical space of compounds and the omics space of target proteins [47]. Cheminformatic studies found that machine learning approaches, such as similarity measure [48], bipartite graph [49] and some classification techniques were useful in finding interaction between drug and its protein targets [11,50]. Most of the classification model was built for single-target protein drug problems. For instance, Support Vector Machine (SVM), as one of machine learning methods, it can be employed to classify whether a compound is drug-like or non-drug like [11]. Decision Tree and Neural Network had also been attempted to distinguish the drug-like compounds from the non-drug like compounds [51–53]. These approaches showed a maximum accuracy up to 83% from a large dataset. In this study, the enhancement of those machine learning methods was done to classify whether a drug compound s with protein target or not.

Dealing with the issue of high dimensional data in the feature space formed by the fingerprint of compounds and the dipeptide descriptors of proteins, many papers show the effectiveness of the embedded capacity of several classifiers [54] such as SVM [55], neural network-based algorithm (MLP) [56], and decision tree-based algorithm (RF) [57] to discard input features. Embedded methods had the advantage that they include interaction with the classification model [54]. In the Random Forest method, we tuned *mtry* that indicated a random selection of m predictors as a subset of split candidates from the full set of predictors when building trees. Thus, in RF, the high dimensionality is reduced by choosing *mtry* smaller than the number of features. However, for MLP and SVM, Even though both of them were able to handle non-linearity (SVM with the kernel; MLP with multilayers), they are still vulnerable to spurious correlation. It meant there were some features that appeared to be highly correlated in training data, but less sensitive in real prediction using testing data. The prediction model generated by SVM showed the tendency. Although all model generated by three methods (MLP, RF, dan SVM) had high accuracy in the validation step, the SVM failed to predict herbal compound. Only very view herbal compound can be predicted by SVM compared to MLP and RF.

The unbalanced dataset probably also contributed to the performance of SVM. The random oversampling was not adequate to improve the performance of SVM because the number of different support vectors did not increase. Thus, the hyperplane was not improved. Oversampling with replacement did not affect the distribution of support vector but affected class probability. Therefore, in this case, RF is more robust than other methods because the oversampling increased the class probability that was required for splitting when building the tree.

However, although the prediction model constructed by SVM could not perform well in predicting herbal compound, our criteria determined that the herbal compounds candidate should be predicted to at least by two methods or should have the average probability score ≥ 0.5. Moreover, the predicted results would be filtered again by comparing to ligand-based and structure-based pharmacophore methods. Thus, our approach provided layered filtering in order to conduct more objective and optimal virtual screening. As stated in [58], machine learning approach can be used to predict the DTI with insufficient known ligands.

A pharmacophore is the pattern of features of a molecule that is responsible for a biological effect, which captures the essential notion that a pharmacophore is built from features rather than defined chemical groups. Every type of atom or group in a molecule that exhibits specific properties related to molecular recognition can be reduced to a pharmacophore feature. These molecular patterns can be labelled as hydrogen bond donors or acceptors, cationic, anionic, aromatic, or hydrophobic, and any possible combinations. Pharmacophore models are very suitable as queries for virtual screening of databases. Pharmacophore models are often utilised as a filter to identify compounds that fulfil simple geometric and chemical functionality requirements of the query, before more complicated and computationally demanding approaches such as molecular docking [59]. Thus, using two

approaches in the methodology, i.e. machine learning and pharmacophore modelling, will lead us to increase the confidence level of the predicted compounds candidates.

Based on the virtual screening on HerbalDB using two prediction approaches, we got 14 compounds that overlap from two method results. Molecular docking algorithms are often calibrated against experimental ligand-protein complex training sets, and the accuracy of these docking programs is often highly dependent on the training sets used [58]. In this case, it is essential to ensure that the docking software used for virtual screening can replicate the binding mode of a known experimental inhibitor for the enzymes studied. From molecular docking analysis, we got six potential compounds, i.e. Hesperidin, Kaempferol-3,4'-di-O-methyl ether (Ermanin); Myricetin-3-glucoside, Peonidine 3-(4'-arabinosylglucoside); Quercetin 3-(2G-rhamnosylrutinoside); and Rhamnetin 3-mannosyl-(1-2)-alloside, that predicted could inhibit the 3CLpro protein of SARS-CoV-2.

After we got this result, we further checked the previous studies to find the biological activities of each compound, so that this research can be useful for the community. We also tried to find from commodity crops. One of the commodity crops in Indonesia is Guava (*Psidium guajava*) that can be harvested continuously in one year. In Indonesia, production of guava in the year 2018 is 230,697 tons, with growth rate from the year 2017 to 2018 is 15.06% [60]. Guava is consumed not only as food but also as a traditional medicine in subtropical areas around the world due to its pharmacologic activities. Based on Herbal Regulation as Healthy Supplement for Fighting COVID-19 in Indonesia published by The Indonesian Food and Drug Authority (BPOM) (May, 2020), we can consume *Psidium guajava* (Guava) 1-4 fruits per day (55-100 gram/fruit) which contain vitamin C 228.3 mg in 100 gram fruit. For the administration, Guava can be eaten directly or processed as juice. There is no case for toxicity for long term consumption, overall this herbal is safe to use as daily nutritional supplement [61]. Phenolic compound from Guajava has been proved as immunomodulator and antioxidant [62,63].

Guava is well known has several flavonoids compounds, i.e. myricetin, quercetin, luteolin, kaempferol, isorhamnetin [64], and Hesperidin [65]. These compounds were also shown in our result, although without the aglycones. Luteolin is known as a furin protein inhibitor [66] which is predicted to be one of the enzymes that break down Coronavirus S (spike) protein as in MERS into units S1 and S2 [67]. In the S1 unit, there is a Receptor Binding Domain (RBD) where the ACE2 peptidase binds so that the virus can bind to the host [67]. Hesperidin / Hesperitin compounds in the *in silico* study are known to inhibit RBD domain binding of the SARS-COV-2 Spike protein with ACE2 receptors in humans so that it is predicted to inhibit the entry of the SARS-COV-2 potentially [5]. It is also known that luteolin is a neuraminidase inhibitor as well as oseltamivir which is currently one of the drugs used in the CDC protocol. Hesperitin (the form of hesperidin aglycone) and quercetin are known to also act as inhibitors of 3Clpro [68,69]. Other compounds in guava such as myricetin are known to act as SARS coronavirus helicase inhibitors [70]. The kaempferol has the potential to be a non-competitive inhibitor of 3CLPro and PLpro as well as quercetin [71]. Another interesting thing is that kaempferol acts as a modulator of autophagy, which can be utilised in strategies to inhibit SARS-COV-2 virus.

Conclusions

We used layered virtual screening with machine learning and pharmacophore modelling approaches that could provide more objective and optimal virtual screening and avoid subjective decision making on research results. Herbal compounds from various plants have potential as antiviral candidates for SARS-CoV-2. Based on our research and literature study, one of Indonesia's potential commodity crops is *Psidium guajava* (guava), and people can use it directly as a preventive effort.

List of abbreviations

3CLPro: 3C-like protease; AUC: Area Under the Curves; COVID-19: Coronavirus Disease 2019; DTI: Drug Target Interactions; DUDE: Database of Useful (Docking) Decoys; EF: Enrichment Factor; LBDD: Ligand-based Drug Design; MLP: Multilayer Perceptron; MERS-CoV: Middle East Respiratory Syndrome coronavirus; PDB: Protein Data Bank; PLPro: Papain-like Protease; RdRp: RNA-dependent RNA Polymerase, RF: Random Forest; SARS-CoV: Severe Acute Respiratory Syndrome coronavirus; SARS-CoV-2: Severe Acute Respiratory Syndrome coronavirus 2; SBDD: Structure-based Drug Design; SVM: Support Vector Machine.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication
Not Applicable

Availability of data and materials
All data and source codes of the machine learning approach used in this research can be accessed at https://github.com/TropBRC-BioinfoLab/virtual-screening-covid19. HerbalDB datasets are available on request to the authors.

Competing interests
The authors declare that they have no competing interests

Funding
This research didn't funded by any research funding.

Authors' contributions
Conceptualization: LE, RIP, FF, AT, WAK; methodology: LE, RIP, WAK; formal analysis: LE, RIP, WAK, NSR, AKN, FKS, AF, KAI; writing—original draft preparation and editing: LE, RIP, WAK, NSR; supervision: FF, AT, IPP, AY. LE, RIP, WAK contributed equally to this work. All authors read and approved the final manuscript.

Acknowledgements
Not Applicable

References

1. Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis. 2020;91:264–6.
2. Zumla A, Chan JFW, Azhar EI, Hui DSC, Yuen KY. Coronaviruses-drug discovery and therapeutic options. Nat Rev Drug Discov. 2016;15(5):327–47.
3. Song Z, Xu Y, Bao L, Zhang L, Yu P, Qu Y, et al. From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses. 2019;11(1).
4. Omrani AS, Saad MM, Baig K, Bahloul A, Abdul-Matin M, Alaidaroos AY, et al. Ribavirin and interferon alfa-2a for severe Middle East respiratory syndrome coronavirus infection: A retrospective cohort study. Lancet Infect Dis. 2014;14(11):1090–5.
5. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B [Internet]. 2020;(PG-). Available from: http://www.sciencedirect.com/science/article/pii/S2211383520302999 NS -
6. Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat Rev Drug Discov. 2020;19:149–50.
7. Ashburn TT, Thor KB. Drug repositioning: Identifying and developing new uses for existing drugs. Vol. 3, Nature Reviews Drug Discovery. Nature Publishing Group; 2004. p. 673–83.
8. Novac N. Challenges and opportunities of drug repositioning. Trends Pharmacol Sci. 2013 May;34(5):267–72.
9. Vyas VK, Goel A, Ghate M, Patel P. Ligand and structure-based approaches for the identification of SIRT1 activators. Chem Biol Interact [Internet]. 2015 Feb;228:9–17. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0009279715000046
10. Fitriawan A, Wasito I, Syafiandini A., Azminah A, Amien M, Yanuar A. Deep Belief Networks for Ligand-Based Virtual Screening of Drug Design. In: Proceeding of 6th International Workshop on Computer Science and Engineering. 2016.
11. Dhanda SK, Singla D, Mondal AK, Raghava GPS. DrugMint: a webserver for predicting and designing of drug-like molecules. Biol Direct. 2013 Dec 5;8(1):28.
12. Liu Y. Machine learning for drug design. Int J Comput Inf Technol. 2015;4(1).
13. Johnson A, Maggiora G. Concepts and Applications of Molecular Similarity. John Willey&Sons. New York; 1990.
14. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Annual Reports in Computational Chemistry. 2008. p. 217–41.
15. Skinnider MA, Dejong CA, Franczak BC, McNicholas PD, Magarvey NA. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. J Cheminform. 2017 Dec 16;9(1):46.

16. Fernández-de Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform. 2017 Dec 6;9(1):9.
17. Gao Q-B, Wang Z-Z, Yan C, Du Y-H. Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett. 2005 Jun 20;579(16):3444–8.
18. Bhasin M, Raghava GPS. Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. J Biol Chem. 2004 May 28;279(22):23262–6.
19. Ong SAK, Lin H, Chen Y, Li Z, Cao Z. Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinformatics [Internet]. 2007;8(1):300. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-300
20. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20(3):273–97.
21. Chang C-C, Lin C-J. LIBSVM. ACM Trans Intell Syst Technol. 2011 Apr;2(3):1–27.
22. Vert J, Tsuda K, Schölkopf B. A Primer on Kernel Methods. Kernel Methods Comput Biol [Internet]. 2004;47:35–70. Available from: https://direct.mit.edu/books/book/3898/chapter/163643/a-primer-on-kernel-methods
23. Apostolidis-afentoulis V. SVM Classification with Linear and RBF kernels. ResearchGate. 2015;
24. Goel E, Abhilasha E. Random Forest: A Review. Int J Adv Res Comput Sci Softw Eng. 2017;7(1).
25. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer (Long Beach Calif). 1996;29(3):31–44.
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2012 Jan 2;
27. Yanuar A, Munim A, Bertha A, Lagho A, Syahdi RR, Rahmat M, et al. Medicinal Plants Database and Three Dimensional Structure of the Chemical Compounds from Medicinal Plants in Indonesia. Int J Comput Sci. 2011;8(5):180–3.
28. Kaserer T, Beck KR, Akram M, Odermatt A, Schuster D, Willett P. Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. Molecules. 2015;20(12):22799–832.
29. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The protein data bank. Acta Crystallogr Sect D Biol Crystallogr. 2002;58.
30. Li Y, Fu X, Duan D, Liu X, Xu J, Gao X. Extraction and Identification of Phlorotannins from the Brown Alga, Sargassum fusiforme (Harvey) Setchell. Mar Drugs. 2017;15(2).
31. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model. 2005;
32. Sanders MPA, Barbosa AJM, Zarzycka B, Nicolaes GAF, Klomp JPG, de Vlieg J, et al. Comparative Analysis of Pharmacophore Screening Tools. J Chem Inf Model. 2012 Jun 25;52(6):1607–20.
33. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020 Jun;582(7811):289–93.
34. Volford A. MarvinSketch. Chem Axon MarvinSketch. 2015;
35. Seidel T, Bryant SD, Ibis G, Poli G LT. 3D pharmacophore modeling techniques in computer-aided molecular design using LigandScout. In: Varnek A, editor. Tutorials in Chemoinformatics. John Wiley & Sons Ltd.; 2017. p. 279–309.
36. Chaput L, Martinez-Sanz J, Saettel N, Mouawad L. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. J Cheminform. 2016 Dec;8(1):56.
37. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29–36.
38. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006 Jun;27(8):861–74.
39. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J Med Chem. 2004 Mar;47(7):1750–9.
40. Steffen C, Thomas K, Huniar U, Hellweg A, Rubner O, Schroer A. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. J Comput Chem. 2010;
41. Liu Z, Zhang C, Zhao Q, Zhang B, Sun W. Comparative Study of Evolutionary Algorithms for Protein-Ligand Docking Problem on the AutoDock. In: Houbing, Song; Dingde J, editor. Simulation Tools and Techniques. Springer International Publishing; 2019. p. 598–607.
42. Wang L, Weng Z, Liang Y, Wang Y, Zhang Z, Di R. Design and Implementation of Parallel Lamarckian Genetic Algorithm for Automated Docking of Molecules. In: 2008 10th IEEE International Conference on High Performance Computing and Communications. IEEE; 2008. p. 689–94.
43. Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SW, et al. Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. J Chem Inf Model. 2009 Feb 23;49(2):444–60.

44. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science (80-). 2020;
45. ul Qamar MT, Alqahtani SM, Alamri MA, Chen L-L. Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants. J Pharm Anal. 2020;In Press.
46. Deng N, Forli S, He P, Perryman A, Wickstrom L, Vijayan RSK, et al. Distinguishing binders from false positives by free energy calculations: Fragment screening against the flap site of HIV protease. J Phys Chem B. 2015;119(3):976–88.
47. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, et al. Toward more realistic drug-target interaction predictions. Brief Bioinform. 2015;16:325–7.
48. Shi J-Y, Yiu S-M, Li Y, Leung HCM, Chin FYL. Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. Methods. 2015 Jul;83:98–104.
49. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug—target network. Nat Biotechnol. 2007 Oct 5;25(10):1119–26.
50. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, et al. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. Altman RB, editor. PLoS Comput Biol [Internet]. 2012 May 10;8(5):e1002503. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002503
51. Gillet VJ, Willett P, Bradshaw J. Identification of biological activity profiles using substructural analysis and genetic algorithms. J Chem Inf Comput Sci [Internet]. 1998;38(2):165–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9538517
52. Wagener M, van Geerestein VJ. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. J Chem Inf Comput Sci. 2000 Mar;40(2):280–92.
53. Frimurer TM, Bywater R, Nærum L, Lauritsen LN, Brunak S. Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds. J Chem Inf Comput Sci. 2000 Nov;40(6):1315–24.
54. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007 Oct 1;23(19):2507–17.
55. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
56. Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, et al. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. Bioinformatics. 2002 Mar 1;18(3):395–404.
57. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics. 2003 Sep 1;19(13):1636–43.
58. Glombitza KW, Keusgen M. Fuhalols and deshydroxyfuhalols from the brown alga sargassum spinuligerum. Phytochemistry. 1995;38(4):987–95.
59. Voet A, Qing X, Lee XY, De Raeymaecker J, Tame J, Zhang K, et al. Pharmacophore modeling: advances, limitations, and current utility in drug discovery. J Receptor Ligand Channel Res. 2014 Nov;2014(7):81.
60. Subdirektorat Statistik Hortikultura. Statistik Tanaman Buah-buahan dan Sayuran Tahunan Indonesia. Jakarta: Badan Pusat Statistik; 2018. 4, 12 p.
61. Badan Pengawas Obat dan Makanan Republik Indonesia. Pedoman Penggunaan Herbal dan Suplemen Kesehatan Dalam Menghadapi COVID-19 di Indonesia. 1st ed. Vol. 1, Badan Pengawas Obat dan Makanan Republik Indonesia. Jakarta: Badan Pengawas Obat dan Makanan Republik Indonesia; 2020. 40–46 p.
62. Laily N, Kusumaningtyas RW, Sukarti I, Rini MRDK. The Potency of Guava Psidium Guajava (L.) Leaves as a Functional Immunostimulatory Ingredient. Procedia Chem. 2015 Jan;14:301–7.
63. Vasconcelos AG, Amorim A das GN, dos Santos RC, Souza JMT, de Souza LKM, Araújo T de SL, et al. Lycopene rich extract from red guava (Psidium guajava L.) displays anti-inflammatory and antioxidant profile by reducing suggestive hallmarks of acute inflammatory response in mice. Food Res Int. 2017 Sep;99:959–68.
64. Hamid Musa K, Abdullah A, Subramaniam V. Flavonoid profile and antioxidant activity of pink guava. ScienceAsia. 2015;41(3):149.
65. Trujillo-Correa AI, Quintero-Gil DC, Diaz-Castillo F, Quiñones W, Robledo SM, Martinez-Gutierrez M. In vitro and in silico anti-dengue activity of compounds obtained from Psidium guajava through bioprospecting. BMC Complement Altern Med. 2019;19(1):1–16.
66. Peng M, Watanabe S, Chan KWK, He Q, Zhao Y, Zhang Z, et al. Luteolin restricts dengue virus replication through inhibition of the proprotein convertase furin. Antiviral Res [Internet]. 2017;143:176–85. Available from: http://dx.doi.org/10.1016/j.antiviral.2017.03.026
67. Kleine-Weber H, Elzayat MT, Hoffmann M, Pöhlmann S. Functional analysis of potential cleavage sites in the MERS-coronavirus spike protein. Sci Rep. 2018;8(1):1–11.
68. Lin CW, Tsai FJ, Tsai CH, Lai CC, Wan L, Ho TY, et al. Anti-SARS coronavirus 3C-like protease effects

of Isatis indigotica root and plant-derived phenolic compounds. Antiviral Res. 2005;68(1):36–42.
69. Nguyen TTH, Woo HJ, Kang HK, Nguyen VD, Kim YM, Kim DW, et al. Flavonoid-mediated inhibition of SARS coronavirus 3C-like protease expressed in Pichia pastoris. Biotechnol Lett. 2012;34(5):831–8.
70. Yu MS, Lee J, Lee JM, Kim Y, Chin YW, Jee JG, et al. Identification of myricetin and scutellarein as novel chemical inhibitors of the SARS coronavirus helicase, nsP13. Bioorganic Med Chem Lett [Internet]. 2012;22(12):4049–54. Available from: http://dx.doi.org/10.1016/j.bmcl.2012.04.081
71. Park JY, Yuk HJ, Ryu HW, Lim SH, Kim KS, Park KH, et al. Evaluation of polyphenols from Broussonetia papyrifera as coronavirus protease inhibitors. J Enzyme Inhib Med Chem. 2017;32(1):504–12.
72. Weston J, Watkins C. Support Vector Machines for Multi-Class Pattern Recognition. Proc 7th Eur Symp Artif Neural Networks. 1999;219–24.
73. Laskowski RA, Swindells MB. LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model. 2011;
74. Chen H, Du Q. Potential Natural Compounds for Preventing 2019-nCoV Infection. Preprints. 2020;2020010358.
75. Kindrachuk J, Ork B, Hart BJ, Mazur S, Holbrook MR, Frieman MB, et al. Antiviral potential of ERK/MAPK and PI3K/AKT/mTOR signaling modulation for Middle East respiratory syndrome coronavirus infection as identified by temporal kinome analysis. Antimicrob Agents Chemother. 2015;59(2):1088–99.
76. Cheung NN, Lai KK, Dai J, Kok KH, Chen H, Chan KH, et al. Broad-spectrum inhibition of common respiratory RNA viruses by a pyrimidine synthesis inhibitor with involvement of the host antiviral response. J Gen Virol. 2017;98(5):946–54.
77. Ahmed-Belkacem A, Colliandre L, Ahnou N, Nevers Q, Gelin M, Bessin Y, et al. Fragment-based discovery of a new family of non-peptidic small-molecule cyclophilin inhibitors with potent antiviral activities. Nat Commun. 2016;7.
78. Hong S-S, Choi JH, Lee SY, Park Y-H, Park K-Y, Lee JY, et al. A Novel Small-Molecule Inhibitor Targeting the IL-6 Receptor β Subunit, Glycoprotein 130. J Immunol. 2015;195(1):237–45.
79. Lim TK, Lim TK. Brassica oleracea (Botrytis Group). In: Edible Medicinal And Non-Medicinal Plants. Springer Netherlands; 2014. p. 571–93.
80. Islam S. Sweetpotato (Ipomoea batatas L.) Leaf: Its Potential Effect on Human Health and Nutrition. J Food Sci. 2006 May;71(2):R13–121.
81. Lim TK, Lim TK. Raphanus raphanistrum subsp. sativus. In: Edible Medicinal and Non Medicinal Plants. Springer Netherlands; 2015. p. 829–69.
82. Koo HJ, Lee S, Shin KH, Kim BC, Lim CJ, Park EH. Geniposide, an anti-angiogenic compound from the fruits of Gardenia jasminoides. Planta Med. 2004 May;70(5):467–9.
83. Parhiz H, Roohbakhsh A, Soltani F, Rezaee R, Iranshahi M. Antioxidant and Anti-Inflammatory Properties of the Citrus Flavonoids Hesperidin and Hesperetin: An Updated Review of their Molecular Mechanisms and Experimental Models. Phyther Res. 2015 Mar;29(3):323–31.
84. Yan Huan, Pan Li-Li, Zhao Qing LH-Y. Chemical Contituents of Coleus forskohlii--. J Yunnan Univ Tradit Chinese Med [Internet]. 2012; Available from: http://en.cnki.com.cn/Article_en/CJFDTotal-YNZY201202006.htm
85. Widyowati R, Agil M. Chemical Constituents and Bioactivities of Several Indonesian Plants Typically Used in Jamu. Chem Pharm Bull. 2018 May;66(5):506–18.
86. (Globinmed) GIHOIM. Plantago major L [Internet]. Global Information Hub On Integrated Medicine (Globinmed). 2016 [cited 2020 Apr 7]. Available from: http://www.globinmed.com/index.php?option=com_content&view=article&id=106097:plantago-major-l&catid=286&Itemid=357
87. Bharat Singh RAS. Secondary Metabolites of Medicinal Plants. John Wiley & Sons Ltd.; 2020. 17 p.
88. Yanuar A, Suhartanto H, Mun'im A, Anugraha BH, Syahdi RR. Virtual Screening of Indonesian Herbal Database as HIV-1 Protease Inhibitor. Bioinformation. 2014;10(2):52–5.
89. Gupta O, Gupta R, Gupta P. Chemical Examination of Flowers of Ipomoea fistulosa. Planta Med. 1980 Feb;38(02):147–50.
90. Nair V, Bang WY, Schreckinger E, Andarwulan N, Cisneros-Zevallos L. Protective Role of Ternatin Anthocyanins and Quercetin Glycosides from Butterfly Pea (Clitoria ternatea Leguminosae) Blue Flower Petals against Lipopolysaccharide (LPS)-Induced Inflammation in Macrophage Cells. J Agric Food Chem. 2015 Jul;63(28):6355–65.
91. Anwar F, Latif S, Ashraf M, Gilani AH. Moringa oleifera: a food plant with multiple medicinal uses. Phyther Res. 2007 Jan;21(1):17–25.
92. Gupta D, Singh J. Flavonoid glycosides from Cassia alata. Phytochemistry. 1991 Jan;30(8):2761–3.
93. Mazumder A, Dwivedi A, du Plessis J. Sinigrin and Its Therapeutic Benefits. Molecules. 2016

Mar;21(4):416.
94. Spiraeoside C21H20O12, FLAVONOID Flavonol - Extrasynthese [Internet]. [cited 2020 Apr 7]. Available from: https://www.extrasynthese.com/spiraeoside-1809.html
95. Pandustore. Ginje (Thevetia peruviana) [Internet]. [cited 2020 Apr 7]. Available from: https://pandustore.co.id/ginje/

Additional files

Additional file 1. List of Potential Virus-based Drug Related to COVID-19
Additional file 2. List of Potential Human-based Drug Related to COVID-19
Additional file 3. Training dataset of Drug Target Interactions
Additional file 4. Training and test dataset for ligand-based method
Additional file 5. The optimal hyper-parameter values of each model
Additional file 6. The prediction results of herbal compounds

Figures

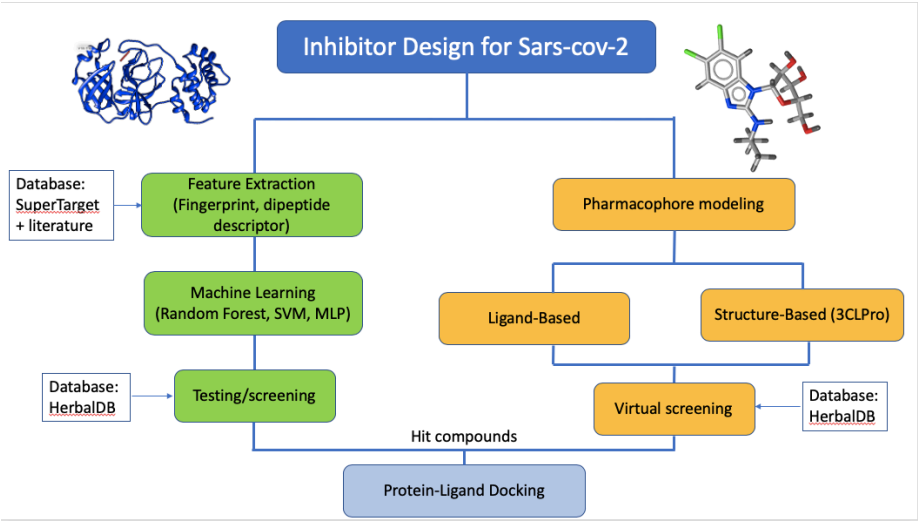

Fig. 1 Graphical methods of this study

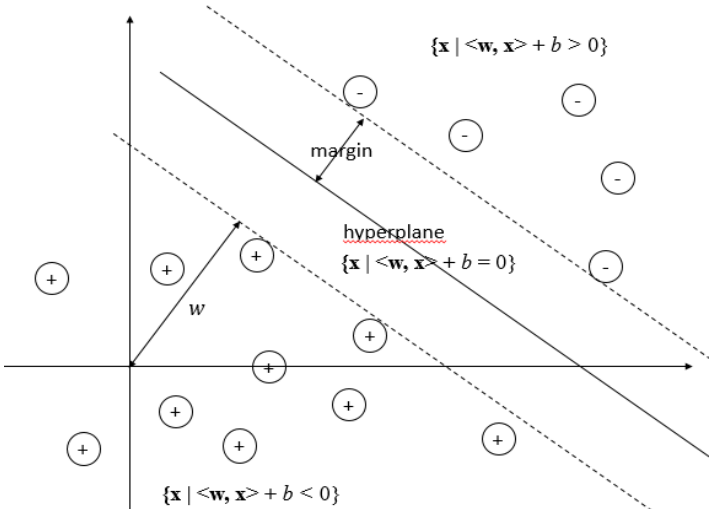

Fig. 2 Support vector machine method (x is the data, w is the weight vector, b is the bias score, ε is the minimum error) [72]

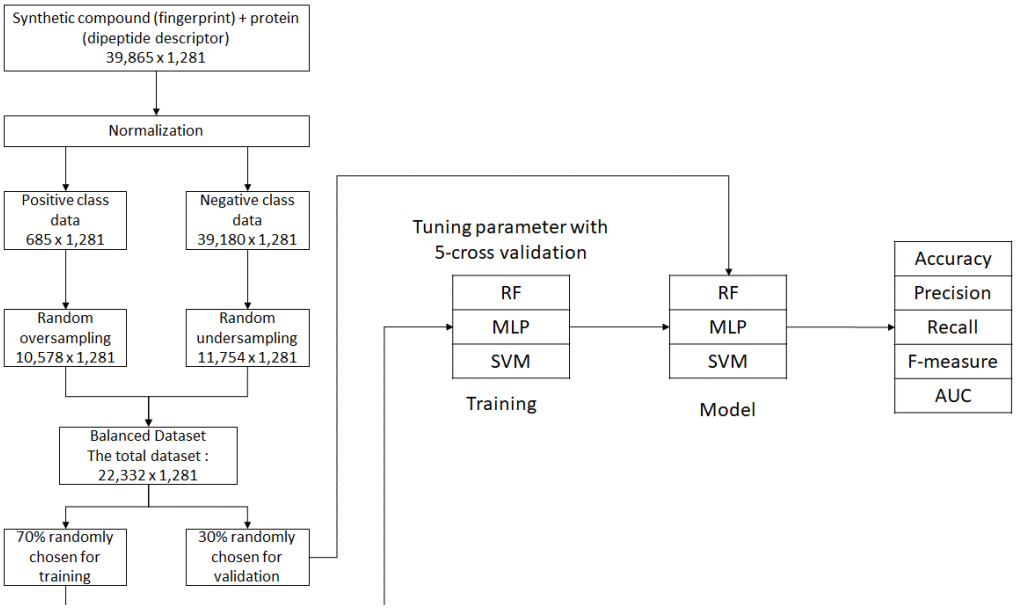

Fig. 3 The scheme of training and validating model of our proposed machine learning approach

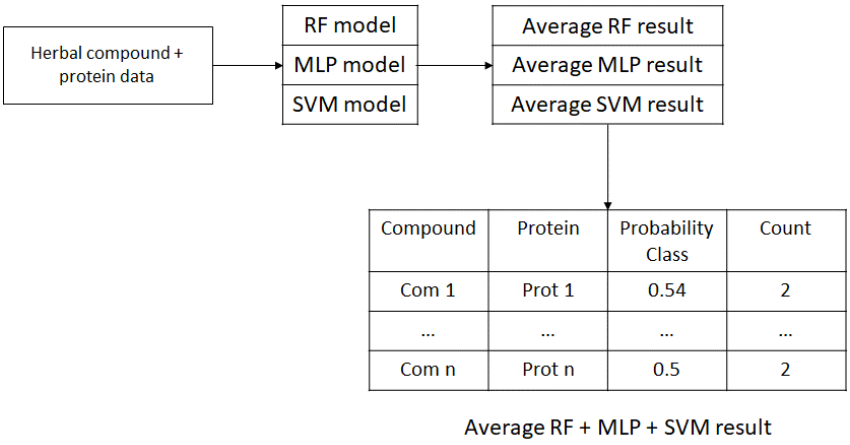

Fig. 4 The scheme of predicting Indonesia herbal compound using the optimal and validated model that had generated by RF, MLP, and SVM in training and validation phase. The class probability score was averaged from three methods. The counter was conducted to indicate the number of methods which predict positive results. The decision was determined based on the criteria of the class probability ≥ 0.5 or at least predicted by two methods.

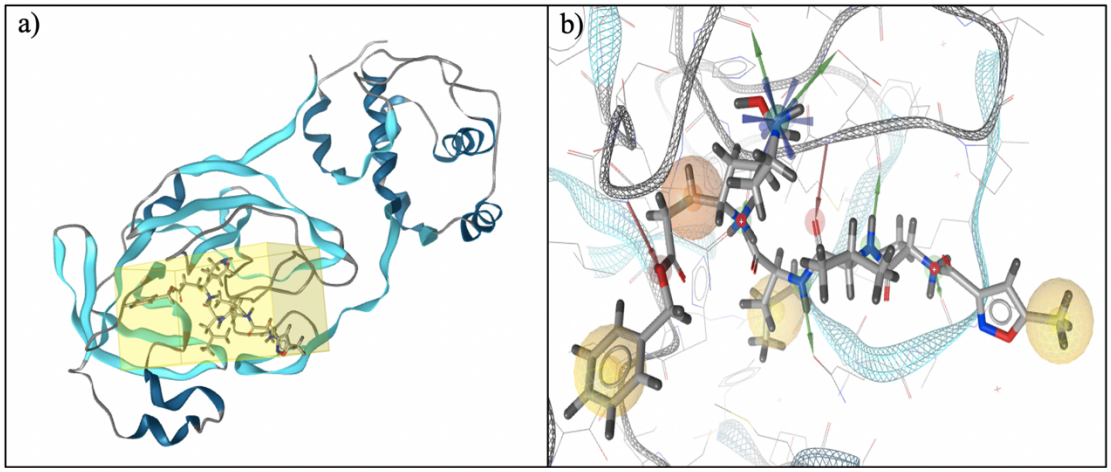

Fig. 5 a) 3D structure complex of the main protease and its ligand, b) pharmacophore feature of the native ligand in the main protease

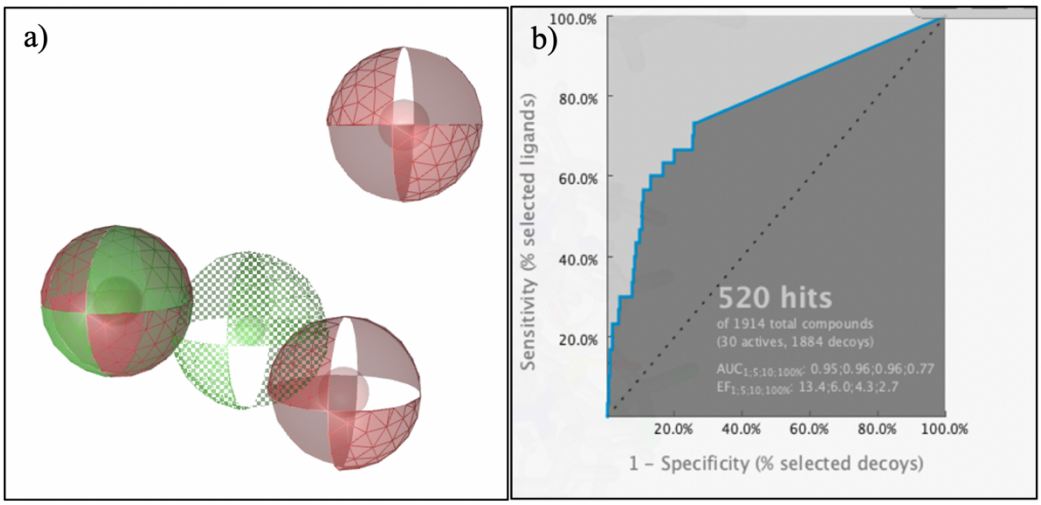

Fig. 6 Pharmacophore model from LBDD analysis. a) Pharmacophore feature of the best pharmacophore model, b) validation parameters of the best pharmacophore model.

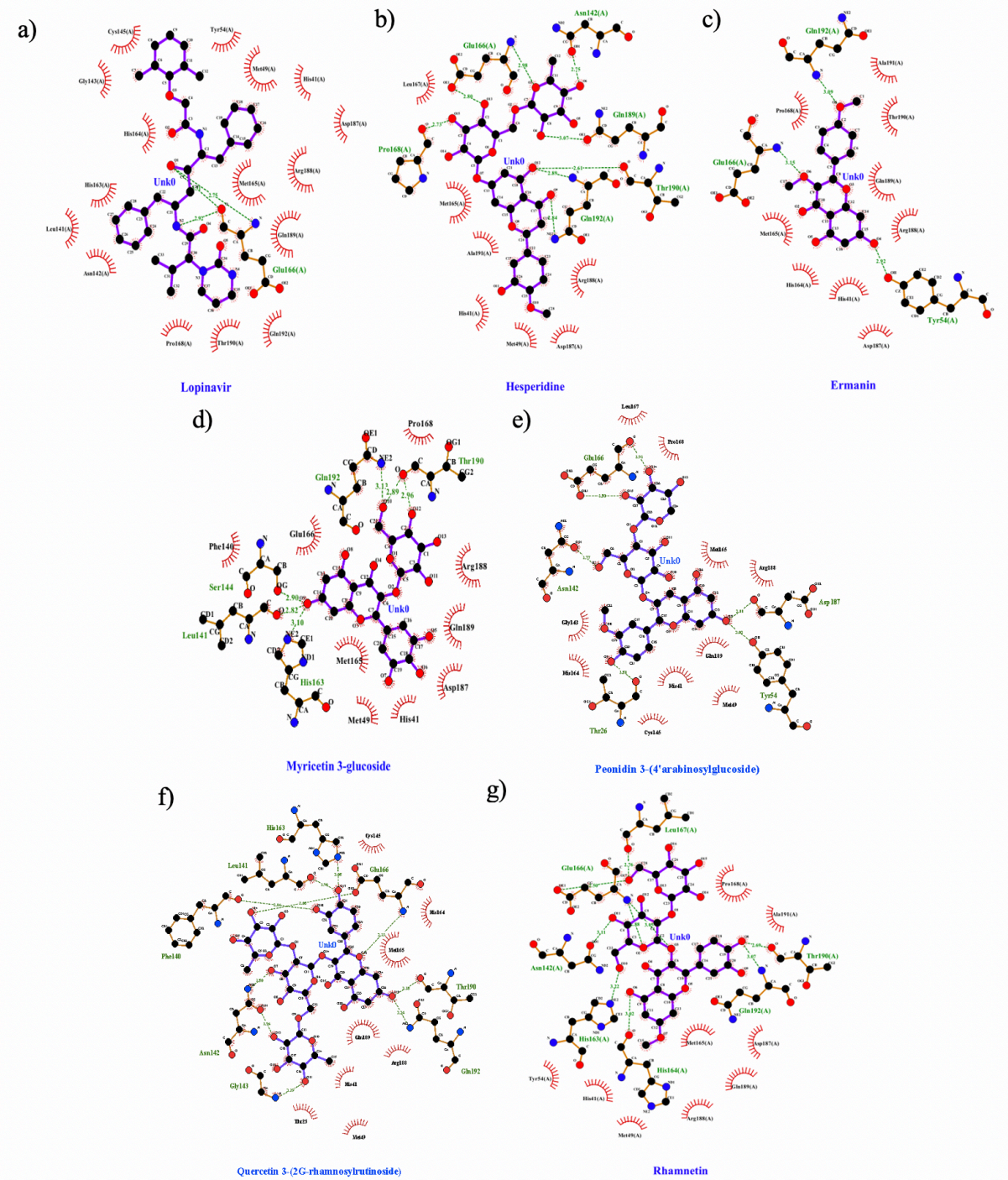

Fig. 7 Interaction of ligands with receptor (3CLpro / main protease); red quarter circles were residue of protein that have non-covalent bond interaction with ligand; residues that written in green colour were residue which had hydrogen bonds interaction with ligand (written with its distance as well). a) Lopinavir; b) Hesperidin; c) Kaempferol-3,4'-di-O-methyl ether (Ermanin); d) Myricetin-3-glucoside; e) Peonidine 3-(4'-arabinosylglucoside); f) Quercetin 3-(2G-rhamnosylrutinoside); g) Rhamnetin 3-mannosyl-(1-2)-alloside (visualization software using LigPlot [73])

Tables

Table 1 List of Potential Protein Target Related to COVID-19

Virus-based protein			Host-based protein			
PDB/Uniprot ID	**Protein**	**Reference**	**Uniprot name**	**Uniprot ID**	**Protein**	**Reference**
6LU7:A	3CLPro	[5]	ACE2	Q9BYF1	ACE2	[53]
PLpro_SARS-CoV-2	PLPro	[5]	AKT1	P31749	AKT	[47]
K4LC41						
			PYRD	Q02127	DHODH	[54]
yp_009725307.1	RdRp	[5]	PPIA	P62937		
			PPIG	Q13427		
6M0J:E	Spike-ACE2	[5]	FKBP5	Q13451		
6LZG:A			FKBP4	Q02790		
6VSB			FKBP2	P26885		
6M0J:A						
			CYP5	P52013	PPIASE	[50]
			FKB1B	P68106		
			PPIB	P23284		
			PPIC	P45877		
			PPIH	O43447		
			FKB1A	P62942		
			IL6RB	P40189	IL-6	[52]

Table 2 List of Potential Drug Explored from SuperTarget Database

Drug	Protein Target
Moexipril hydrochloride	ACE2
Arsentrioxide	AKT1
Arthrocine	
Celecoxib	
Erlotinib	
Gefitinib	
Imatinib Mesylate	
Lapatinib ditosylate	
Simvastatin	
Sorafenibum	
Sunitinib	
Atovaquone	PYRD, PPIA, PPIG
Essigsaeure	
Huanghuahaosu	
Hydroxycinchophene	
Leflunomide	
Rapamycin	FKBP5, FKBP4, FKBP2, FKB1B, FKB1A
Athylenglykol	FKBP4
Methylsulfinylmethane	
Dithiothreitol	CYP5_CAEEL
Carboxypyrrolidine	PPIB, PPIC, PPIH
Pimecrolimus	FKB1A
Tacrolimus	
Thiabendazole	

Table 3 The performance of each model calculated using 30% of dataset that was excluded from training set

Method	Performance Measure	Value
Multilayer Perceptron (MLP)	AUC	0.98405
	F-measure	0.98254
	Precision	0.96628
	Recall	0.99936
	Accuracy	0.98321
Random Forest (RF)	AUC	0.98734
	F-measure	0.98608
	Precision	0.97255
	Recall	1
	Accuracy	0.98665
Support Vector Machine (SVM)	AUC	0.99919
	F-measure	0.99911
	Precision	0.99847
	Recall	0.99975
	Accuracy	0.99915

Table 4 The Predicted Potential Compounds Targeting 3CLPro, PLPro, and RdRp

No	Protein Target	Herbal Compound
1.	3CLPro	Amaranthine, Methylthio, Arabinopyrano, Peonidin-3, Quercetin-3, Sinigrin, Heperidine, Myricetin-3, (+)-2,3-Dihydro-9-hydroxy, Cyanidin-3, Scutellarein, Spiraeoside, Glucoputran, Isoforskolin, Kaempferol-3
2.	PLPro	Methylthio, Sinigrin, Glucoputran
3.	RdRp	Methylthio, Arabinopyrano, Peonidin-3, Quercetin-3, Theviridoside, Sinigrin, Heperidine, Myricetin-3, (+)-2,3-Dihydro-9-hydroxy, Cyanidin-3, Catalpol, Scandoside, Scutellarein, Spiraeoside, Geniposide, Oleoside, Majoroside, Glucoputran, Isoforskolin, Kaempferol-3

Table 5 The top-30 of hit compounds from LBDD methods

No	Name Compound	No	Name Compound
1	Kaempferol 3-alpha-D-arabinopyranoside	16	Catalpol
2	Isoforskolin	17	Cyanidin-3-sophoroside-5-glucoside
3	Glucoputranjivin	18	(+)-2,3-Dihydro-9-hydroxy-2 [1-(6-sinapinoyl)beta-D-glucosyloxy-1-methylethyl]-7H-propanoat
4	Loganic Acid	19	Myricetin 3-glucoside
5	Majoroside	20	Hesperidin
6	Oleoside	21	Azadirachtin A
7	Geniposide	22	1-Caffeoyl-beta-D-glucose
8	Glucobrassicin	23	Sinigrin
9	Spiraeoside	24	Theviridoside
10	Alizarin	25	Quercetin 3-(2G-rhamnosylrutinoside)
11	Morindone	26	Peonidin 3-(4'arabinosylglucoside)
12	Casuarinin	27	trans-p-Sinapoyl-b-D-glucopyranoside
13	Scutellarein-6,4'-dimethyl ether-7-(6"-acetylglucoside)	28	6,8-Di-C-beta-D-arabinopyranosyl apigenin
14	Scandoside methyl ester	29	8-Methylthio-octyl glucosinolate
15	beta-Glucogallin	30	Amaranthine

Table 6 Molecular docking results of 14 hit (overlapped) compounds against the main protease of SARS-CoV-2

No	Compound name	Binding Energy (ΔG) (kcal/mol)	Sources
1	Cyanidin-3-sophoroside-5-glucoside	-6.52	*Brassica Oleracea* [79]*; Ipomoea Batatas*[80]*; Raphanus Sativus* [81]
2	Geniposide	-7.04	*Gardenia jasminoides* [82]
3	Hesperidin	-8.72	*Psidium guajava* [65] *Citrus aurantium* [83]
4	Isoforskolin	-6.88	*Coleus forskohlii* [84]
5	Kaempferol 3,4'-di-O-methylether (Ermanin)	-8.51	*Zingiber aromaticum* [85]
6	Majoroside	-7.03	*Plantago major* [86]
7	Myricetin-3-glucoside	-8.26	*Moringa oleifera* [87]
8	Oleoside	-6.52	*Oleaceae familia (e.g. Jasminum sambac)* [88]
9	Peonidine 3-(4'-arabinosylglucoside)	-8.52	*Ipomoea fistulosa* [89]
10	Quercetin 3-(2G-rhamnosylrutinoside)	-8.56	*Clitoria Ternatea* [90]
11	Rhamnetin 3-mannosyl-(1-2)-alloside	-8.48	*Moringa oleifera* [91] *Cassia alata* [92]
12	Sinigrin	-5.19	*Brassica nigra* [93]
13	Spiraeoside	-7.97	*Filipendula ulmaria* [94]
14	Theviridoside	-7.13	*Thevetia peruviana* [95]
15	Lopinavir	-9.41	Antiviral drug (positive control)