



OPEN ACCESS

EDITED BY

Parimal Kar,
Indian Institute of Technology Indore,
India

REVIEWED BY

Kemal Yelekci,
Kadir Has University, Türkiye
Md. Fulbabu Sk,
University of Illinois at Urbana-Champaign,
United States

*CORRESPONDENCE

Junjian Hu,
✉ hujunjian79@t63.com
Abdul Wadood,
✉ awadood@awkum.edu.pk

SPECIALTY SECTION

This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 04 October 2022

ACCEPTED 11 January 2023

PUBLISHED 07 March 2023

CITATION

Samad A, Ajmal A, Mahmood A, Khurshid B, Li P, Jan SM, Rehman AU, He P, Abdalla AN, Umair M, Hu J and Wadood A (2023), Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation. *Front. Mol. Biosci.* 10:1060076. doi: 10.3389/fmolb.2023.1060076

COPYRIGHT

© 2023 Samad, Ajmal, Mahmood, Khurshid, Li, Jan, Rehman, He, Abdalla, Umair, Hu and Wadood. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation

Abdus Samad¹, Amar Ajmal¹, Arif Mahmood^{2,3}, Beenish Khurshid¹, Ping Li⁴, Syed Mansoor Jan¹, Ashfaq Ur Rehman⁵, Pei He⁶, Ashraf N. Abdalla⁷, Muhammad Umair⁸, Junjian Hu^{9*} and Abdul Wadood^{1*}

¹Department of Biochemistry, Abdul Wali Khan University, Mardan, KPK, Pakistan, ²Center for Medical Genetics and Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China, ³Institute of Molecular Precision Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China, ⁴Institutes of Biomedical Sciences, Shanxi university, Taiyuan, China, ⁵Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, United States, ⁶Department of Obstetrics and Gynecology, Nanfang Hospital, Southern Medical University, Guangzhou, China, ⁷Department of Pharmacology and Toxicology, College of Pharmacy, Umm Al-Qura University, Makkah, Saudi Arabia, ⁸Department of Life Sciences, School of Science, University of Management and Technology (UMT), Lahore, Pakistan, ⁹Department of Central Laboratory, SSL Central Hospital of Dongguan City, Affiliated Dongguan Shilong People's Hospital of Southern Medical University, Dongguan, China

The new coronavirus SARS-COV-2, which emerged in late 2019 from Wuhan city of China was regarded as causing agent of the COVID-19 pandemic. The primary protease which is also known by various synonymous i.e., main protease, 3-Chymotrypsin-like protease (3CL^{PRO}) has a vital role in the replication of the virus, which can be used as a potential drug target. The current study aimed to identify novel phytochemical therapeutics for 3CL^{PRO} by machine learning-based virtual screening. A total of 4,000 phytochemicals were collected from deep literature surveys and various other sources. The 2D structures of these phytochemicals were retrieved from the PubChem database, and with the use of a molecular operating environment, 2D descriptors were calculated. Machine learning-based virtual screening was performed to predict the active phytochemicals against the SARS-CoV-2 3CL^{PRO}. Random forest achieved 98% accuracy on the train and test set among the different machine learning algorithms. Random forest model was used to screen 4,000 phytochemicals which leads to the identification of 26 inhibitors against the 3CL^{PRO}. These hits were then docked into the active site of 3CL^{PRO}. Based on docking scores and protein-ligand interactions, MD simulations have been performed using 100 ns for the top 5 novel inhibitors, ivermectin, and the APO state of 3CL^{PRO}. The post-dynamic analysis i.e., Root means square deviation (RMSD), Root mean square fluctuation analysis (RMSF), and MM-GBSA analysis reveal that our newly identified phytochemicals form significant interactions in the binding pocket of 3CL^{PRO} and form stable complexes, indicating that these phytochemicals could be used as potential antagonists for SARS-COV-2.

KEYWORDS

SARS-CoV-2, COVID, 19, machine learning, molecular docking, MD simulation, Corona virus

1 Introduction

SARS-CoV-2 is a single-strand RNA, positive sense, and enveloped beta coronavirus that causes respiratory, nervous, hepatic, and human gastrointestinal diseases (Tahir ul Qamar et al., 2020). Wuhan, a city in China, was the first city to be infected by the virus in December 2019 (Zhu et al., 2019; Zhou et al., 2020). COVID-19 outbreak was declared a pandemic by the World Health Organization (WHO). The infection spreads rapidly across the World. By the end of October 2020, more than 60 million people were infected by COVID-19, resulting in more than 1.4 million fatalities. The number of patients and fatalities was rising, posing a major threat to global health. High temperature, coughing, shortness of breath, and severe cases that can result in renal failure and even death are some of the symptoms of COVID-19 infections (Rothan and Byrareddy, 2020; Asif et al., 2022), until now, there is no effective treatment available yet.

SARS-CoV-2 is a member of the beta coronavirus family (Marty and Jones, 2020), usually, during the process of transcription, beta coronaviruses produce an 800 kDa polypeptide (Xu et al., 2020). The genome of the novel SARS-CoV-2 was recently sequenced and compared with those of existing coronaviruses (CoVs) by Wu et al. who identified that the novel SARS-CoV-2 belonged to the β -CoVs, which were initially discovered in bats and have now evolved to infect humans (Wu et al., 2020a). The SARS-CoV-2 genome is approximately 30 kb in size and is comprised of at least six open reading frames (ORFs) which are responsible for encoding the whole proteome of the virus. The coding RNA contains the structural, non-structural protein (Nsp) coding regions and the accessory protein-coding region (Durojaiye et al., 2020). The genes on the 3'-terminus encode the four structural proteins including the spike protein, membrane, envelope, nucleocapsid, and many accessory proteins. The membrane, envelope, and nucleocapsid protein protect the virus before entering the host cell. The Spike protein of SARS-CoV-2 comprises S1 and S2 subunits. The receptor-binding domain is a part of the S1 subunit that plays role in the attachment of the virus with the receptor while viral cell membrane fusion is mediated by the S2 subunit, thus facilitating the virus entry (Alanagreh et al., 2020; Jackson et al., 2021). The SARS-CoV-2 virus's replication and ability to spread are facilitated by numerous crucial proteins and enzymes. Two essential proteases, main protease (3CL^{PRO}) and papain-like protease (PLpro) are necessary for viral replication (Huang et al., 2020; Mouffouk et al., 2021). The non-structural proteins nsp1, nsp2, and nsp3 are known to be cleaved by PLpro, while the remaining 13 are cleaved by 3CL^{PRO} (Klemm et al., 2020). The 3CL^{PRO} cleaves polypeptide sequences after a glutamine residue, making it a perfect drug target as no human host-cell proteases with this cleavage specificity are identified (Hilgenfeld and Hilgenfeld, 2014; Ullrich and Nitsche, 2020).

The structure of the 3CL^{PRO} comprises three important domains, domain-I ranges from 8–101, while domains-II corresponds to position 102–184, followed by the connecting loop from 185–200, which links domain-II and domain-III, domain-III has a total number of 103 residues which lies after the connecting loop from 201–303 (Wu et al., 2020b). Furthermore, the His-41 and Cys-145 form an essential catalytic dyad (Kneller et al., 2020). Small compounds that target conserved viral proteases, such as the major protease, may thus be able to inhibit crucial phases of the SARS-CoV-2 life cycle while causing few adverse effects (Mengist et al., 2021). Approved drugs have been developed for viral infections such as those caused by Hepatitis C virus and human immunodeficiency virus for the target's serine proteases and

aspartyl protease respectively which employ that viral proteases are well-established therapeutic targets (Agbowuro et al., 2018). Antiviral drugs are required in this situation to prevent infection in high-risk populations as well as to treat infected patients. Developing inhibitors that stop coronavirus replication can recover millions of people globally. In the clinical investigations, efforts to repurpose the majority of approved drugs have discovered several promising candidates (such as remdesivir and hydroxychloroquine) but these drugs had little to no effect on mortality and the duration of hospital stay (Luttens et al., 2022). Hence, it is crucial to find new drug candidates that would target various SARS-CoV-2 proteins for increased COVID-19 therapeutic effectiveness (Elmaaty et al., 2022). Despite the significant cost and time required for the development of the new drug, clinical trials only yield a 13 percent success rate, while in 40%–60% of cases, drugs failed to reach the market because of the lack of optimum pharmacokinetic properties (Gurung et al., 2021).

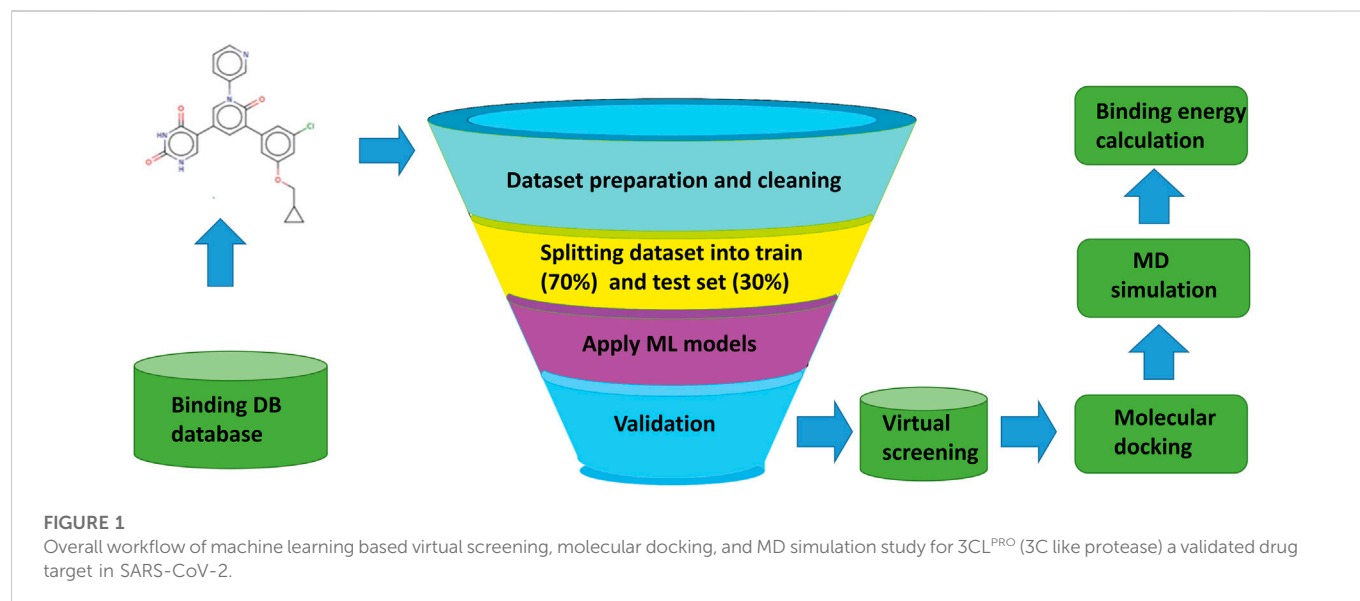
The use of computer-aided drug discovery (CADD) tools helps to accelerate the process of drug discovery and to reduce costs (Macalino et al., 2015). In addition, the advent of supercomputing facilities, algorithms, and tools has enhanced lead identification in pharmaceutical research (Macalino et al., 2018). Artificial intelligence (AI) and machine learning approaches have substantially assisted the analysis of pharmaceutical-related large data in the drug discovery process (Floresta et al., 2022). Furthermore, the structure-based drug development method is specific and successful in identifying lead compounds and optimizing them, and it has aided in the understanding of disease at the molecular level (Yang et al., 2022). In the current study, we employed different machine learning (ML) models for the virtual screening of phytochemicals against the 3CL^{PRO} drug target in SARS-CoV-2. The active hits obtained from ML-based were passed through an electronic filter called PAINS filter and their ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties were examined. The active phytochemicals that passed through the PAINS filter and have enhanced properties were further considered for the molecular docking analysis. Furthermore, the stability and binding energy of these compounds in the active site of 3CL^{PRO} were investigated by 100 ns of MD simulations. Based on our findings we suggest these phytochemicals as potent inhibitors of SARS-CoV-2 3CL^{PRO}. *In vitro* evaluation of these compounds, is essential for the understanding of their action and mechanism to cope with such a pandemic.

2 Methodology

The overall workflow of the current study, from the collection and preparation of the dataset of active and inactive compounds, screening of compounds, molecular docking, and binding energy calculations are represented in Figure 1.

2.1 Preparing and cleaning the dataset

From the binding DB database (Sandhu et al., 2022) a total of 101 molecules were retrieved for 3CL^{PRO} (3C like protease) a drug target in SARS-CoV-2. A total of 500 decoys molecules, which are considered to be inactive, were generated using the DUDE database (Mysinger et al., 2012). Out of the total 601 compounds (Supplementary Table S1), 101 compounds from the binding DB



database were labeled as “1” active, and the 500 decoys were labeled as “0” inactive. The Pandas library of python was used for data preprocessing and data cleaning (Santos et al., 2020). The dataset was split into train set (70%) and a test set (30%).

2.2 Features calculation

The 2D features of all the compounds were calculated using MOE (2016) software (Wadood et al., 2022a). Total 206 features were calculated. Feature with 0 or null values were removed from the dataset to reduce the computation time.

2.3 Principal component analysis (PCA)

The dataset was uploaded to iRaPCA v1.0 implemented in the LideB tool in CSV format. The optimum subsets of descriptors were selected from the dataset. The dimensionality was reduced by performing the PCA. The process is based on the principle of feature bagging (Prada Gori et al., 2022). The conventional feature extraction and data representation method used extensively in the fields of pattern recognition is principal component analysis (PCA), generally called as Karhunen-Loeve expansion. PCA is a method for reducing high-dimension data to low-dimension while preserving the majority of the relevant data. The main benefits of PCA are its low noise sensitivity, lower capacity and memory requirements, and increased performance (Karamizadeh et al., 2013).

2.4 Machine learning models

2.4.1 K nearest neighbor model

The distance-based classification algorithm is called k-Nearest Neighbors (kNN), which is an effective and simple machine learning algorithm widely used for the classification of active and inactive

compounds in the dataset (Wadood et al., 2022b). The accuracy of the kNN model depends entirely on the quality of the data. One of the most difficult parts of KNN is figuring out how many neighbors to consider. The KNN can be used for both classification and regression (Sarker, 2021a).

2.4.2 Support vector machine (SVM)

SVM is generally used for the classification of data. SVM is based on the principle of calculating margins between two classes. This classifier reduced the error by drawing the margins in a manner where the distance between the margin and the classes is as large, as possible (Noreen et al., 2016). The SVM classifier depends on the kernel function and is more effective for high-dimensional data classification. When the dataset contains additional noise, such as overlapping target classes, SVM does not perform effectively (Sarker, 2021b).

2.4.3 Random forest

Random forest (RF) is an ensemble algorithm made up of several decision trees, similar to how a forest is made up of many trees (Breiman, 2001). To train, the decision trees of a random forest various subsets of the training dataset are used. To classify a new sample, the sample's input vector must be passed down from each decision tree of the forest. This algorithm classifies the data using majority voting. In terms of performance, RF performs better than a decision tree. For huge datasets, it works effectively. The classifier also calculates which variables or attributes are most significant in the classification (Ul Hassan et al., 2018). The sklearn library of python was used for developing the three machine learning models.

2.4.4 Naïve bayes

The naive Bayesian algorithm is based on the Bayes theorem and is a reliable classification method. A data set can be classified by NB classifier assuming that every feature contributes equally and independently (Patel et al., 2020). In this study, the NB classifier was built using python v.3.9.

2.4.5 Cross-validation and performance evaluation

We used 10-fold cross-validation in this study. The performance of the models was accessed by using several statistical parameters including accuracy, sensitivity, specificity, F1 score, MCC (Ahmad et al., 2021).

2.5 Virtual screening of the asian phytochemicals

A list of Asian plants with notable therapeutic properties was compiled, and then a thorough literature search was performed to determine the phytochemical contents. The compound collection was carried out using Google Scholar, PubMed, MEDLINE, and other web-based resources. A total of 4,000 phytochemical libraries was generated, and the 2D structure of these phytochemicals was retrieved from the PubChem database. Before adding to the library all these phytochemicals were cleaned and energy minimized using the mmff94 force field.

2.6 PAIN filter

Pre-filtering large databases using appropriate molecular properties is a typical approach to reduce computing and get rid of unwanted compounds (Baell and Holloway, 2010). All the active hits were filtered by an online tool PAINS (Wadood et al., 2022c) and only those compounds were further selected for docking that was passed from the PAINS filters.

2.7. Molecular docking study

2.7.1 Preparation and validation of target protein

The 3D structure of SARS-CoV-2 3CL^{PRO} (PDB ID: 6LU7; Resolution: 2.16 Å; Organism: SARS-CoV-2; Method: X-ray diffraction) was downloaded from the RCSB Protein Data Bank (Hatada et al., 2020). There are two chains in the crystal structure: A and C. The macromolecule chain A was chosen as the target receptor. Pymol was used to remove water molecules and heteroatoms from the protein structure (Janson et al., 2017). The structure was then energy minimized using ff14sb implemented in the molecular operating environment (MOE) (Ashraf et al., 2021). The PROCHECK (Laskowski et al., 1996) and ERRAT (Colovos and Yeates, 1993) tools from the Structural Analysis and Verification Server (SAVES) (<http://nihserver.mbi.ucla.edu/SAVES>) were used to validate the crystal structure. The stereo chemical quality of the protein structure was evaluated using PROCHECK.

2.7.2 Molecular docking protocol

All the phytochemicals predicted as active by the machine learning method were docked into the active site of a SARS-CoV-2 3CL^{PRO} for molecular interaction studies. The crystal structure of the SARS-CoV-2 3CL^{PRO} (PDB ID: 6LU7) is complex with an N3 inhibitor was retrieved from the PDB database. The Inhibitor N3 is linked to the protease at site one of this crystal structure, which contains five cavities for ligand binding (Das et al., 2021). We used the N3 binding site (site 1) for virtual screening of these phytochemicals' library. For the molecular docking study, MOE

v2016 was used to run a docking protocol using rigid and ligand-based docking parameters. The Triangular Matching docking method (default) was used and a total of ten poses were generated for each Phytochemical (Thuy et al., 2020). The best S score hits against 3CL^{PRO} were considered for the molecular interactions study and their 3D images were generated by PyMol software. A total of 05 top-ranked compounds were shortlisted for further molecular dynamic simulations analysis based on the docking score. These phytochemicals are structurally diverse, effective, and new inhibitors for the main protease, according to the docking score, binding mode, and visual ligand interaction.

2.8 MD simulations

Molecular dynamics simulation is a powerful tool to understand the dynamics and interaction behavior of the reference complex and the selected top hits were used. The ff14SB protein force field in Amber 20 package was employed (Salomon-Ferrer et al., 2013a). For solvation of each system, the tip3p water model with box dimension 8.0 was used. All of the systems were adequately solvated and neutralized by adding four Na⁺ ions to counterbalance the charges on the systems. Afterward, energy minimization for 6,000 steps of neutralized complexes was carried out using the steepest descent minimization algorithm, then progressively heated to 300 K before equilibrating density for 2 ns with weak constraints. The whole system was equilibrated at constant pressure for another 2 ns. A Langevin thermostat was used to control the temperature 300 K. Further, a 100-ns MD was performed on the equilibrated systems. For long-range electrostatic interactions, Particle Mesh Ewald (PME) algorithm was used (Darden et al., 1998). For covalent bonds including hydrogen, the SHAKE algorithm was utilized. Finally, a 100 ns MD simulation of all equilibrated complexes at constant pressure and temperature was carried out by using PMEMD.cuda (Salomon-Ferrer et al., 2013b).

2.9 DCCM

The dynamic cross-correlation analysis is useful for explaining the correlation among the residues represented by a three-dimensional matrix. The cross-correlation was calculated by the formula (Junaid et al., 2018)

$$C_{ij} = \langle \Delta r_i \Delta r_j \rangle / (\langle \Delta r_i^2 \rangle \langle \Delta r_j^2 \rangle)^{(1/2)} \quad (1)$$

Where the mean position of *i*th and *j*th atom is represented by Δr_i , Δr_j , respectively. Where the angular brackets are used to measure the average time of the entire trajectories produced as a result of MD simulations. Positive Correlated movement such as movement in the same direction is represented by the positive value of C_{ij} , while the negative value of C_{ij} reflects strong anti-correlation movements between the residues. Cpptraj was used to perform DCCM analysis while origin 2021 was used for graphical representations (Perez-Lemus et al., 2022).

2.10 Binding affinity calculations

To study the interaction between protein and ligand, binding free energy calculations play an important role. Using MMPBSA. PY

TABLE 1 Train and test set used in the study.

Dataset	Inhibitors	Non-inhibitors	Total
Train	32	388	420
Test	33	148	181

script, the binding free energy between main protease and phytochemicals inhibitors was calculated (Gul et al., 2021). The following equation was used to calculate the free energy of each energy term:

$$\Delta G_{bind} = \Delta G_{complex} - [\Delta G_{receptor} + \Delta G_{ligand}] \quad (2)$$

In the equation, ΔG_{bind} represents the total binding free energy, $\Delta G_{complex}$ denotes the free energy of complex, $\Delta G_{receptor}$ and ΔG_{ligand} represents the free energy of receptor protein and ligand respectively. The following equation was used to calculate the individual free energy of complex, protein and ligand.

$$G_X = E_{MM} - (TS) + (G_{solvation}) \quad (3)$$

Where x denotes complex, protein or ligand, the average molecular mechanic potential in a vacuum is given by E_{MM} , the entropic and temperature contribution is represented by TS, while the free energy of the solvation is given by $G_{solvation}$.

3 Results

3.1 Data preparation

A total of 101 molecules were retrieved from the binding databank database for 3CL^{PRO} a drug target in SARS-CoV-2. The 101, molecules were categorized as active molecules. The remaining 500 decoys molecules were labeled as inactive. The dataset was split into a train set (70%) and test set (30%). Out of the total 601 molecules, the train set contains 420 compounds while the test set contains 181 compounds. The active and inactive compounds of the train and test set are present in Table 1.

3.2 Principle component analysis

Total 208 2D features were calculated with the help of MOE software. The feature with 0 values were removed. As, not every extracted feature will necessarily depict the optimal properties of molecules. Therefore, optimization was carried out to get rid of duplication. Additionally, after applying the PCA the features that have higher significance were used to train the models (Araki et al., 2016). After applying PCA the data size (N) of the dataset was decreased. To evaluate how the PCA manages to maintain the dominant properties throughout the classification tasks. The models were generated by using the entire dataset without optimum features selection and the performance of the models was evaluated. It was found that the accuracy of SVM was very low as 61% and the MCC was 0.27. The accuracy of KNN model was 70% with an MCC value of 0.58 while the accuracy of RF model was 90% with an MCC value of 0.78. However, after the optimum features selection and the reduction of

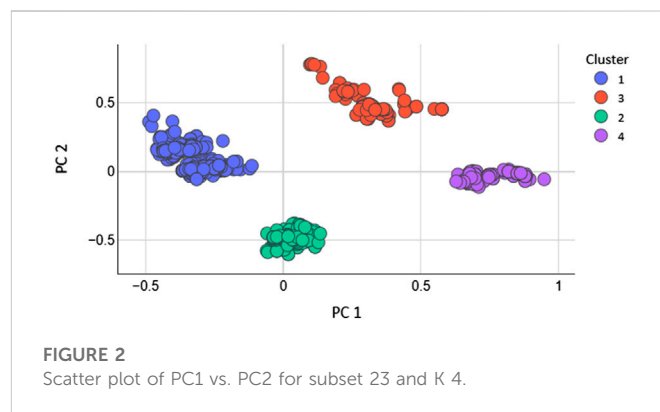


FIGURE 2
Scatter plot of PC1 vs. PC2 for subset 23 and K 4.

the dimension of the dataset the performance of all the models was greatly improved. If we want to reveal variance in a dataset having x-y coordinates, PCA finds a new coordinate system in which x, y coordinates have a different value. A new coordinate is created by the axes PC1 and PC2. These are combinations of the x-y coordinate system. Figure 2 shows the scatter plot of PC1 vs. PC2 for K = 4.

3.2.1 Chemical space and diversity analysis

The machine learning model's accuracy is predicted by the chemical diversity of the samples from the training and test sets. The applicability of machine learning models is restricted by a small number of samples. As a result, in the present study's physiochemical distribution analysis of the training set and test set for the molecular weight (MW) and LogP was conducted (Figures 3, 4) with MW ranging from 50 to 800 Da and LogP ranging from -2 to 15.

3.3 Models generation and validation

Machine learning algorithms such as kNN, SVM, RF and GNB were used for the classification of the active inhibitors against 3CL^{PRO}. The sklearn library of python was used for developing the models. All the models were trained on the dataset downloaded from the binding DB database. The performance of the models was accessed by using a number of statistical parameters including accuracy, sensitivity, specificity, and MCC. Table 2 displays the over-all performance of the models on the train set while Table 3 displays the performance of all the models on the test set.

Compared to other machine learning models random forest model achieved better accuracy and MCC value. Model performance is proportional to the area under the curve (AUC). RF has the highest AUC, followed by SVM on the training and test set Figures 5, 6. Further, we used RF model to classify the active phytochemicals against the 3CL^{PRO} enzyme. Out of 4,000 phytochemicals, a total of 26 phytochemicals were predicted as active against the 3CL^{PRO}.

3.4 PAIN filter

Using the online PAINS tool all the hits were examined for their ADMET (absorption, distribution, metabolism, excretion, and toxicity) (Supplementary Table S2) properties. A total of seven compounds were passed from the PAINS filter and only two

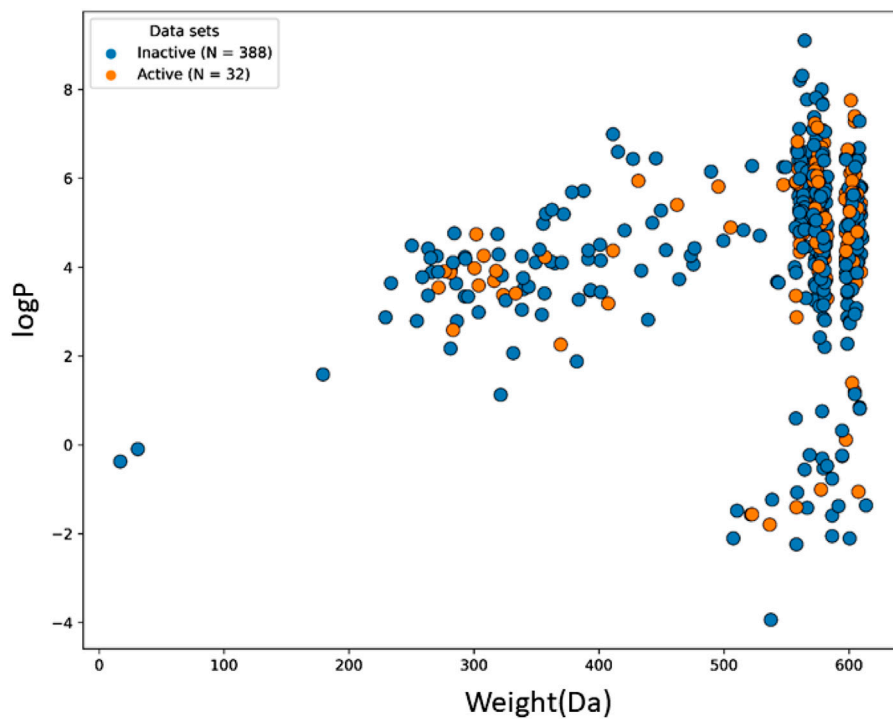


FIGURE 3

The chemical space and diversity distribution of the train set. The molecular weight and LogP define the chemical space.

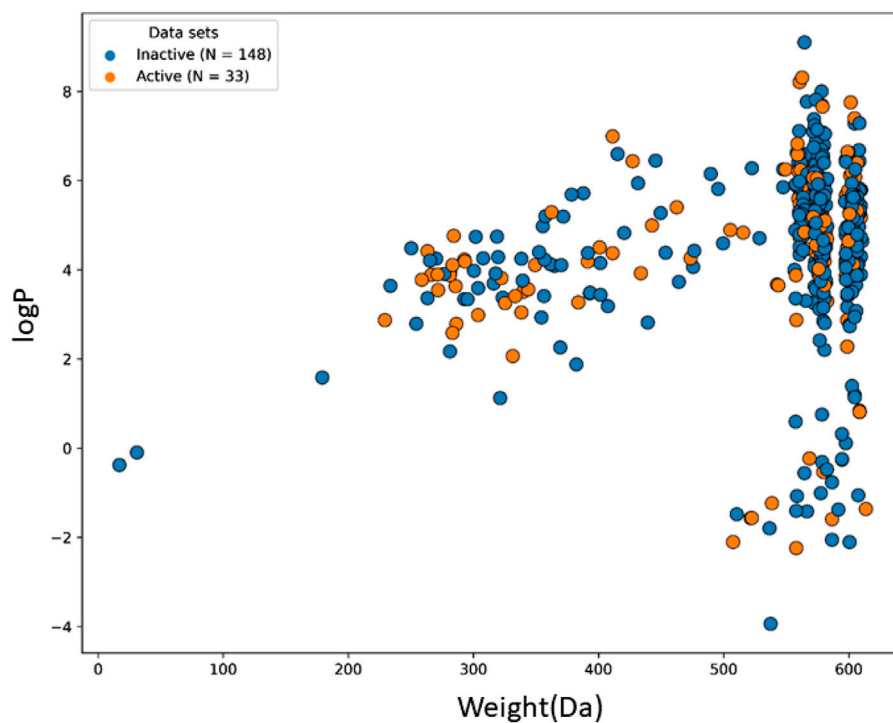


FIGURE 4

The chemical space and diversity distribution of the test set. The molecular weight and LogP define the chemical space.

TABLE 2 Overall performance of machine learning models on the train set.

Model	Accuracy (%)	Sensitivity	Specificity	MCC
KNN	97	0.88	0.99	0.91
SVM	98	0.90	0.99	0.93
RF	98	0.97	0.99	0.96
GNB	94	0.83	0.96	0.79

TABLE 3 Performance of models on the test set.

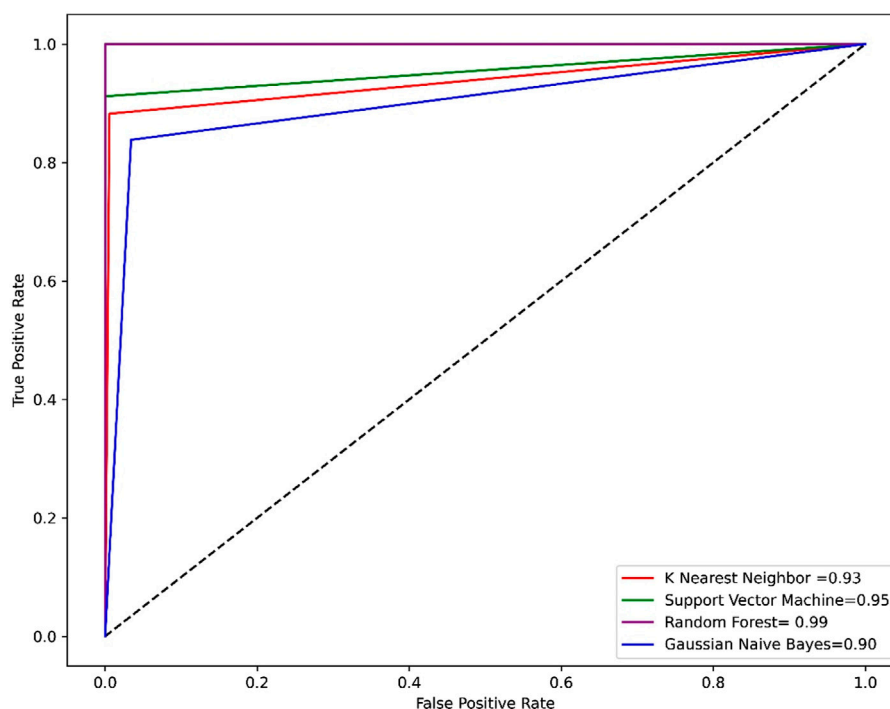
Model	Accuracy (%)	Sensitivity	Specificity	MCC
KNN	94	0.75	0.98	0.78
SVM	96	0.82	0.99	0.87
RF	98	0.95	0.99	0.95
GNB	96	0.86	0.98	0.85

compounds were out of the limit. The structure of compound along with IUPAC name of the compounds passed from the PAIN filter are given in Table 4.

3.5 Molecular docking analysis

The hits obtained from ML based virtual screening were further used for molecular docking study. The crystal structure of the SARs-CoV-2

3CL^{PRO} (PDB ID: 6LU7) is complex with an N3 inhibitor was retrieved from the PDB database. PROCHECK tool was used to assess the 3D model's quality of the 3CL^{PRO} structure using the Ramachandran plot (Figure S2a). The Ramachandran plot for the 3CL^{PRO} structure showed that 84.5% of residues were in the most favored region, while 14.3% were in the additional allowed region, 1.1% residues were in the generously allowed region and 0% residues were in the disallowed region demonstrating the high quality of the 3CL^{PRO} structure. For non-bonded atomic interactions, ERRAT is also known as the "overall quality factor," with higher scores reflecting the high quality. For a high-quality model, the accepted range is > 50 (Messaoudi et al., 2013) The ERRAT server predicted an overall quality factor of 85.90 for the 3CL^{PRO} structure used in our study (Figure S2b). The interaction of top hits and the reference compound were analyzed, and it was found that all of the compounds have potent inhibitory effects on 3CL^{PRO}. In order to study the interactions of these compounds in detail, the 3D visualization and compound interaction analysis was carried out. According to the interaction details Table 5, Compound 1 has stronger interaction among all of the docked compounds, it has 04 hydrogen bond donor interactions with the active site residues i.e., CYS145, SER46, and MET49, with four hydrogen bond acceptor interactions with HIS41, LEU141, and HIS163, along with one π -stacking interaction with residue THR25 with the docking score of -12.0321 . Similarly, the interactions details of Compound 2 reveal that it shares five hydrogen bond donor interactions with key active site residues of the main protease i.e., THR26, MET49, ASN142, CYS145, and MET165, and two π -H interactions with residues with SER46 and THR90 respectively. The interaction table indicates that Compound 3 forms 6 hydrogen bond interactions with His41, Met49, Cys145, His163, and Gln189, and one π -H interaction with Leu 141. Compound 4 shows 04 hydrogen bond donor interactions with the catalytic residues i.e., Thr 25, Thr26, Met49, and His164, and one

**FIGURE 5**

The ROC-AUC curve of all the models on the train set. The graph shows the TP against FP rate.

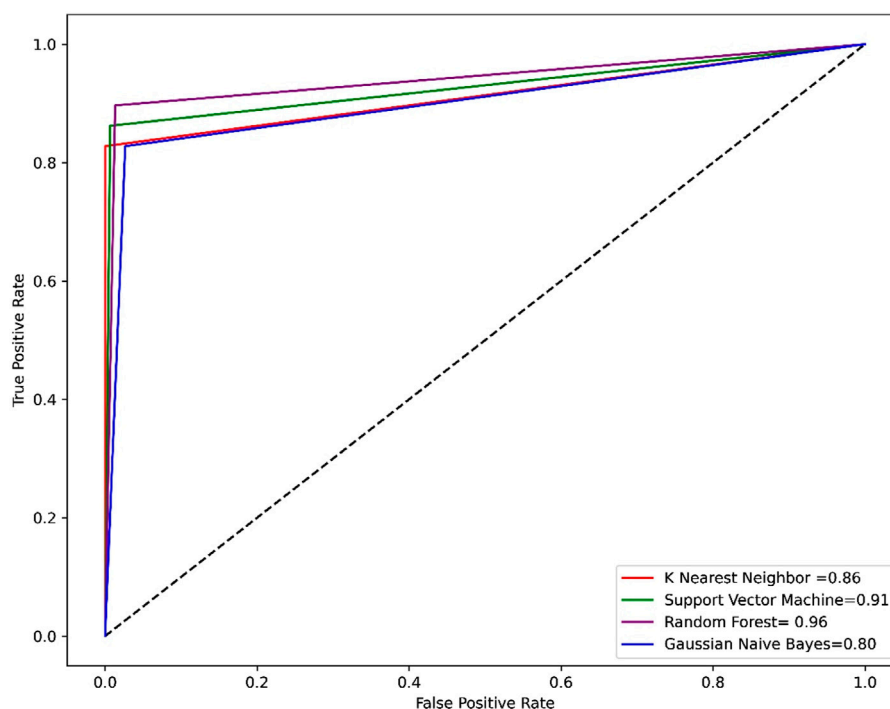


FIGURE 6
ROC-AUC curve of all the models on the test set. The graph shows the TP against FP rate.

hydrogen bond acceptor interaction with Gly143, with one π - π interaction with residue His41. Afterward, we analyzed the interaction of Compound 5, the finding of interaction analysis indicates that Compound 5 interacts *via* four hydrogen bond donor interactions with the key residues including Thr26, Met49, Asn142, and Gln189, while Thr26, and His41 were found in hydrogen bond donor interactions with Compound 5 with a docking score of -10.7164 . It has recently been demonstrated that ivermectin inhibits SARS-CoV-2 by up to 5000-fold *in vitro* with an IC₅₀ value of $\sim 2 \mu\text{M}$ (Jan et al., 2021; Kaur et al., 2021). In the docking study, ivermectin was selected as a standard reference inhibitor. The interaction details for the control compound are listed in Figure 7H. The control compound forms 05 hydrogen bonds with the key catalytic residues of main protease Asn119, Cys145, and Met165. The co-crystallized ligand (PDB ID: 6LU7) was removed from the active site and re-docked into the binding site of 3CL^{PRO} in order to evaluate the precision of MOE-Dock. The RMSD value between the top-ranked docked conformation and the co-crystallized ligand was 0.6532 (Figure S3), indicating the strong accuracy of the MOE-Dock procedure (Wadood et al., 2022c).

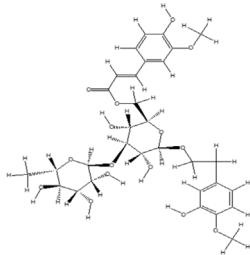
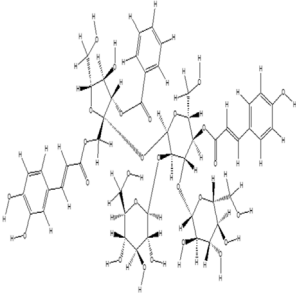
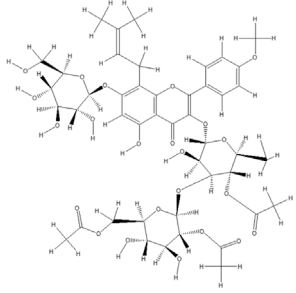
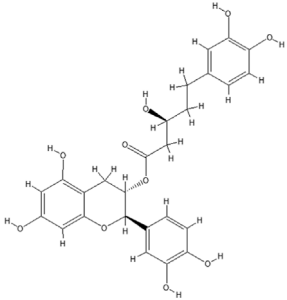
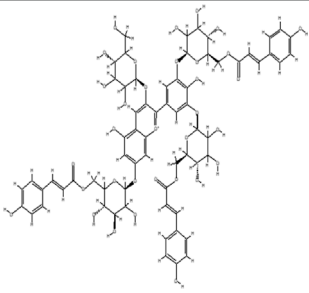
3.6 MD simulation analysis

3.6.1 Root means square deviation

Root means square deviation (RMSD) analysis was performed to calculate the stability of the top five phytochemicals and reference compound (ivermectin) in the active site of the main protease. We examined and compared the stability of these compounds with the reference and APO protein. The RMSD finding indicates that all these

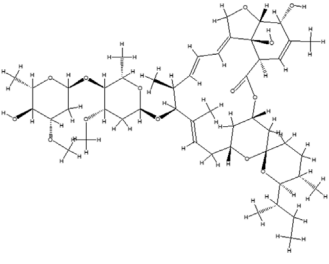
five phytochemicals show stable behavior but some minor deviation. For all the systems the averaged RMSD ranges between 1 and 3 Å. The average RMSD for ivermectin was initially 2.0 Å. Then a small increase was observed in RMSD up to 40 ns, soon after reaching 40ns the system acquired stability and remained stable for the rest of the simulation period. The complex Compound 1 shows significant stability as can be observed, however after 60 ns, the system briefly displayed a small variation. Then the system achieved stability and moved into the production phase. For Compound 2, RMSD reveals that the system shows highly stable behavior in the entire period of simulation, at 20ns minor fluctuations from its mean position were observed, afterward, the system gained stability and no more significant deviations were observed with the average RMSD value of 1.7 Å. For complex Compound 3, the system initially shows stable behavior, at around 15 ns a gradual increase in the RMSD curve was observed followed by a slight decrease in the RMSD curve at 20 ns. After that the system equilibrates with an average RMSD value of 2.1 Å as shown in Figure 8. The Compound 4 complex RMSD analysis reveals that the system initially shows an increase in the RMSD curve but soon after reaching 25 ns the system equilibrates and no significant fluctuations were observed for the rest of the simulation period which indicates the stable binding of Compound 4 compound in the catalytic pocket of 3CL^{PRO} with the average RMSD value of 1.4 Å. Afterward, we analyzed the RMSD of Compound 5 in the active site of 3CL^{PRO}, the RMSD curve of the corresponding complex has minor fluctuations at different time intervals, with an average RMSD value of 1.7 Å. The backbone RMSD for the phytochemical bound 3CL^{PRO} was slightly lower than the control indicating the stable binding of these phytochemicals which was further validated by RMSF and MM-GBSA analysis.

TABLE 4 PubChem ID of the compound, IUPAC name of compound and the PAIN filter result of the compounds.

Compound ID	Structure	IUPAC name	PAINS filter
91895373 (Compound 1)		[(2R,3R,4S,5R,6R)-3,5-dihydroxy-6-[2-(3-hydroxy-4-methoxyphenyl) ethoxy]-4-[(2S,3R,4R,5R,6S)-3,4,5-trihydroxy-6-methyloxan-2-yl] oxyoxan-2-yl] methyl (E)-3-(4-hydroxy-3-methoxyphenyl) prop-2-enoate	Passed
10606127 (Compound 2)		[(2S,3S,4R,5R)-2-[[[(E)-3-(3,4-dihydroxyphenyl) prop-2-enoyl] oxymethyl]-4-hydroxy-5-(hydroxymethyl)-2-[(2R,3R,4S,5R,6R)-6-(hydroxymethyl)-5-[(E)-3-(4-hydroxyphenyl) prop-2-enoyl] oxy-3,4-bis[[[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxy] oxan-2-yl] oxyoxolan-3-yl] benzoate	Passed
5318857 (Compound 3)		(5R,10S,13R,16R,19S)-10-[[[(4S,5S)-4-[(4S,6R)-4,5-dihydroxy-6-(hydroxymethyl)-3-[(2S,3R,5S)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxyoxan-2-yl] oxy-3,5-dihydroxyoxan-2-yl] oxy-16,19-dihydroxy-4,5,9,9,13,19,20-heptamethyl-21-oxahexacyclo [18.2.2.01,18.04,17.05,14.08,13] tetracos-17-en-22-one	Passed
457885 (Compound 4)		[(2R,3S)-2-(3,4-dihydroxyphenyl)-5,7-dihydroxy-3,4-dihydro-2H-chromen-3-yl] (3S)-5-(3,4-dihydroxyphenyl)-3-hydroxypentanoate	Passed
44256914 (Compound 5)		[(3S,4S,6S)-3,4,5-trihydroxy-6-[5-hydroxy-2-[4-hydroxy-3,5-bis[[[(2S,5S,6R)-3,4,5-trihydroxy-6-[[[(E)-3-(4-hydroxyphenyl) prop-2-enoyl] ox methyl] oxan-2-yl] oxy] phenyl]-3-[(2S,5S)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxychromenylium-7-yl] oxyoxan-2-yl] methyl (E)-3-(4-hydroxyphenyl) prop-2-enoate	Passed

(Continued on following page)

TABLE 4 (Continued) PubChem ID of the compound, IUPAC name of compound and the PAIN filter result of the compounds.

Compound ID	Structure	IUPAC name	PAINS filter
6321424 (Reference compound)		(1R,4S,5'S,6R,6'R,8R,10E,12S,13S,14E,16E,20R,21R,24S)-6'-[(2S)-butan-2-yl]-21,24-dihydroxy-12-[[[2R,4S,5S,6S)-5-[(2S,4S,5S,6S)-5-hydroxy-4-methoxy-6-methyloxan-2-yl]oxy-4-methoxy-6-methyloxan-2-yl]oxy-5',11,13,22-tetramethylspiro[3,7,19-trioxatetracyclo[15.6.1.14,8.020,24]]pentacos-10,14,16,22-tetraene-6,2'-oxane]-2-one	Passed

3.6.2 Root mean square fluctuation

The individual amino acid fluctuations of the main protease in complex with ligands were computed by RMSF analysis to assess the stability of the active site residues toward the compounds in the entire 100 ns MD trajectories (Figure 9). The RMSF of the main protease in the APO state, reference compound, and all five phytochemicals bounds to the main protease were analyzed and compared to each other, the black line in each plot represents the apo state while the red line indicated the residual flexibility of reference compound bounds to the target protein. Figure 9 shows that residues 51 and 250–260 show higher fluctuations. All these fluctuating residues were not found in the active site and these residues were far away from the active site indicating the stable binding of phytochemicals in the active site of the target protein.

3.6.3 Radius of gyration

The radius of gyration is useful for exploring the compactness and folding of the protein, Higher Rg values are indicative of less compactness (more unfolded), while lower Rg values indicate more structural rigidity and strong compactness. The MD simulation study serves to illustrate the effects of inhibitors binding upon the conformation of protein molecules. As illustrated in Figure 10 the results of Rg analysis indicate that these phytochemicals bound to 3CL^{PRO} have less radius of gyration values compared to the apo state, which demonstrates the 3CL^{PRO} stability, and compactness after ligand binding. The reference compound, Compound 1, and Compound 4 have almost similar Rg values, with an average Rg value of 22–22.3 and 22–22.4 Å while Compound 2, Compound 3, and Compound 5 showed an average gyration of 22–22.5, 22–23.3 and 22–22.4 Å respectively. The compactness of the protein was significantly affected by the binding and unbinding of these phytochemical inhibitors.

3.6.4 Dynamic cross-correlation matrix (DCCM) analysis

The extent of correlation motion between the residues imposed by the binding of compounds in the active site of 3CL^{PRO} was elucidated by the inter-residue correlation analysis. The results indicate that compound 1 in complex with the receptor active site residues showed significantly stronger parallel correlations motions in comparison with the control compound, which further validates that these positive correlation motions may be induced by the acquired interaction of these compounds with the key residues (25–50, 141–145,163), like hydrophilic, hydrogen and hydrophobic. Overall, the DCCM findings demonstrate that the control compound and our identified compound displayed comparable patterns of highly positive correlation. Furthermore, for compound 3 and compound

5 the nearby loops regions were also found in strong positive correlations as shown in Figure 11. The dark green color demonstrates a positive correlation in residues of protein while the dense brown color indicates a negative correlation between the protein residues. The negatively correlated residues move in an anti-parallel direction while the positively correlated residues move in a parallel direction.

3.7 GBSA results

3.7.1 MM-GBSA analysis

Protein-ligand complexes from the MD simulation trajectories were used to calculate the energy parameters to assess the energetics of 3CL^{PRO} to the ligands. The binding free energies of each system were calculated using the MM-GBSA method. Table 6 display the computed average binding free energies and specific energetic contribution components of the final 500 frames. As can be observed, compound 1 has smaller free energy (–56.94 kcal/mol) followed by compound 2 (–55.65 kcal/mol), compound 3 (–53.58 kcal/mol), and compound 4 (–46.95 kcal/mol). It was observed that, as compared to the control system, all the ligands in complex with 3CL^{PRO} revealed high binding affinity demonstrating that all the systems are stable. Out of all, the binding affinity of system one was very high for the receptor. This outcome is consistent with the conclusion drawn from the earlier RMSD and docking analysis i.e., compound 1 showed stable dynamic behavior and established a greater number of non-covalent interactions (Figure 8A; Table 5).

4 Discussion

The increased mortality rate of SARS-CoV-2 has created a pandemic situation globally, no effective drugs and treatments are available to treat COVID-19, however, many clinical trials are undergoing. New infectious agents, like SARS and MERS, have emerged in the last 20 years and have created epidemics. The functional significance of 3CL^{PRO} in the viral life cycle and the lack of closely comparable human homologs make 3CL^{PRO} an attractive target for the development of antiviral medications (Jin et al., 2020). By targeting the 3CL^{PRO} most of the natural compounds play a significant role in the treatment of COVID-19 infections (Jin et al., 2020; Mengist et al., 2020). *In vitro*, animal models, and clinical trials are all used to study natural compounds that are extracted from medicinal plants, animals, and marine species for the treatment of COVID-19 (Wu et al., 2019; Wei et al., 2020; Sahoo et al., 2021). One of the most promising and

TABLE 5 Docking score and interaction of top five hits against the 3CL^{pro}.

C. No	Docking score	Ligand	Receptor	Residues	Interaction	Distance	Energy (kcal/mol)
Compound 1	-12.0321	O 4	SG	CYS 145	H-donor	4.06	-0.5
		O 8	SG	CYS 145	H-donor	4.04	-0.8
		O 14	OG	SER 46	H-donor	2.96	-0.6
		C 28	SD	MET 49	H-donor	3.89	-0.8
		O 2	NE2	HIS 41	H-acceptor	3.29	-0.7
		O 8	NE2	HIS 163	H-acceptor	3.05	-0.7
		O 9	NE2	HIS 163	H-acceptor	3.28	-1.8
		O 11	CA	LEU 141	H-acceptor	3.49	-0.6
		6-ring	CA	THR 25	π -H	4.07	-0.6
Compound 2	-11.4527	O 13	SG	CYS 145	H-donor	4.40	-0.7
		O 15	SD	MET 49	H-donor	3.84	-0.5
		O 18	O	THR 26	H-donor	2.86	-1.4
		O 21	OD1	ASN 142	H-donor	2.84	-0.6
		O 25	SD	MET 165	H-donor	3.60	-1.2
		O 12	NE2	HIS 41	H-acceptor	2.96	-0.8
		O 19	NE2	HIS 163	H-acceptor	3.07	-1.9
		6-ring	N	SER 46	π -H	4.24	-1.4
		6-ring	N	THR 90	π -H	4.33	-0.6
Compound 3	-11.2783	O 8	SD	MET 49	H-donor	3.79	-0.5
		O 22	SG	CYS 145	H-donor	3.19	-1.1
		C 26	OE1	GLN 189	H-donor	3.13	-0.9
		O 22	NE2	HIS 41	H-acceptor	3.15	-1.0
		O 23	NE2	HIS 163	H-acceptor	3.19	-1.0
		6-ring	CA	LEU 141	π -H	3.80	-0.5
Compound 4	-10.9628	O 4	O	THR 26	H-donor	2.80	-2.2
		O 6	ND1	HIS 164	H-donor	2.95	-1.8
		O 9	OG1	THR 25	H-donor	3.05	-1.6
		C 13	SD	MET 49	H-donor	3.81	-0.6
		O 5	N	GLY 143	H-acceptor	3.16	-2.7
		6-ring	5-ring	HIS 41	π - π	3.27	-0.0
Compound 5	-10.7164	O 10	OD1	ASN 142	H-donor	3.11	-1.9
		O 15	O	GLN 189	H-donor	3.07	-1.0
		O 18	O	THR 26	H-donor	3.01	-1.8
		C 57	SD	MET 49	H-donor	3.94	-0.6
		O 18	N	THR 26	H-acceptor	2.95	-0.9
		O 30	NE2	HIS 41	H-acceptor	3.10	-0.6
IVERMECTIN	-9.5398	O 5	SG	CYS 145	H-donor	3.77	-0.6
		O 6	O	ASP 187	H-donor	2.91	-0.4
		C 35	SD	MET 165	H-donor	3.81	-0.5
		C 45	SD	MET 49	H-donor	3.49	-0.2
		O 13	ND2	ASN 119	H-acceptor	3.43	-0.6

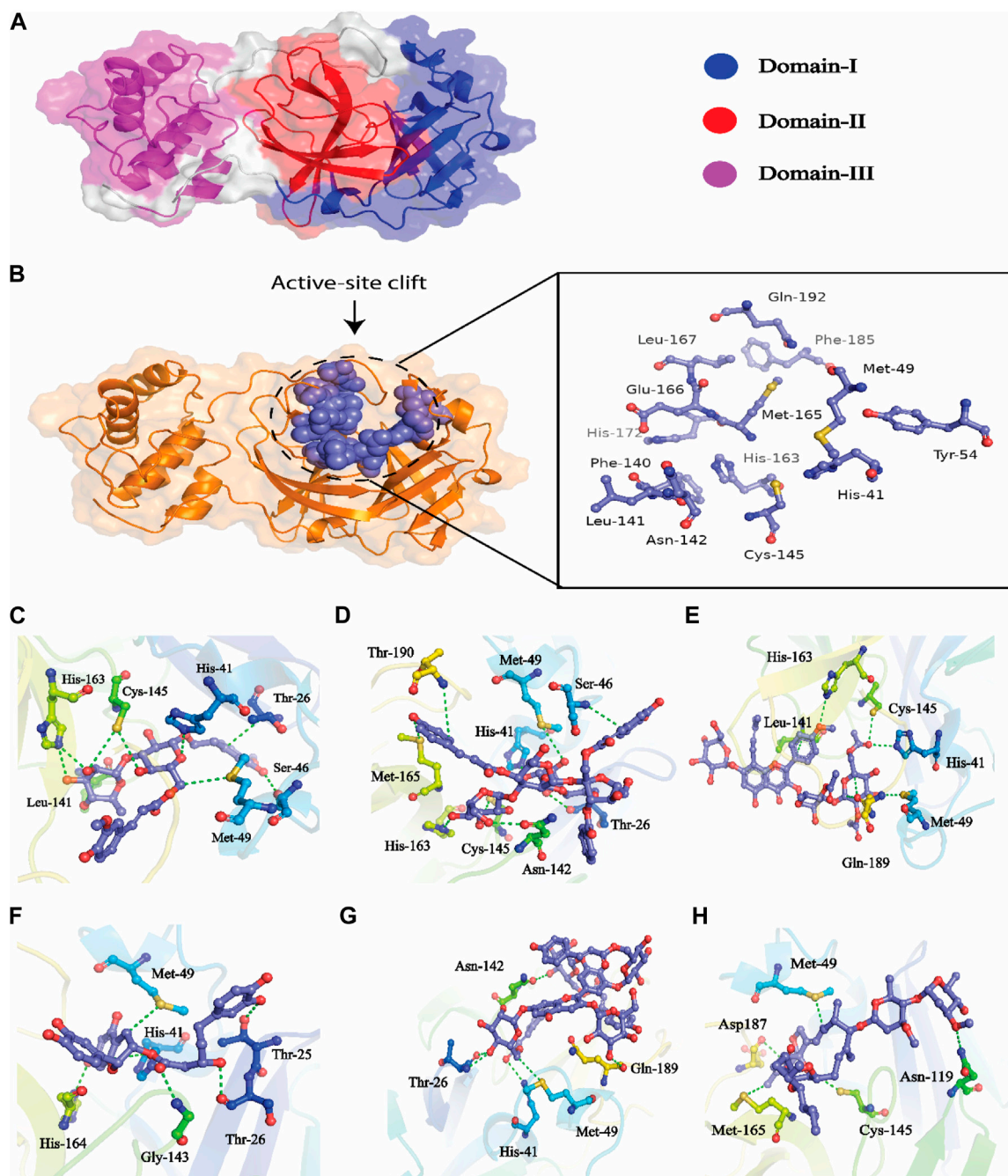


FIGURE 7

(A) All the three domains of 3CL^{PRO}, (B) active site of the main protease and (C) indicates the interaction of Compound 1 in the active site of 3CL^{PRO}, (D) represents the 3D interactions of Compound 2, (E) indicates the 3D interaction of Compound 3, (F) indicates interactions of Compound 4, (G) indicates the interaction of Compound 5, (H) indicating the three-dimensional interactions of the Control compounds (Ivermectin) with the 3CL^{PRO}.

effective strategies for combating the current pandemic is still seen to be the use of natural products (ying et al., 2001). Extractions from medicinal plants and their secondary metabolites frequently show strong antiviral properties. Some *in vitro* studies showed that PSM and viral incubation had direct interference. The viral protein, its lipid layers, and the cell's lysis can be destroyed by the plants' metabolites (Akram et al., 2018). There are about six to seven thousand different plant species in Pakistan, of which 700 are regularly used as medicines (Khan et al., 2022). The SARS CoV 2 RdRp was chosen as a receptor for computational drug development in the previous study in which 200 phytochemicals were used for virtual

screening. The top 10 ligands among 200 total ligands were chosen based on drug discovery criteria such as S-score, ligand interactions, hydrophobic interactions, and drug-likeness (Mahrosh and Mustafa, 2021).

Developing a new drug against the virus is time-consuming and costly. The ability of computer-aided drug design, on the other hand, to screen a large library of small molecules quickly and accurately may help the researcher to develop a new therapeutic agent against SARS-CoV-2 (Wang, 2020). The virtual screening workflow has made it possible to screen the enormous, diverse chemical library for

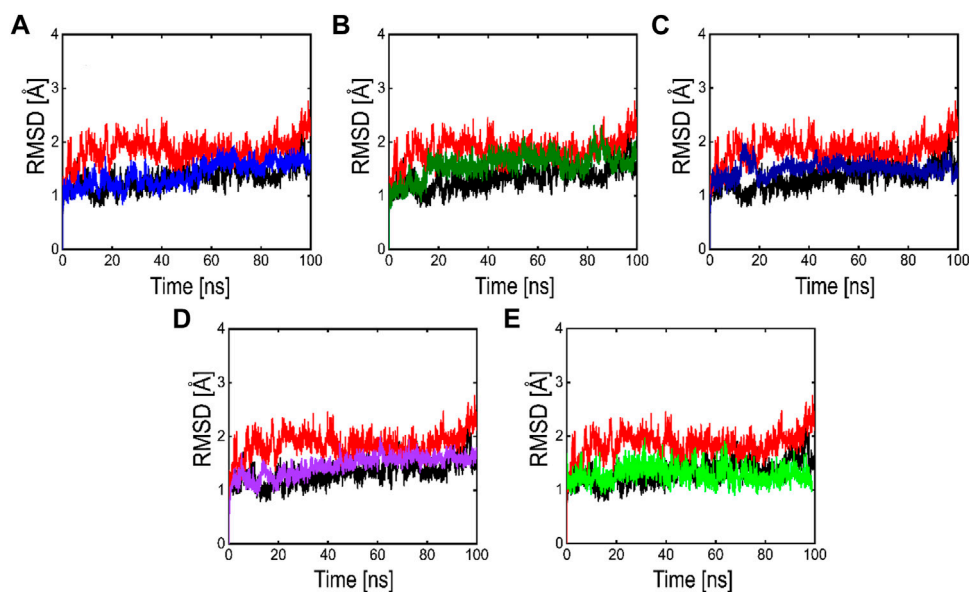


FIGURE 8

RMSD plots of the APO form (Black color), reference complex (Red color) and the top active phytochemicals (A) Compound 1 (B) Compound 2 (C) Compound 3 (D) Compound 4 and (E) Compound 5 bound to 3CL^{PRO}.

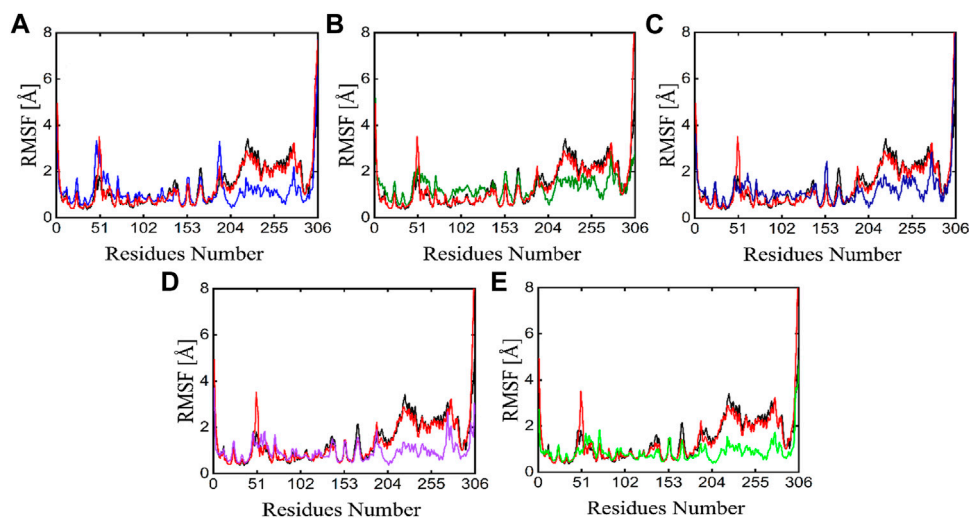


FIGURE 9

RMSF plots of the APO state (Black), control (Red) and the potent phytochemicals (A) Compound 1 (B) Compound 2 (C) Compound 3 (D) Compound 4 and (E) Compound 5.

the identification of powerful inhibitors (Neves et al., 2018). In the drug development processes, machine learning (ML) techniques are frequently used to categorize compounds as potentially active or inactive against a given protein target (Patel et al., 2020). Structure and ligand-based virtual screening frequently yield a high proportion of false positive hits (Deng et al., 2015). To reduce the false positive hits in this work, we used to machine-learning-base virtual screening for the prediction of new inhibitors against the 3CL^{PRO}. K-nearest neighbor (KNN), support vector machine (SVM), and Random Forest (RF) algorithm three of the most popular ML algorithms were chosen for virtual screening workflow. In general, classifier

performance is evaluated in terms of accuracy. KNN achieved 0.93% accuracy SVM achieved 0.96% accuracy, whereas RF produced 0.99% accuracy on the train set. Our results revealed the best performance of the RF model, so we used the RF model to classify the Asian phytochemicals. Out of 4,000 phytochemicals, a total of 26 phytochemicals were predicted as active against the 3CL^{PRO}. These active hits were further docked into the active site of the main protease. Among the 26 docked compounds, Compound 1 was found as the most potent with a docking score of -12.03 and it formed four H-donor interaction with CYS145, SER46, MET49, and four H-acceptor interactions with HIS41, HIS163, LEU141 one pi-H

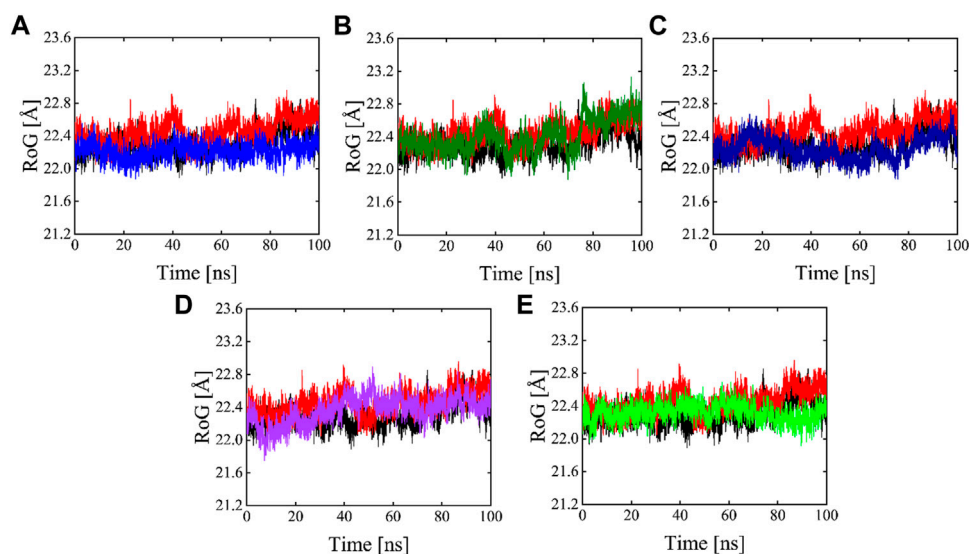


FIGURE 10

Rg plots of Apo (Black), red (reference), and Compound 1-5 are labeled different colors as (A–E) Respectively.

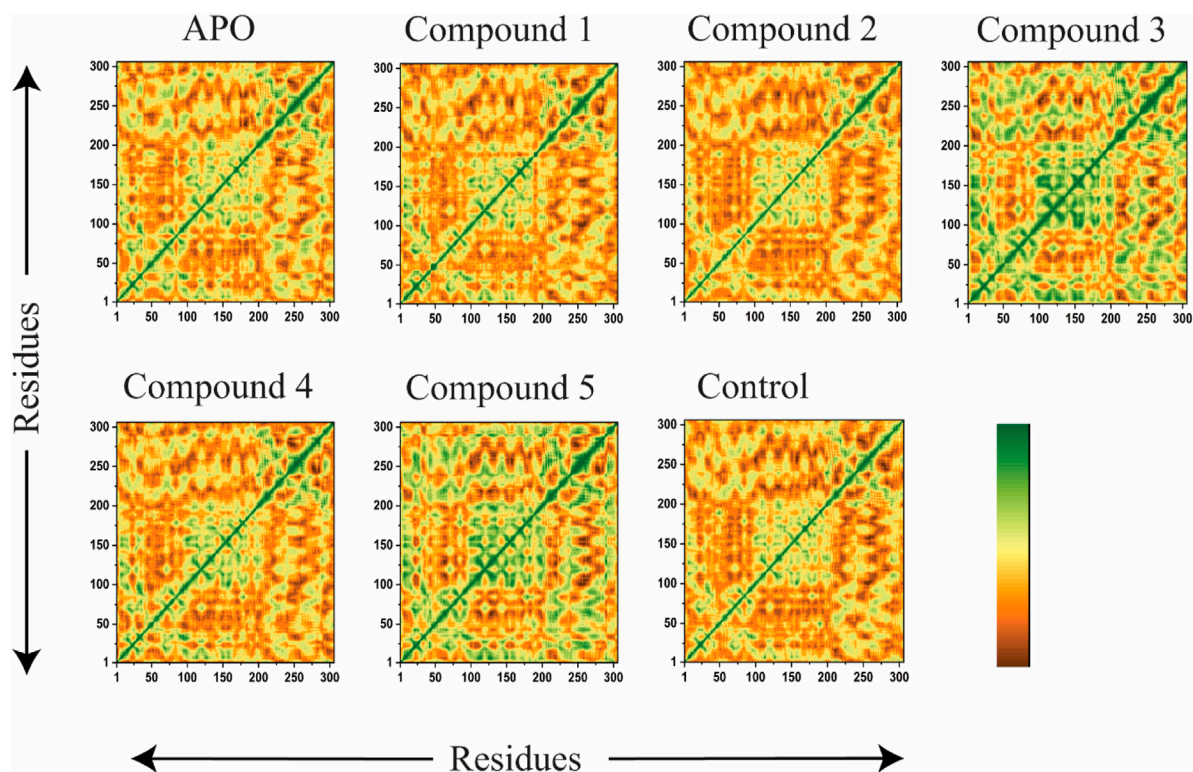


FIGURE 11

DCCM of the Apo state, Compound 1, Compound 2, Compound 3, Compound 4, Compound 5, and ivermectin (control) bound to 3CL^{PRO}. The positively correlated movement is represented by green color, while negatively correlated motion is indicated by deep brown color.

interaction with THR25 active site residues. Compound 2 was found as the second most potent hit with a docking score of -11.45 followed by Compound 3. Compound 2 formed a total of five hydrogen bonds donor interactions with the active site residues including CYS145, MET49, THR26, ASN142, MET165, and two H-acceptor

interactions with HIS41, and HIS163. The docking scores as well as interactions of Compound 3, 4 and 5 were also good as compared to the standard compound. The docking score of reference compound ivermectin was -9.53 and it formed a total of four H-donor interactions with CYS 145, MET 165 and one

TABLE 6 Represents MMGBSA Binding Free Energy (kcal/mol) calculation for the selected phytochemicals and control compound.

S. No	Compound name	VDWAALS	EEL	EGB	ESURF	-TAS	DELTA TOTAL
1	Compound 1	-83.4745	-20.3304	56.6693	-9.8094	-18.4312	-56.9450
2	Compound 2	-79.3325	-20.6400	52.7843	-8.4635	-17.8254	-55.6517
3	Compound 3	-73.1537	-19.5693	51.8532	-8.5177	-19.2984	-53.5835
4	Compound 4	-64.4348	-16.3432	41.7462	-6.8571	-13.9835	-46.9500
5	Compound 5	-42.2227	-4.3191	13.2240	-4.7141	-10.8921	-38.0319
6	Ivermectin	-38.9027	-6.3834	20.7589	-4.3827	-14.5924	-28.9100

vdW = the van der Waals energy, EEL, electrostatic energy; ESURF, surface areas energy; EGB, the electrostatic contribution to the solvation free energy calculated by GB.

H-acceptor interaction with ASN 119 active site residue. Additionally, dynamics simulation was carried out to comprehend and support the molecular docking study. For all the systems the averaged RMSD was found between 1 and 3 Å. The averaged RMSD for ivermectin was 2.0 Å, initially, up to 40 ns the system undergoes raised up in the RMSD value up to 40 ns, and soon after reaching 40 ns the system acquired stability and remained stable for the rest of the simulation period. The complex Compound 1 shows significant stability as can be observed, however after 60 ns, the system briefly displayed a tolerable variation. The system thereafter became stable and moved into the production phase. For Compound 2, the finding of the stability index in terms of RMSD reveals that the system shows highly stable behavior in the entire period of simulation, at 20 ns minor fluctuations from its mean position were observed, afterward, the system gained stability and no more significant deviations were observed with the average RMSD value of 1.7 Å. For complex Compound 3, the system initially shows invariant behavior, up to 15 ns a gradual increase in the RMSD curve was observed followed by a slight decrease in the RMSD curve at 20 ns afterward the system attains the equilibrated with the averaged RMSD value of 2.1 Å. The protein structure's compactness as a function of time can be evaluated by the radius of gyration (Ajmal et al., 2022). The RoG analysis revealed compound 1, and compound 4 have almost similar Rg values, with an average Rg value of 22–22.3 and 22–22.4 Å while compound 2, compound 3, and compound 5 showed an average gyration of 22–22.5, 22–23.3 and 22–22.4 Å respectively. The Rg analysis of all the simulated complexes revealed that these phytochemicals formed stable and compact complexes with 3CL^{PRO}. All the short-listed phytochemicals revealed good binding affinity for 3CL^{PRO}. Compound 1 has smaller free energy (-56.94 kcal/mol) followed by compound 2 (-55.65 kcal/mol), compound 3 (-53.58 kcal/mol), and compound 4 (-46.95 kcal/mol). It was observed that, as compared to the control system, all the ligands in complex with 3CL^{PRO} revealed high binding affinity demonstrating that all the systems are stable. The RMSF analysis revealed that Domain II had a stable behavior, whereas Domain I and Domain III's amino acid residues had more flexibility in the helix and turn regions. The overall finding of RMSD and binding energy indicates that our novel phytochemicals have higher binding capacity toward the active site of the main protease. ML-based workflow combined with molecular docking and molecular dynamics approach reveals that the predicted new active phytochemicals may disrupt the SARS-CoV-2 3CL^{PRO} function.

5 Conclusion

We used *in silico* machine learning tools for drug designing against the SARS-CoV-2 3CL^{PRO}. The phytochemical dataset with more than 4,000 chemicals derived from the PubChem database was used for virtual screening against 3CL^{PRO}. Furthermore, the compound's inhibitory potential was explored using the molecular docking and MD simulation study. Using these advanced approaches, we found high-potential therapeutic compounds that can possibly inhibit SARS-CoV-2 pathogenesis. The virtual screening process, which includes MM-GBSA methods assists in reducing the list from over 4,000 possible lead compounds to 26 compounds. This research relies only on various computational tools and further it is recommended to evaluate the *in-vitro* inhibitory potential of these short-listed compounds. Successful assessment and *in vitro* evaluation of these compounds will help us to use them as a therapeutic option to treat and cope with COVID-19.

Data availability statement

Data will be provided upon reasonable request from the corresponding author of this manuscript. Requests to access these datasets should be directed to awadood@awkum.edu.pk.

Author contributions

AS, AJ, and SMJ performed experiments, analyzed data, and drafted the manuscript. AM and BK analyzed data, interpreted the results, drafted the manuscript, and revised the manuscript. AM, PL, AR, AA, MU, and PH revised the manuscript, drafted the methods, performed proofreading and improved discussion. MU and AS draw figures and tables. HJ, AM, and AW, designed, conceptualized, drafted the manuscript, analyzed and interpreted the results and revised the manuscript.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: (22UQU4331128DSR60).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmolb.2023.1060076/full#supplementary-material>

References

- Agbowuro, A. A., Huston, W. M., Gamble, A. B., and Tyndall, J. D. A. (2018). Proteases and protease inhibitors in infectious diseases. *Med. Res. Rev.* 38, 1295–1331. doi:10.1002/med.21475
- Ahmad, A., Akbar, S., Khan, S., Hayat, M., Ali, F., Ahmed, A., et al. (2021). Deep-AntiFP: Prediction of antifungal peptides using distant multi-informative features incorporating with deep neural networks. *Chemom. Intelligent Laboratory Syst.* 208, 104214. doi:10.1016/j.chemolab.2020.104214
- Ajmal, A., Ali, Y., Khan, A., Wadood, A., and Rehman, A. U. (2022). Identification of novel peptide inhibitors for the KRas-G12C variant to prevent oncogenic signaling. *J. Biomol. Struct. Dyn.*, 1–10. doi:10.1080/07391102.2022.2138550
- Akram, M., Tahir, I. M., Shah, S. M. A., Mahmood, Z., Altaf, A., Ahmad, K., et al. (2018). Antiviral potential of medicinal plants against HIV, HSV, influenza, hepatitis, and coxsackievirus: A systematic review. *Phytotherapy Res.* 32(5), 811–822. doi:10.1002/ptr.6024
- Alanagreh, L., Alzoughool, F., and Atoum, M. (2020). The human coronavirus disease COVID-19: Its origin, characteristics, and insights into potential drugs and its mechanisms. *Pathogens* 9, 331. doi:10.3390/pathogens9050331
- Araki, T., Ikeda, N., Shukla, D., Jain, P. K., Londhe, N. D., Shrivastava, V. K., et al. (2016). PCA-based polling strategy in machine learning framework for coronary artery disease risk assessment in intravascular ultrasound: A link between carotid and coronary grayscale plaque morphology. *Comput. Methods Programs Biomed.* 128, 137–158. doi:10.1016/j.cmpb.2016.02.004
- Ashraf, S., Ranaghan, K. E., Woods, C. J., Mulholland, A. J., and Ul-Haq, Z. (2021). Exploration of the structural requirements of Aurora Kinase B inhibitors by a combined QSAR, modelling and molecular simulation approach. *Sci. Rep.* 11, 18707. doi:10.1038/s41598-021-97368-3
- Asif, A., Ilyas, I., Abdullah, M., Sarfraz, S., Mustafa, M., and Mahmood, A. (2022). The comparison of mutational progression in SARS-CoV-2: A short updated overview. *J. Mol. Pathology* 3, 201–218. doi:10.3390/jmp3040018
- Baell, J. B., and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740. doi:10.1021/jm901137j
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 261–277. doi:10.1023/a:1017934522171
- Colovos, C., and Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci.* 2(9), 1511–1519. doi:10.1002/pro.5560020916
- Darden, T., York, D., and Pedersen, L. (1998). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98 (12), 10089–10092. doi:10.1063/1.464397
- Das, P., Majumder, R., Mandal, M., and Basak, P. (2021). *In-silico* approach for identification of effective and stable inhibitors for COVID-19 main protease (Mpro) from flavonoid based phytochemical constituents of *Calendula officinalis*. *J. Biomol. Struct. Dyn.*, 39(16), 6265–6280. doi:10.1080/07391102.2020.1796799
- Deng, N., Forli, S., He, P., Perryman, A., Wickstrom, L., Vijayan, R. S. K., et al. (2015). Distinguishing binders from false positives by free energy calculations: Fragment screening against the flap site of HIV protease. *ACS Publ.* 119(3), 5. doi:10.1021/jp506376z
- Durojaiye, A. B., Clarke, J. R. D., Stamatiades, G. A., and Wang, C. (2020). Repurposing cefuroxime for treatment of COVID-19: A scoping review of *in silico* studies. *J. Biomol. Struct. Dyn.* 39 (6489), 1–8. doi:10.1080/07391102.2020.1777904
- Elmaaty, A. A., Darwish, K. M., Khattab, M., Elhady, S. S., Salah, M., Hamed, M. I. A., et al. (2022). In a search for potential drug candidates for combating COVID-19: Computational study revealed salvianolic acid B as a potential therapeutic targeting 3CLpro and spike proteins. *J. Biomol. Struct. Dyn.* 40 (19), 8866–8893. doi:10.1080/07391102.2021.1918256
- Floresta, G., Zagni, C., Gentile, D., Patamia, V., and Rescifina, A. (2022). Artificial intelligence technologies for COVID-19 De Novo drug design. *Int. J. Mol. Sci.* 23, 3261. doi:10.3390/ijms23063261
- Gul, S., Ozcan, O., Asar, S., Okyar, A., Baris, I., and Kavakli, I. H. (2021). *In silico* identification of widely used and well-tolerated drugs as potential SARS-CoV-2 3C-like protease and viral RNA-dependent RNA polymerase inhibitors for direct use in clinical trials. *J. Biomol. Struct. Dyn.*, 39(17), 6772–6791. doi:10.1080/07391102.2020.1802346
- Gurung, A. B., Ali, M. A., Lee, J., Farah, M. A., and Al-Anazi, K. M. (2021). An updated review of computer-aided drug design and its application to COVID-19. *Biomed. Res. Int.* 2021, 8853056. doi:10.1155/2021/8853056
- Hatada, R., Okuwaki, K., Mochizuki, Y., Handa, Y., Fukuzawa, K., Komeiji, Y., et al. (2020). Fragment molecular orbital based interaction analyses on COVID-19 main protease - inhibitor N3 complex (PDB ID: 6LU7). *J. Chem. Inf. Model.* 60, 3593–3602. doi:10.1021/acs.jcim.0c00283
- Hilgenfeld, R., and Hilgenfeld, C. R. (2014). From SARS to MERS: Crystallographic studies on coronavirus proteases enable antiviral drug design. *FEBS J. [Internet]* 281 (18), 4085–4096. doi:10.1111/febs.12936
- Huang, J., Song, W., Huang, H., and Sun, Q. (2020). Pharmacological therapeutics targeting RNA-dependent RNA polymerase, proteinase and spike protein: From mechanistic studies to clinical trials for COVID-19. *J. Clin. Med.* 9, 1131. doi:10.3390/jcm9041131
- Jackson, C. B., Farzan, M., Chen, B., and Choe, H. (2021). Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell. Biol.* 23, 3–20. doi:10.1038/s41580-021-00418-x
- Jan, J. T., Cheng, T. J. R., Juang, Y. P., Ma, H. H., Wu, Y. T., Yang, W. B., et al. (2021). Identification of existing pharmaceuticals and herbal medicines as inhibitors of SARS-CoV-2 infection. *Proc. Natl. Acad. Sci. U. S. A.* 118(5), e2021579118. doi:10.1073/pnas.2021579118
- Janson, G., Zhang, C., Prado, M. G., and Paiardini, A. (2017). PyMod 2.0: Improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics* 33(3), 444–446. doi:10.1093/bioinformatics/btw638
- Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al. (2020). Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582, 289–293. doi:10.1038/s41586-020-2223-y
- Junaid, M., Shah, M., and Khan, A. (2018). CLJ of, 2018 undefined. Structural-dynamic insights into the *H. pylori* cytotoxin-associated gene A (CagA) and its abrogation to interact with the tumor suppressor protein ASP2 using decoy. *Taylor Francis* 37(15), 4035–4050. doi:10.1080/07391102.2018.1537895
- Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., and Hooman, A. (2013). An overview of principal component analysis. *J. Signal Inf. Process.* 04 (3), 173–175. doi:10.4236/jsip.2013.43b031
- Kaur, H., Shekhar, N., Sharma, S., Sarma, P., Prakash, A., and Medhi, B. (2021). Ivermectin as a potential drug for treatment of COVID-19: An in-silico review with clinical and computational attributes. *Pharmacol. Rep.* 73(3), 736–749. doi:10.1007/s43440-020-00195-y
- Khan, M., Zaem, A., Munir, A., Ulfat, A., and Ampak, J. B. (2022). Undefined. Plants secondary metabolites (psms), as an investigational source against Covid-19 from flora of Pakistan. *Pakbs. Org.*, 1485–1493. Available from: <https://pakbs.org/pjbot/papers/1650356086.pdf>.
- Klemm, T., Ebert, G., Calleja, D. J., Allison, C. C., Richardson, L. W., Bernardini, J. P., et al. (2020). Mechanism and inhibition of the papain-like protease, PLpro, of SARS-CoV-2. *EMBO J.* 39 (18), e106275. doi:10.15252/embj.2020106275
- Kneller, D. W., Phillips, G., Weiss, K. L., Pant, S., Zhang, Q., O'Neill, H. M., et al. (2020). Unusual zwitterionic catalytic site of SARS-CoV-2 main protease revealed by neutron crystallography. *J. Biol. Chem.* 295, 17365–17373. doi:10.1074/jbc.ac120.016154
- Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* 8(4), 477–486. doi:10.1007/BF00228148
- Luttens, A., Gullberg, H., Abdurakhmanov, E., Vo, D. D., Akaberi, D., Talibov, V. O., et al. (2022). Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors

- with broad-spectrum activity against coronaviruses. *J. Am. Chem. Soc.* 144 (7), 2905–2920. doi:10.1021/jacs.1c08402
- Macalino, S. J. Y., Gosu, V., Hong, S., and Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives Pharmacol. Res.* 38, 1686–1701. doi:10.1007/s12272-015-0640-5
- Macalino, S. J. Y., Basith, S., Clavio, N. A. B., Chang, H., Kang, S., and Choi, S. (2018). Evolution of *in silico* strategies for protein-protein interaction drug discovery. *Molecules* 23, 1963. doi:10.3390/molecules23081963
- Mahrosh, H. S., and Mustafa, G. (2021). An *in silico* approach to target RNA-dependent RNA polymerase of COVID-19 with naturally occurring phytochemicals. *Environ. Dev. Sustain* 23 (11), 16674–16687. doi:10.1007/s10668-021-01373-5
- Marty, A. M., and Jones, M. K. (2020). The novel Coronavirus (SARS-CoV-2) is a one health issue. *One Health* 9. doi:10.1016/j.onehlt.2020.100123
- Mengist, H. M., Fan, X., and Jin, T. (2020). Designing of improved drugs for COVID-19: Crystal structure of SARS-CoV-2 main protease Mpro. *Signal Transduct. Target. Ther.* 5, 67. doi:10.1038/s41392-020-0178-y
- Mengist, H. M., Dilnessa, T., and Jin, T. (2021). Structural basis of potential inhibitors targeting SARS-CoV-2 main protease. *Front. Chem.* 9, 7. doi:10.3389/fchem.2021.622898
- Messaoudi, A., Belguith, H., and ben Hamida, J. (2013). Homology modeling and virtual screening approaches to identify potent inhibitors of VEB-1 β -lactamase. *Theor. Biol. Med. Model.* 10(1), 1–10. doi:10.1186/1742-4682-10-22
- Mouffouk, C., Mouffouk, S., Mouffouk, S., Hambaba, L., and Haba, H. (2021). Flavonols as potential antiviral drugs targeting SARS-CoV-2 proteases (3CLpro and PLpro), spike protein, RNA-dependent RNA polymerase (RdRp) and angiotensin-converting enzyme II receptor (ACE2). *Eur. J. Pharmacol.* 891, 173759. doi:10.1016/j.ejphar.2020.173759
- Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi:10.1021/jm300687e
- Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., and Andrade, C. H. (2018). QSAR-based virtual screening: Advances and applications in drug discovery. *Front. Pharmacol.* 9, 1275. doi:10.3389/fphar.2018.01275
- Noreen, K., Azween, A., Belhaouari, S. B., Sellapan, P., Saeed, A. B., and Nilanjan, D. (2016). Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. *J. Med. Imaging Health Inf.* 6 (1), 78–87. doi:10.1166/jmhi.2016.1593
- Patel, L., Shukla, T., Huang, X., Ussery, D. W., and Wang, S. (2020). Machine learning methods in drug discovery. *Molecules* 25, 5277. doi:10.3390/molecules25225277
- Perez-Lemus, G. R., Menéndez, C. A., Alvarado, W., Byléhn, F., and de Pablo, J. J. (2022). Toward wide-spectrum antivirals against coronaviruses: Molecular characterization of SARS-CoV-2 NSP13 helicase inhibitors. *Sci. Adv.* 8. doi:10.1126/sciadv.abj4526
- Prada Gori, D. N., Llanos, M. A., Bellera, C. L., Talevi, A., and Alberca, L. N. (2022). iRaPCA and SOMoC: Development and validation of web applications for new approaches for the clustering of small molecules. *J. Chem. Inf. Model.* 62 (12), 2987–2998. doi:10.1021/acs.jcim.2c00265
- Rothan, H. A., and Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J. Autoimmun.* 109, 102433. doi:10.1016/j.jaut.2020.102433
- Sahoo, A., Fuloria, S., Swain, S. S., Panda, S. K., Sekar, M., Subramaniyan, V., et al. (2021). Potential of marine terpenoids against sars-cov-2: An *in silico* drug development approach. *Biomedicine* 9(11), 1505. doi:10.3390/biomedicine9111505
- Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3(2), 198–210. doi:10.1002/wcms.1121
- Salomon-Ferrer, R., Götz, A. W., Poole, D., le Grand, S., and Walker, R. C. (2013). Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. *J. Chem. Theory Comput.* 9, 3878–3888. doi:10.1021/ct400314y
- Sandhu, H., RajaramKumar, N., and Garg, P. (2022). Machine learning-based modeling to predict inhibitors of acetylcholinesterase Graphic abstract Keywords Acetylcholinesterase, *Mol. Divers.*, 26, 1, 1–10. doi:10.1007/s11030-021-10223-5
- Santos, B. S., Silva, L., Ribeiro-Dantas, M. da C., Alves, G., Endo, P. T., and Lima, L. (2020). COVID-19: A scholarly production dataset report for research analysis. *Data Brief* 32, 106178. doi:10.1016/j.dib.2020.106178
- Sarker, I. H. (2021). CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. *Internet Things* 14, 100393. doi:10.1016/j.iot.2021.100393
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* 2(3), 1–21. doi:10.1007/s42979-021-00592-x
- Tahir ul Qamar, M., Alqahtani, S. M., Alamri, M. A., and Chen, L. L. (2020). Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants. *J. Pharm. Anal.* 10 (4), 313–319. doi:10.1016/j.jpha.2020.03.009
- Thuy, B. T. P., My, T. T. A., Hai, N. T. T., Hieu, L. T., Hoa, T. T., Thi Phuong Loan, H., et al. (2020). Investigation into SARS-CoV-2 resistance of compounds in garlic essential oil. *ACS Omega* 5(14), 8312–8320. doi:10.1021/acsomega.0c00772
- Ul Hassan, C. A., Khan, M. S., and Shah, M. A. (2018). “Comparison of machine learning algorithms in data classification” in Proceedings of the ICAC 2018 - 2018 24th IEEE International Conference on Automation and Computing: Improving Productivity through Automation and Computing, September 2018, Newcastle Upon Tyne, UK. doi:10.23919/ICAC.2018.8748995
- Ullrich, S., and Nitsche, C. (2020). The SARS-CoV-2 main protease as drug target. *Bioorg Med. Chem. Lett.* 30 (17), 127377. doi:10.1016/j.bmcl.2020.127377
- Wadood, A., Ajmal, A., Junaid, M., Rehman, A. U., Uddin, R., Azam, S. S., et al. (2022). Machine learning-based virtual screening for STAT3 anticancer drug target. *Curr. Pharm. Des.* 28, 3023–3032. doi:10.2174/1381612828666220728120523
- Wadood, A., Ajmal, A., Junaid, M., Rehman, A. U., Uddin, R., Azam, S. S., et al. (2022). Machine learning-based virtual screening for STAT3 anticancer drug target-. *Curr. Pharm. Des.* 28, 3023–3032. doi:10.2174/1381612828666220728120523
- Wadood, A., Shareef, A., Ur Rehman, A., Muhammad, S., Khurshid, B., Khan, R. S., et al. (2022). Silico drug designing for ala438 deleted ribosomal protein S1 (RpsA) on the basis of the active compound Zrl15. *ACS Omega* 7(1), 397–408. doi:10.1021/acsomega.1c04764
- Wang, J. (2020). Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. *J. Chem. Inf. Model.* 60, 3277–3286. doi:10.1021/acs.jcim.0c00179
- Wei, X., Zhao, M., Zhao, C., Zhang, X., Qiu, R., Lin, Y., et al. (2020). TMR modern herbal medicine. *Glob. registry COVID-19 Clin. trials Indic. Des. traditional Chin. Med. Clin. trials* 3 (3), 140.
- Wu, H., Gao, S., and Terakawa, S. 2019 Inhibitory effects of fucoidan on NMDA receptors and l-type Ca²⁺ channels regulating the Ca²⁺ responses in rat neurons. *Pharm. Biol.* 57(1), 1–7. doi:10.1080/13880209.2018.1548626
- Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269. doi:10.1038/s41586-020-2008-3
- Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., et al. (2020). Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm. Sin. B* 10 (5), 766–788. doi:10.1016/j.apsb.2020.02.008
- Xu, X., Chen, P., Wang, J., Feng, J., Zhou, H., Li, X., et al. (2020). Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* 63, 457–460. doi:10.1007/s11427-020-1637-5
- Yang, J., Cai, Y., Zhao, K., Xie, H., and Chen, X. (2022). Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov. Today* 27 (11), 103356. doi:10.1016/j.drudis.2022.103356
- Ying, A. T., Qiu, H. R., Yang, Z., Zhang, D. K., Ren, L. G., Cheng, Y. Y., et al. (2001). Alkaloids from *Cynanchum komarovii* with inhibitory activity against the tobacco mosaic virus. *Phytochemistry* 58 (8), 1267–1269. doi:10.1016/s0031-9422(01)00382-x
- Zhou, P., Lou, Y. X., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2019). A novel coronavirus from patients with pneumonia in China. *N. Engl. J. Med.* 382 (8). doi:10.1056/nejmoa200107