Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Chemical Physics 564 (2023) 111709

Contents lists available at ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

Searching for potential inhibitors of SARS-COV-2 main protease using supervised learning and perturbation calculations

Trung Hai Nguyen [a,b,*], Nguyen Minh Tam [a,b], Mai Van Tuan [c], Peng Zhan [d], Van V. Vu [e], Duong Tuan Quang [f,*], Son Tung Ngo [a,b,*]

[a] Laboratory of Theoretical and Computational Biophysics, Advanced Institute of Materials Science, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[b] Faculty of Pharmacy, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[c] Department of Microbiology, Hue Central Hospital, Hue City, Viet Nam
[d] Department of Medicinal Chemistry, Key Laboratory of Chemical Biology (Ministry of Education), School of Pharmaceutical Sciences, Cheeloo College of Medicine, Shandong University, 44 West Culture Road, 250012 Jinan, Shandong, PR China
[e] NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City, Viet Nam
[f] Department of Chemistry, Hue University, Thua Thien Hue Province, Hue City, Viet Nam

ARTICLE INFO

Keywords:
SARS-CoV-2 Mpro
LIE
FEP
Machine learning
Supervised learning

ABSTRACT

Inhibiting the biological activity of SARS-CoV-2 Mpro can prevent viral replication. In this context, a hybrid approach using knowledge- and physics-based methods was proposed to characterize potential inhibitors for SARS-CoV-2 Mpro. Initially, supervised machine learning (ML) models were trained to predict a ligand-binding affinity of ca. 2 million compounds with the correlation on a test set of $R = 0.748 \pm 0.044$. Atomistic simulations were then used to refine the outcome of the ML model. Using LIE/FEP calculations, nine compounds from the top 100 ML inhibitors were suggested to bind well to the protease with the domination of van der Waals interactions. Furthermore, the binding affinity of these compounds is also higher than that of nirmatrelvir, which was recently approved by the US FDA to treat COVID-19. In addition, the ligands altered the catalytic triad Cys145 - His41 - Asp187, possibly disturbing the biological activity of SARS-CoV-2.

1. Introduction

Coronaviruses have the largest genomes among RNA viruses (26–32 kb) encrypting structural and nonstructural proteins [1,2]. Coronaviruses have been infecting humans and normally cause mild respiratory syndrome [3]. However, in 2002, the severe acute respiratory syndrome coronavirus (SARS-CoV) was first recognized in Guandong, China, and was associated with 774 deaths over 8096 infected cases [4]. The Middle East respiratory syndrome coronavirus (MERS-CoV) was first reported in 2012 to be able to transfect animals to humans and lead to severe cases of respiratory syndromes and deaths [5]. This shows that coronavirus can induce severe symptoms and potential pneumonia and death. A novel coronavirus, SARS-CoV-2, causes severe acute respiratory syndromes and is related to millions of deaths worldwide since it initially spread in December 2019 in Wuhan, Hubei Province, China [6-9]. The virus has been suggested to initiate from bats and can quickly transfect between humans [10]. The spreading speed is tremendously high since the virus can exist in an aerosol [11]. Despite efforts to reduce the viral outbreak, more than 450 million people have been infected to date. The viral outbreak effectuated the COVID-19 pandemic. Therefore, the development of therapy is crucial for community health. In this context, remdesivir was first approved as an antiviral drug for treating COVID-19 [12]. However, it is considered a controversial decision [13] since the drug showed disappointing trials [14,15]. After that, Pfizer's Paxlovid, which combined nirmatrelvir and ritonavir, was authorized as the first oral antiviral drug to treat COVID-19 by the FDA [16]. Although Paxlovid effectiveness is as high as 89% in reducing hospitalization or death compared placebo, its components might cause a severe interaction with widely used medications such as statins, blood thinners, and some antidepressants. Moreover, new SARS-CoV-2 variants in the UK, South Africa, and the US [17,18] have prompted scientists to search for more COVID-19 drugs since emerging mutations can reduce the effectiveness

Abbreviations: ML, Machine Learning; FEP, Free Energy Perturbation; SL, Supervised Learning; Mpro, SARS-CoV-2 Mpro; Docking, Simulation; LIE, Linear Interaction Energy.

* Corresponding authors at: Ton Duc Thang University, Ho Chi Minh City, Viet Nam (T.H. Nguyen, S.T. Ngo) and Hue University, Hue City, Viet Nam (D.T. Quang).
E-mail addresses: nguyentrunghai@tdtu.edu.vn (T.H. Nguyen), dtquang@hueuni.edu.vn (D.T. Quang), ngosontung@tdtu.edu.vn (S.T. Ngo).

https://doi.org/10.1016/j.chemphys.2022.111709
Received 3 May 2022; Received in revised form 11 August 2022; Accepted 21 September 2022
Available online 26 September 2022

of current treatments. Indeed, some variants have recently been reported to be able to escape from neutralizing antibodies [19,20]. Therefore, developing an appropriate treatment for COVID-19 is urgently needed.

Among more than 20 structural proteins and nonstructural proteins (nsp) encoded by SARS-CoV-2 genomes, the main protease (Mpro), as well as a 3-chymotrypsin-like protease (3CLpro), are known as crucial enzymes related to the replication and proliferation of a new virus [1,21]. In particular, SARS-CoV-2 Mpro, corresponding to nsp5, and papain-like protease (PLpro), corresponding to nsp3, first autocleave themselves from the synthesis of messenger RNA (mRNA) translation. The proteases then cleave polyproteins to polypeptides, leading to the replication and functionalities of a new virus. The cleaved proteins involve endoribonuclease (NendoU), exonuclease (exoN), helicase (Hel), RNA-dependent RNA polymerase (RdRp), and an *S*-adenosyl-methionine-dependent ribose 2′-*O*-methyltransferase (2′-*O*-MT). It should be noted that whereas PLpro is responsible for the formation of nsp1-3, Mpro determines the formation of nsp4-16 [22]. Therefore, 3CLpro/Mpro is one of the most appropriate targets for COVID-19 drug design.

3CLpro/Mpro is a homodimer protein consisting of two chains each consisting of 306 residues divided into three domains I, II, and III [23]. Indeed, the active site of Mpro comprises His41 and Cys145 located in the cleft between domains I and II [24,25]. Strong binding inhibitors to Mpro often form tight hydrogen bond (HB) and nonbonded (NBC) contacts with these residues [26]. Moreover, the important residues controlling the ligand-binding affinity to SARS-CoV-2 Mpro include *Thr26*, *Ser46*, *Asn142*, *Gly143*, *His164*, *Glu166*, and *Gln189* [26]. Although the catalytic activity of SARS-CoV-2 Mpro rigidly relies on dimerization [23], investigation of the ligand-binding affinity *in silico* can be performed based on the monomeric form [27].

Because of the well-characterized structure and great interest in designing inhibitors for SARS-CoV-2 Mpro, numerous computational and experimental studies have been carried out to estimate efficient inhibitors to block the biological activity of Mpro [24,25,28-37]. Among these studies, computational approaches are widely used to speed up the screening of potential SARS-CoV-2 inhibitors since several thousand compounds can be tested over a short period of time [27,38-40]. Indeed, computer-aided drug design (CADD) has arisen as a robust protocol for high-throughput screening of thousands/million compounds for potential inhibitors of enzymes since October 5, 1981. An article entitled "Next Industrial Revolution: Designing Drugs by Computer at Merck" was published in *Fortune* magazine [41]. The power of CADD has been increasingly demonstrated because the method remarkably reduces the time and cost of drug development [42]. For example, 81 inhibitors were screened over 400 000 tested compounds, yielding a hit rate of only 0.02% [43]. CADD is used not only to screen for new inhibitors but also to test existing drugs for repurposing targets [34,40]. Therefore, numerous drugs have been discovered thanks to the contribution of CADD. Examples of some of the earliest successes of CADD include the carbonic anhydrase inhibitor dorzolamide, which was authorized in 1995 [44,45], and saquinavir, ritonavir, and indinavir, which were authorized for inhibiting human immunodeficiency virus 1 (HIV-1) protease in 1995, 1996, and 1996, respectively [41].

Typically, the computational approach is utilized to determine promising agents that can bind well to a protein target. Investigation of the ligand-binding free energy is thus one of the most important tasks in CADD [46]. Several methods have been developed to unravel the physical/chemical process [47]. Among these, molecular docking [48] or quantitative structure–activity relationship (QSAR) [49] methods can be used to characterize the ligand-binding affinity of several thousand/million ligands. More accurate methods such as the fast pulling of ligand (FPL) [50], linear interaction energy (LIE) [51,52], and molecular mechanism/Poisson-Boltzmann surface area (MM/PBSA) [53-55] are then used to refine the docking/QSAR outcomes. The free energy perturbation (FEP) method was finally performed to validate the list of promising inhibitors [56-59]. Moreover, recent advancements in machine learning (ML) methods have benefited many areas of science and technology. In particular, ML has been used in CADD for drug discovery and repurposing [60,61]. The most common task of ML in CADD [62] is to predict ligand binding affinities from features extracted from molecular properties including physical, chemical, and structural terms. This is a typical supervised regression problem where ML approaches such as random forest, gradient boosting, and deep learning can result in good prediction accuracy. ML was used to repurpose existing drugs for SARS-CoV-2 treatment [40].

In this work, we combined ML models with atomistic simulations to screen a large database of approximately two million compounds for potential inhibitors of SARS-CoV-2 Mpro. In particular, ML models were trained to predict ligand binding free energies for the whole database, and top-lead ligands with the strongest predicted binding affinity were selected. These top-lead compounds were subsequently subjected to physics-based calculations, including molecular docking and molecular dynamics (MD) simulations, to validate their binding mechanisms to SARS-CoV-2 Mpro. Our resulting list of promising inhibitors for SARS-CoV-2 Mpro can serve as a foundation for further experimental investigations and contribute to the rapid development of SARS-CoV-2 therapy.

2. Materials and methods

2.1. Computational scheme

The computational scheme used to investigate potential inhibitors for SARS-CoV-2 Mpro from ChEMBL [63], a database of bioactive molecules with drug-like properties, is described in Fig. 1. In particular, ML models were trained and tested to accurately determine the binding free energy of ChEMBL compounds to SARS-CoV-2 Mpro. The top 100 potential inhibitors for SARS-CoV-2 Mpro, which were estimated by ML models, were docked to the protease. The structural change of the SARS-CoV-2 Mpro + inhibitor complexes was then investigated using unbiased MD simulations. The Gibbs free energy difference between the *unbound* and *bound* states of the top 100 ligands to Mpro was revealed via LIE and FEP calculations. A list of compounds for inhibiting SARS-CoV-2 Mpro from the ChEMBL database was thus obtained.

2.2. SARS-CoV-2 Mpro and ligands

The list of available inhibitors of SARS-CoV-2 Mpro (**Table S1**) was collected from literature reviews, which involved 571 compounds with the corresponding values of the half-maximal inhibitory concentration (IC_{50}). The experimental ligand-binding free energy was calculated as $\Delta G_{EXP} = RT\ln IC_{50}$, which was based on the approximation that IC_{50} equals the inhibition constant (k_i) and was used as the label for training models. **Table S1** contains the labeled data and SMILES strings of the corresponding compounds. The distribution of the ΔG_{EXP} values is shown in **Figure S1** of the Supporting Information. A total of 451 and 120 compounds were randomly chosen for training and testing, respectively. The performance metrics used for model selection include RMSE, Pearson's R, and Spearman's ρ correlation coefficients. The best model was used to predict ligand-binding affinity for the ChEMBL database, which includes ca. 2 million bioactive compounds with drug properties. Furthermore, the SARS-CoV-2 Mpro structure was downloaded from the Protein Data Bank with ID 7JYC [64].

2.3. Supervised learning calculation

Regression models were trained to predict ligand-binding free energy from molecular features. They included linear regression (LR), random forest (RF), extreme gradient boosting (XGBoost) [65], and a deep learning model based on convolutional networks on graphs (GraphConv) [66]. The LR model, which is simple and therefore less prone to

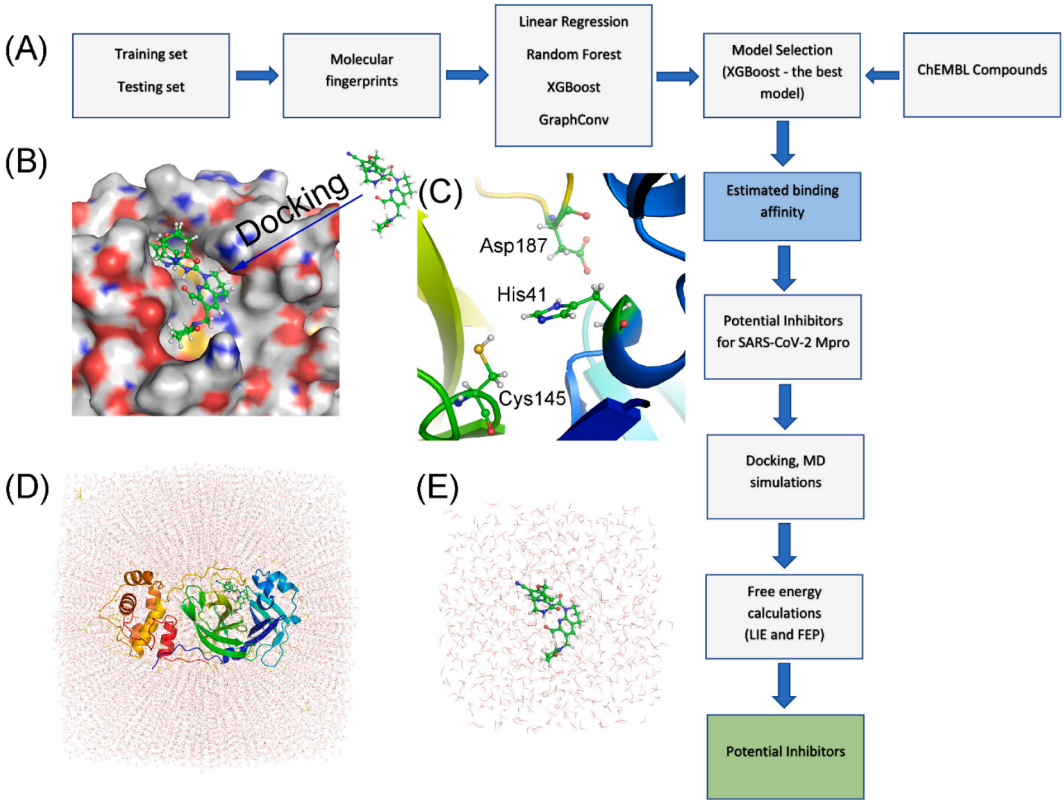

Fig. 1. Computational strategy. (A) Computational approach utilized to search promising inhibitors for SARS-CoV-2 Mpro by using hybrid approach involving supervised machine learning and atomistic simulations. (B) Ligand-binding pose was preliminarily predicted via AutoDock Vina. (C) Configuration of catalytic triad Cys145 - His41 - Asp187. (D) + (E) Initial conformations of SARS-CoV-2 Mpro + inhibitor and individual inhibitors in solution.

overfitting, was used as a baseline model. Model hyperparameters were tuned by using a tenfold cross-validation technique. Optimal values of hyperparameters that minimize mean square error (MSE) were searched by using the Hyperopt library [67]. For LR, only the L2 regularization strength (alpha) was tuned. For RF, the following hyperparameters were tuned: *max_depth*, min_samples_split, *min_samples_leaf* and *max_features*. For XGBoost, *max_depth*, *min_child_weight*, *subsample*, *colsample_bytree*, *reg_lambda*, and *learning_rate* were optimized. For the GraphConv model, the number of units in the graph_cov layers and dense layer, learning rate, and dropout rates were tested. The Python library Scikit-Learn [68] was used to train the LR and RF models. To train XGBoost and GraphConv [66], we used the XGBoost and DeepChem [69] libraries, respectively.

Molecular features were extracted by using the RDKitDescriptors tool kit implemented in Deepchem [69]. It computed 200 physical and chemical properties such as molecular weight, polar surface area, number of valence electrons, numbers of HB donors and acceptors, and maximum and minimum partial charge. To make the models more robust and less prone to overfitting, we reduce the number of features as follows. Some features have zero values for almost all compounds in the training set. Removing features having zero for more than 99% training compound resulted in a reduction of 68 features. Strongly correlated features (absolute value of Pearson's R $>$ 0.95) were also removed, which resulted in a further reduction of 27 features. Finally, 105 features were used as input to the ML models. Some features may have no numerical value for certain compounds. These missing values were imputed by the median of the feature. Before inputting into the models, all features were standardized to have a zero mean and standard deviation of one. These 105 features were used to train the LR, RF, and XGBoost models. GraphConv uses convolutional networks on graphs to automatically extract useful features from the ligand molecule, which is represented as an undirected graph [66]. It is implemented as an embedding layer that accepts the SMILES string as input and outputs a fixed-length vector (called the molecular fingerprint). The fixed-length vector is then fed into a densely connected layer. Both the embedding vectors and weights of densely connected layers are learned simultaneously during training of the model. The Python code for carrying out ML modeling is available online at this GitHub URL https://github.com/nguyentrunghai/SARS-CoV-2Mpro_inhibitor_ML.

2.4. Docking simulations

AutoDock Vina [70] using modified empirical parameters [71] was utilized to investigate the ligand binding poses to SARS-CoV-2 Mpro. In particular, AutoDockTools [72] was used to parameterize both the ligands and receptor. The docking grid center was picked as the center of mass of the native inhibitor. The grid size was picked as 24 $\times$ 24 $\times$ 24 Å, referring to the previous assessment [26]. The exhaustiveness parameter of modified Vina was selected as the default value referring to the previous assessment [71]. The largest energy variability between docking modes was appointed as 7 kcal mol^{-1}. The docking structure with the lowest docking energy was used as the starting conformation of MD simulations.

2.5. Molecular dynamics simulations

MD simulations were performed to refine the docking outcome using GROMACS 5.1.5 [73]. In particular, the Amber99SB-iLDN force field [74] was employed to topologize the protease and charge-neutralizing ions. The catalytic triad Cys145-His41-Asp187 may play a vital role in the binding process of a ligand to SARS-CoV-2 Mpro [75]. The protonation state of these residues was assigned as shown in Fig. 1**C**. Water molecules were parameterized via the TIP3P water model [76]. In addition, the general Amber force field [77] was utilized to parameterize

the ligands using the AmberTools18 and ACPYPE packages [78,79]. Among these, the chemical information, including geometrical parameters and charges, was obtained from quantum mechanics calculations using density functional theory (DFT) with the B3LYP functional and 6-31G(d,p) basis set. The restrained electrostatic potential approach was employed to assign the atomic charges [77].

A dodecahedron periodic boundary condition (dPBC) box with a volume of 669 nm^3 was used to situate the SARS-CoV-2 Mpro + inhibitor complex. In particular, the minimum distance between the complex and the boundary is 1.0 nm. The solvated complex thus comprises ca. 66,000 atoms, respectively, including 1 protease, 1 ligand, 20,400 water molecules, and counterbalanced ions Na^+. In addition, a dPBC box with a capacity of ca. 41 nm^2 was employed to place the ligand, in which the minimum distance between the ligand and the boundary was 1.0 nm. The solvated ligand system hence consists of ca. 3700 atoms, including 1 ligand, 1200 water molecules, and neutralized ions Na^+.

MD parameters were obtained according to previous simulations [26,36]. In particular, nonbonded interactions were cut off at 0.9 nm. The fast particle–mesh Ewald (PME) electrostatics method [80] was utilized to mimic electrostatic (cou) interactions. Moreover, the cutoff scheme was employed to calculate the van der Waals (vdW) interaction. Equilibration simulations were attempted over three steps involving energy minimization, NVT, and NPT simulations. The length of the NVT and NPT simulations is 100 ps. The last snapshot of NPT simulations was then used as the starting shape of MD simulations. The lengths of the MD simulations are 5.0 and 20.0 ns, corresponding to the solvated ligand and complex systems, respectively. The simulation for each system was repeated twice with different random initializations.

2.6. Free energy calculations

Linear interaction energy (LIE) calculation. The binding free energy (ΔG_{LIE}) of a ligand to SARS-CoV-2 Mpro via the LIE approach [81] was computed as the difference between the average of the cou and vdW interactions of the ligand with neighboring atoms over various systems, including the ligand in the solvated complex (*bound* state - noted as subscript *b*) and the ligand in solution (*unbound* state - noted as subscript *u*). ΔG_{LIE} can be determined by.

$$\Delta G_{LIE} = \alpha\left(\left\langle V_{l-s}^{vdW}\right\rangle_b - \left\langle V_{l-s}^{vdW}\right\rangle_u\right) + \beta\left(\left\langle V_{l-s}^{cou}\right\rangle_b - \left\langle V_{l-s}^{cou}\right\rangle_u\right) + \gamma \quad (1)$$

where the coefficients were selected as $\alpha = 0.288$, $\beta = -0.049$, and $\gamma = 5.880$ by referring to the previous assessment [26,33].*Free energy perturbation (FEP) simulation.* The FEP approach [82] was finally used to provide a more accurate estimation of the ligand binding free energy. λ-alteration simulations [83] were employed to change the ligand-binding system from *bound* ($\lambda = 0$) to *unbound* ($\lambda = 1$) states. The free energy alteration between two states, $\Delta G_{\lambda=0\to1} = -k_B T ln\langle e^{-\frac{\Delta H}{k_B T}}\rangle_{\lambda=0}$, can be calculated via several values of the coupling parameter λ over MD simulations. The ligand changing from *bound* to *unbound* states over λ-alteration simulations can be called the ligand-demolition process, in which the energy change can be computed by using the Bennett acceptance ratio (BAR) method [84]. The binding free energy of a ligand to a protein is thus determined as the gap of energy changes over two ligand-demolition processes, including annihilating the ligand in the soluble complex ($\Delta G_{\lambda=0\to1}^{Comp}$) and individual ligand ($\Delta G_{\lambda=0\to1}^{lig}$), as follows:

$$\Delta G_{FEP} = \Delta G_{\lambda=0\to1}^{Comp} - \Delta G_{\lambda=0\to1}^{lig} \quad (2)$$

2.7. Analysis tools

The chemicalize webapp, an online tool of ChemAxon, was utilized to predict the protonation states of ligands. The computed error (success-docking rate and correlation coefficients) was assessed via 1000 rounds of the bootstrapping method [85]. The ligand-binding diagram was produced by using the free version of Maestro [86]. The nonhydrogen atom root-mean-square deviation (RMSD) between docked and experimental binding poses was calculated via GROMACS tools "gmx rms" [73]. A hydrogen bond (HB) contact was counted if the angle ∠ [acceptor (A)-hydrogen (H)-donor (D)] was larger than 135° and the A-D distance was less than 0.35 nm. A nonbonded contact was counted if the distance between two nonhydrogen atoms was lower than 4.5 Å. The clustering calculation was performed via GROMACS tools "gmx cluster." The collective-variable FEL was constructed by using GROMACS tools "gmx sham.".

3. Results and discussion

3.1. Prediction of potential inhibitors from supervised ML models

To rapidly screen the ChEMBL library, four regression models were trained: linear regression (LR), random forest (RF), extreme gradient boosting (XGBoost) [65], and convolutional networks on graphs (GraphConv) [66]. The performance metrics of the test set are listed in Table 1. Due to its simplicity, the LR model is the least accurate among the four models we trained in this work. This is not unexpected since LR is not able to capture nonlinear relationships between features and targets. Nevertheless, the LR model served as a baseline for more sophisticated models. The best model by all three metrics is XGBoost, which gives an RMSE of 1.125 ± 0.095, Pearson's R of 0.748 ± 0.044 and Spearman's ρ of 0.765 ± 0.048 (Table 1). However, it is not better than the second-best model, which is random forest by a large margin. The GraphConv model comes in third place, although its performance is very close to that of random forest. From this assessment, the XGBoost model was selected to predict the binding free energy for nearly 2 million compounds in the ChEMBL database. Fig. 2 shows a comparison of the binding free energy between the experiment and prediction made by XGBoost for 120 compounds in the test set.

The distribution of the predicted binding free energy by the XGBoost model (ΔG_{XGB}) is shown in **Figure S1** of the Supporting Information. The predicted binding free energy for the ChEMBL ligand ranges from −5.78 to −10.02 kcal mol^{-1} with a mean value of −7.47 ± 0.46 kcal mol^{-1}. The top 100 compounds with ΔG_{XGB} values from −9.63 to −10.02 kcal mol^{-1} may be good candidates for SARS-CoV-2 Mpro inhibitors because their predicted binding affinities are comparable to that of nirmatrelvir with $\Delta G_{XGB} = -9.57$ kcal mol^{-1}. The mean value of ΔG_{XGB} is −9.72 ± 0.01 kcal mol^{-1}. The predicted binding free energy is probably underestimated since the ML model was trained and tested on k_i, which was approximated by IC_{50}. The list of compounds is reported in **Table S2** of the Supporting Information. Although the ML model adopted a good correlation coefficient with the respective experiments (cf. Fig. 2), physical-based methods were also carried out to further evaluate the binding process of these ligands to SARS-CoV-2 Mpro.

3.2. Estimation of docking pose of ligands to SARS-CoV-2 Mpro

As mentioned above, although the ML method was shown to be effective in predicting potential inhibitors, physical-based approaches were also utilized to refine the predicted binding affinity and to provide physical and chemical insights into the binding process [87]. In this work, the binding process of SARS-CoV-2 Mpro and its inhibitors was

Table 1
Performance metrics of regression models in predicting binding free energy of 120 tested ligands to SARS-CoV-2 Mpro. Numbers in parentheses are error bars estimated by bootstrapping.

Model	RMSE (kcal mol^{-1})	Pearson's R	Spearman's ρ
Linear Regression	1.299 ± 0.104	0.631 ± 0.070	0.708 ± 0.053
Random Forest	1.157 ± 0.093	0.737 ± 0.046	0.753 ± 0.045
XGBoost	1.125 ± 0.095	0.748 ± 0.044	0.765 ± 0.048
GraphConv	1.161 ± 0.088	0.735 ± 0.050	0.749 ± 0.043

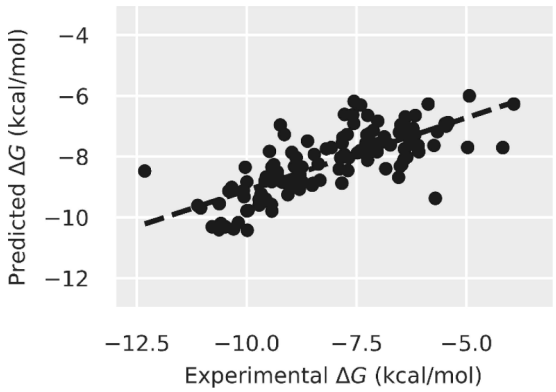

Fig. 2. Predicted binding free energy versus experiment for 120 test compounds. Prediction was made using XGBoost model.

investigated via two steps: molecular docking and MD simulations. In particular, the ligand-binding pose to SARS-CoV-2 Mpro was preliminarily probed via AutoDock Vina [70]. Note that AutoDock Vina provided an appropriate success-docking rate, $\widehat{\rho} = 67\%$, when the docking pose of nine ligands was compared to the respective experiments [26]. In addition, recently, the modified version of AutoDock Vina [71] was benchmarked using the different empirical parameters, which provided a better correlation to the respective experiments. However, the performance of the approach on SARS-CoV-2 Mpro was not considered. Therefore, in this work, the $\widehat{\rho}$ value of AutoDock Vina with original and modified empirical parameters was reevaluated with a larger set involving 40 different complexes (**Table S3** of the Supporting Information). The obtained results were consistent with those of previous work [26]. In particular, the conformations docked via AutoDock Vina differed from the experimental structures by an amount of $RMSD = 1.85 \pm 0.15$ such that the $\widehat{\rho}$ value was 62.5 ± 7.8 %. Interestingly, the corresponding values of the modified AutoDock Vina are 1.80 ± 0.14 Å and 65.0 ± 7.6 %, respectively, which are better than those of AutoDock Vina. Here, a successfully docked shape was counted if the RMSD was lower than 2.00 Å. AutoDock Vina with the modified empirical parameters was thus employed to find the binding pose of the top 100 ligands to SARS-CoV-2 Mpro. The docking outcomes are shown in Table 2 and **Table S5** of the Supporting Information. The ligand-binding free energy ΔG_{Dock} diffuses in the range from −11.3 to −17.6 kcal mol^{-1}, in which the mean value is −14.89 ± 0.10 kcal mol^{-1}. The mean value is significantly smaller than that of nirmatrelvir, $\Delta G_{\text{Dock}} = -13.8$ kcal mol^{-1}, which was considered a positive control. Note that the modified AutoDock Vina often offered an overestimated value of ligand-binding affinity [71]. Moreover, as mentioned above, the predicted value ΔG_{ML} is probably underestimated. This may explain why the obtained docking results were significantly larger than those obtained by the ML models.

3.3. MD-Refined simulations

Docking approaches normally utilize several approximations to speed up the computation [70,72]. Therefore, the obtained outcomes are thus refined by using more accurate and precise approaches such as Steered-MD (SMD), MD, and replica-exchange MD (REMD) simulations [34,37,89]. In this work, although AutoDock Vina with the modified empirical parameters formed the success-docking rate $\widehat{\rho} = 65.0 \pm 7.6$, unbiased MD simulations were employed to clarify the docking results. According to previous work [26], the solvated SARS-CoV-2 Mpro + inhibitor complex was equilibrated by using 20.0 ns unbiased MD simulations for each trajectory, which is long enough to equilibrate the complex conformations. The all-atom RMSD of the complexes almost achieved stable states just after 5.0 ns (cf. **Table S4** of the Supporting Information). The equilibrated conformations of the solvated complex and ligand systems were then used for free energy calculations via LIE/FEP approaches.

The protease individual residues were calculated to reveal the ligand-binding mechanism by analyzing intermolecular NBC and HB contacts between top-lead compounds (Table 2 and discussed below). The probability of NBC and HB contacts of ligands to SARS-CoV-2 Mpro residues was described in Fig. 3. In particular, top-lead ligands adopted intermolecular contacts with 43 residues. The residues formed intermolecular NBCs to top-lead ligands over more than 46% of the evaluated conformations (20,000 shapes in total). In addition, only 24/43 residues adopted intermolecular HB to top-lead compounds, which occupied 6% of all investigated shapes. Note that the outcomes are obviously larger than those of twenty available inhibitors [26]. It may be argued that the top-lead compounds form a stronger ligand-binding affinity than twenty available inhibitors [26]. Moreover, the residues that were able to adopt more than 6% HB and 46% NBC to ligands are critical elements

Table 2
Calculated binding free energy of top-lead compounds to SARS-CoV-2 Mpro via different approaches.

N^0	Compound name	ΔG_{XGB}	ΔG_{Vina}	$\langle V_{l-s}^{\text{cou}} \rangle_b - \langle V_{l-s}^{\text{cou}} \rangle_u$	$\langle V_{l-s}^{\text{vdW}} \rangle_b - \langle V_{l-s}^{\text{vdW}} \rangle_u$	ΔG_{LIE}	ΔG_{cou}	ΔG_{vdW}	ΔG_{FEP}	ΔG_{EXP} [a]
1	CHEMBL3815050	−10.02	−15.7	7.16	−36.37	−16.71 ± 0.36	−2.06	−10.06	−12.12 ± 1.12	
2	CHEMBL4300604	−9.98	−14.7	10.66	−33.23	−15.97 ± 0.49	−9.48	−9.49	−18.97 ± 3.59	
3	CHEMBL3945443	−9.98	−15.1	9.58	−30.44	−15.12 ± 0.03	−16.92	−14.73	−31.65 ± 0.92	
4	CHEMBL3678802	−9.97	−17.1	5.23	−32.15	−15.40 ± 1.39	−9.47	−8.72	−18.19 ± 0.69	
5	CHEMBL4170638	−9.91	−14.0	3.62	−31.78	−15.21 ± 0.36	−3.68	−11.19	−14.87 ± 0.33	
6	CHEMBL4111845	−9.79	−15.3	3.79	−35.84	−16.39 ± 0.58	−12.95	−8.46	−21.41 ± 0.69	
7	CHEMBL580289	−9.75	−15.0	7.88	−32.08	−15.51 ± 0.78	−17.79	−9.45	−27.24 ± 2.87	
8	CHEMBL3640406	−9.74	−15.8	15.84	−29.33	−15.10 ± 0.14	9.97	−14.96	−4.99 ± 1.26	
9	CHEMBL538763	−9.74	−15.1	11.20	−30.55	−15.23 ± 0.78	6.02	−12.13	−6.11 ± 3.77	
10	CHEMBL176909	−9.73	−15.3	14.99	−35.52	−16.84 ± 0.36	−6.99	−12.47	−19.45 ± 1.07	
11	CHEMBL4095929	−9.72	−17.2	18.81	−32.15	−16.06 ± 0.21	9.13	−14.07	−4.95 ± 0.42	
12	CHEMBL3640394	−9.69	−14.4	15.34	−29.54	−15.14 ± 0.30	7.79	−14.77	−6.98 ± 0.57	
13	CHEMBL3657195	−9.68	−15.0	7.87	−30.99	−15.19 ± 0.26	5.07	−8.63	−3.56 ± 1.93	
14	CHEMBL416434	−9.68	−15.9	11.72	−32.35	−15.77 ± 1.97	−7.06	−11.11	−18.17 ± 2.52	
15	CHEMBL1471687	−9.67	−13.8	−8.97	−38.03	−16.39 ± 0.45	−7.19	−13.76	−20.95 ± 0.08	
16	CHEMBL285908	−9.67	−11.3	11.07	−33.31	−16.01 ± 0.99	1.60	−10.40	−8.80 ± 2.00	
17	CHEMBL3673817	−9.67	−16.3	2.23	−32.67	−15.40 ± 0.27	3.57	−12.21	−8.64 ± 1.74	
18	CHEMBL4101092	−9.66	−17.1	14.50	−32.10	−15.83 ± 0.7	−1.73	−16.77	−18.50 ± 1.08	
19	CHEMBL20260	−9.64	−15.8	−3.65	−34.18	−15.55 ± 1.34	1.54	−12.55	−11.01 ± 2.88	
20	Nirmatrelvir	−9.57	−13.8	2.21	−29.80	−14.57 ± 0.06	−4.56	−9.78	−14.35 ± 0.04	−10.46 [88]

[a] ΔG_{EXP} value obtained based on IC50 value, in which term was approximately equal to inhibition constant K_i and contribution of covalent binding energy is assumed to be small [88]. Unit is kcal mol^{-1}.

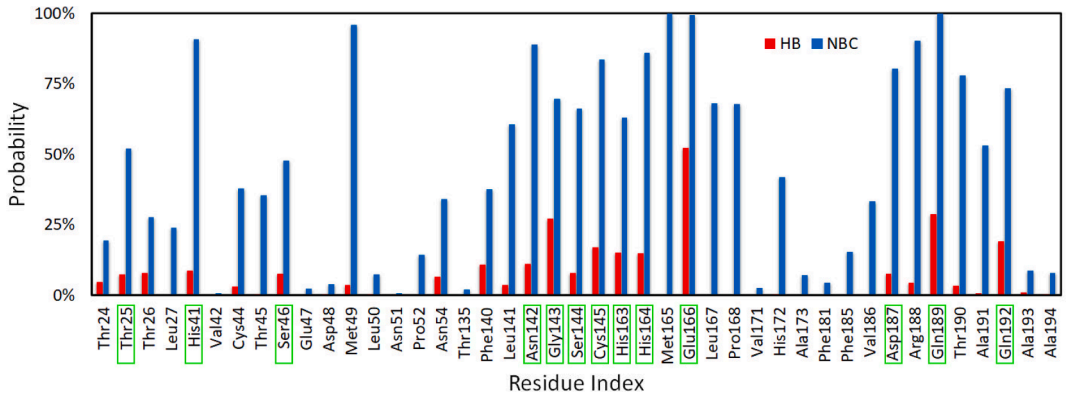

Fig. 3. Probability of NBC and HB contacts between SARS-CoV-2 Mpro individual residues and top-lead compounds. Green rectangles denote residues that formed more than 6% HB and 46% NBC to ligands.

controlling the ligand-binding process to the protease. There are 13 residues satisfying the criteria (Fig. 3). The residues are *Thr25*, *His41*, *Ser46*, *Asn142*, *Gly143*, *Ser144*, *Cys145*, *His163*, *His164*, *Glu166*, *Asp187*, *Gln189*, and *Gln192*. Interestingly, top-lead compounds frequently created intermolecular contacts with the catalytic triad *Cys145-His41-Asp187*, which may prevent the biological activity of SARS-CoV-2 Mpro. In addition, the residues *Met49*, *Leu141*, *Met165*, *Leu167*, *Pro168*, *Arg188*, *Thr190*, and *Ala191* also contribute a large amount of the binding energy because they rigidly adopt NBC to ligands. Overall, top-lead ligands can bind well to Mpro by forming a large number of contacts with several residues located in the active sites. This may prevent the reduction in ligand-binding affinity when mutations of SARS-CoV-2 Mpro occur.

As mentioned above, the catalytic triad Cys145-His41-Asp187 may play a crucial role in the binding process of a ligand to SARS-CoV-2 Mpro [75]. The distance between the Cys145-Sγ and His41-Nε ($d(S\gamma_{Cys145} - N\varepsilon_{His41})$) and His41-Nδ vs. Asp187-Oδ ($d(N\delta_{His41} - O\delta_{Asp187})$) atoms was considered to mediate the influence of ligands on SARS-CoV-2 Mpro. Moreover, SARS-CoV-2 Mpro in solution without ligand was also simulated over 220 ns (**Figure S2** of the Supporting Information) to compare the difference. The collective-variable free energy landscape (FEL) using $d(N\delta_{His41} - O\delta_{Asp187})$ and ($d(S\gamma_{Cys145} - N\varepsilon_{His41})$) as reaction coordinates was constructed over equilibrium snapshots of SARS-CoV-2 Mpro with top-lead inhibitors (Table 1) and without an inhibitor. The outcome is shown in Fig. 4. The presence of inhibitors clearly altered the FEL of SARS-CoV-2 Mpro, implying a change in enzymatic biological activity. In particular, the SARS-CoV-2 Mpro without the presence of ligands formed two minima noted as **A1-2** at ($d(N\delta_{His41} - O\delta_{Asp187})$, ($d(S\gamma_{Cys145} - N\varepsilon_{His41})$) coordinates of (0.50, 0.35) and (0.73, 0.33). In addition, the SARS-CoV-2 Mpro + ligands adopted two minima denoted as **B1-2** at ($d(N\delta_{His41} - O\delta_{Asp187})$, ($d(S\gamma_{Cys145} - N\varepsilon_{His41})$) coordinates of (0.51, 0.34) and (0.29, 0.33).

3.4. Binding free energy calculation via the LIE method.

The binding free energy of ligands and SARS-CoV-2 Mpro can be calculated using the LIE approach with a correlation coefficient of $R_{LIE} = 0.73 \pm 0.09$ [26]. The binding free energy ΔG_{LIE} was thus performed to assess the top 100 inhibitors, which was suggested by the XGBoost model. The mean gap of the cou and vdW interaction energies between a ligand and surrounding molecules over *bound* and *unbound* states was calculated and is shown in Table 2 **and S2** of the Supporting Information. The corresponding values over 100 ligands are 7.18 ± 0.55 and -26.48 ± 0.49 kcal mol^{-1}, respectively. The obtained outcome suggests the domination of vdW over electrostatic interactions in the binding process of ligands to SARS-CoV-2 Mpro, which is in good agreement with previous works [34,36]. This also explained why the empirical parameters for amyloid beta systems involving $\alpha = 0.288$, $\beta = -0.049$, and $\gamma = -5.880$ were successfully applied for computing the binding affinity of a ligand to SARS-CoV-2 Mpro. The negative metrics β and γ correspond to the loss of electrostatic interaction during the ligand association, and the hydrophobic interaction is strong as the domination of the vdW term.

The obtained ΔG_{LIE} is shown in Table 2 **and S2** of the Supporting Information. Over the top 100 compounds, the metric varies in the range from -7.51 ± 3.55 to -16.84 ± 0.36 kcal mol^{-1}, in which the average

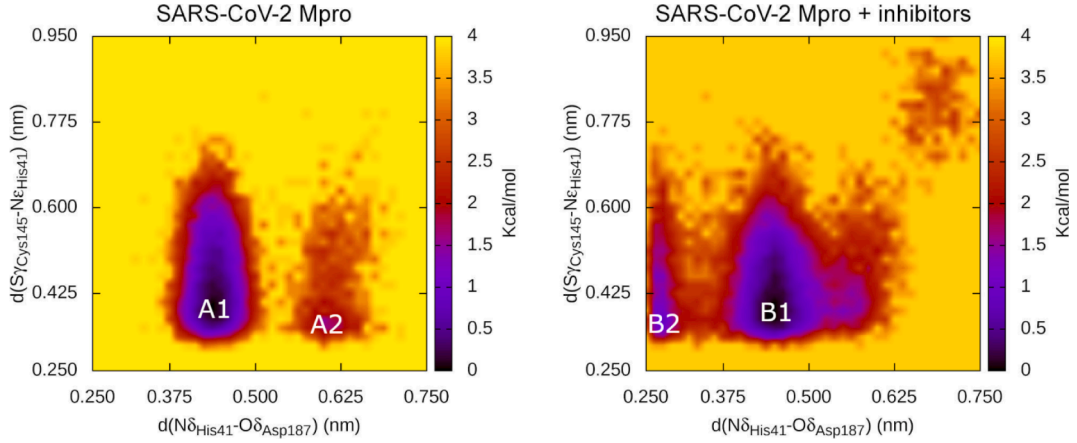

Fig. 4. Collective-variable FEL of SARS-CoV-2 Mpro in present and absent ligands. Distances $d(N\delta_{His41} - O\delta_{Asp187})$ and $(d(S\gamma_{Cys145} - N\varepsilon_{His41})$, which are associated with catalytic triad Cys145 - His41 - Asp187, were utilized as reaction coordinates.

value is -13.87 ± 0.14 kcal mol^{-1}. The magnitude of ΔG_{LIE} is larger than that of ΔG_{ML} by an amount of 4.14 kcal mol^{-1}, which is consistent with the previous work that the ΔG_{LIE} is smaller than ΔG_{EXP} by a value of 3.89 kcal mol^{-1}. Note that the overestimation of ΔG_{LIE} compared with ΔG_{ML} was observed because the ML model was trained using IC_{50} as an approximation to k_i in calculating the experimental binding free energy. Moreover, the obtained ΔG_{LIE} are in good agreement with the docking simulation, which is $\Delta G_{Dock} = -14.89 \pm 0.10$ kcal mol^{-1}. The RMSE between ΔG_{Dock} and ΔG_{LIE} is 1.88 kcal mol^{-1}. Furthermore, 19% of the ligands formed a strong binding free energy to the protease, in which the calculated ΔG_{LIE} was lower than -15.00 kcal mol^{-1} (Table 2). Interestingly, the top 19 inhibitors formed a lower ΔG_{LIE} than that of the positive control nirmatrelvir, whose ΔG_{LIE} is -14.57 ± 0.06 kcal mol^{-1} (Table 2). They were thus selected for further investigation to validate the outcome via the perturbation method [82].

3.5. Investigation of ligand-binding affinity via perturbation simulations

The perturbation simulation is known as one of the most accurate approaches thus far [26,47,90,91]. In a recent report [26], this approach was indicated as the most accurate method for determining the binding free energy of SARS-CoV-2 Mpro and its ligands. The FEP approach formed the highest correlation coefficient, $R_{FEP} = 0.85 \pm 0.06$ [26], for the respective experiment in comparison with other methods such as LIE, fast pulling of ligand (FPL) [50], and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) [53,92]. The perturbation simulations were successfully applied to characterize several potential inhibitors for SARS-CoV-2 Mpro [34,35,90]. The approach was thus utilized to refine the binding free energy of the top 19 inhibitors, which adopted an ΔG_{LIE} smaller than -15.00 kcal mol^{-1}.

The obtained ΔG_{FEP} was mentioned in Table 2. In particular, the value fell in the range from -3.56 ± 1.93 to -31.65 ± 0.92 kcal mol^{-1} with a mean of -14.56 ± 1.74 kcal mol^{-1}. This may suggest that nine compounds having $\Delta G_{FEP} < -15.00$ kcal mol^{-1} would be highly potent inhibitors for SARS-CoV-2 Mpro (Table 2). These compounds include CHEMBL4300604, CHEMBL3945443, CHEMBL3678802, CHEMBL4111845, CHEMBL580289, CHEMBL176909, CHEMBL416434, CHEMBL1471687, and CHEMBL4101092. Interestingly, among nine compounds, CHEMBL3945443, which formed the strongest binding affinity to SARS-CoV-2 Mpro, has a nitrile group, the same as nirmatrelvir; unfortunately, the MD-refined structure (Fig. 5) of the complex did not reveal this issue. This may imply the inaccuracy of conventional MD simulations since they cannot represent the covalent binding between proteins and ligands. Further computation using a quantum chemical approach would thus be employed to reveal the problem. Moreover, using the same approach, the available inhibitors can form -9.18 ± 2.48 (α-ketoamide 13b), -7.73 ± 1.77 (ritonavir), and -14.56 ± 2.65 (proscillaridin) kcal mol^{-1} [26,36]. Furthermore, the binding free energy of PF-07321332 (nirmatrelvir) to SARS-CoV-2 Mpro is -14.35 ± 0.04 kcal mol^{-1}, which was calculated as a positive control. The compound CHEMBL3945443 can thus form a covalent bond to Cys145-Sγ of SARS-CoV-2 Mpro. Furthermore, over all considered systems, the mean electrostatic free energy difference ΔG_{cou} is -2.66 ± 1.91 kcal mol^{-1}, while the average vdW free energy difference ΔG_{vdW} is -11.89 ± 0.54 kcal mol^{-1}. The obtained results indicated that the vdW interaction rules over electrostatic interactions in the binding process of the considered inhibitors to SARS-CoV-2 Mpro. It should be noted that this result is consistent with previous works [26,36].

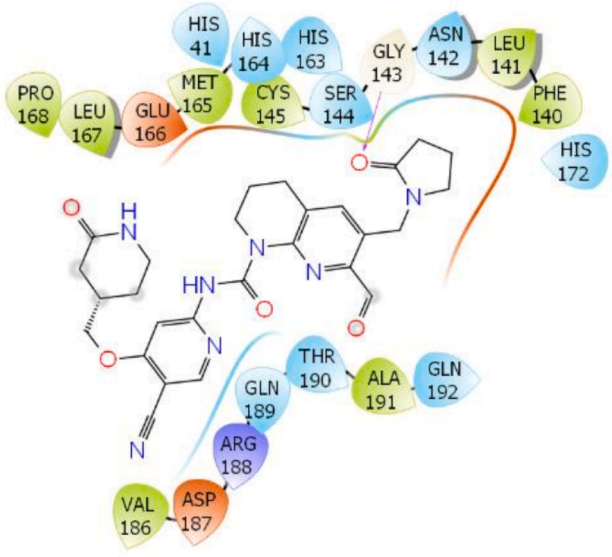

Fig. 5. Interaction diagram between SARS-CoV-2 Mpro + CHEMBL3945443. Complexed structure was obtained by calculating clustering of all equilibrium snapshots of complex with cutoff of 0.2 nm.

4. Conclusions

In this work, a hybrid approach using knowledge- and physical-based methods was proposed to characterize potential inhibitors of SARS-CoV-2 Mpro. Initially, the XGBoost model was trained to screen ligand-binding affinity over a large database of compounds. The predicted binding affinity for a test set strongly correlates with experimental data with a correlation coefficient of $R = 0.748 \pm 0.044$. The ligand-binding free energy ΔG_{XGB} of drug-like compounds from the ChEMBL database [63] was then predicted by the XGboost model.

The top 100 compounds that formed the largest values of ΔG_{ML} were further investigated via atomistic simulations. The binding pose of these ligands to the protease was preliminarily estimated using the modified AutoDock Vina with a successful docking rate of 65.0 ± 7.8 %. The complex was then refined via unbiased MD simulations. Moreover, the equilibrium conformations of the complex were used for binding free energy calculation via the LIE approach. A short list including 19 compounds was suggested for further analysis via the FEP method since adopting $\Delta G_{LIE} < -15.00$ kcal mol^{-1}. Furthermore, the perturbation simulations indicated that 9 compounds (CHEMBL4300604, CHEMBL3945443, CHEMBL3678802, CHEMBL4111845, CHEMBL580289, CHEMBL176909, CHEMBL416434, CHEMBL1471687, and CHEMBL4101092) can bind well to SARS-CoV-2 Mpro with a binding free energy of $\Delta G_{FEP} < -15.00$ kcal mol^{-1}, which is lower than that of nirmatrelvir,$\Delta G_{FEP} = -14.35$ kcal mol^{-1}. It may be argued that these compounds may act as highly potent inhibitors of SARS-CoV-2 Mpro.

The obtained results also suggested that the vdW interaction may play an important role in the binding process of SARS-CoV-2 Mpro + inhibitors. Moreover, the residues *Thr25*, *His41*, *Ser46*, *Asn142*, *Gly143*, *Ser144*, *Cys145*, *His163*, *His164*, *Glu166*, *Asp187*, *Gln189*, and *Gln192* play an important role since HB and NBC were rigidly adopted as ligands. Furthermore, the residues *Met49*, *Leu141*, *Met165*, *Leu167*, *Pro168*, *Arg188*, *Thr190*, and *Ala191* also contribute a large amount of the binding energy since rigidly adopting NBC to ligands. In addition, the ligands inhibit the biological activity of SARS-CoV-2 Mpro by altering FEL construction via $d(N\delta_{His41} - O\delta_{Asp187})$ and $(d(S\gamma_{Cys145} - N\varepsilon_{His41}))$. The catalytic triad Cys145 - His41 - Asp187 was thus disturbed.

gmentsnts

CRediT authorship contribution statement

Trung Hai Nguyen: Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Nguyen Minh Tam:** Methodology, Formal analysis. **Mai Van Tuan:** Project administration, Resources. **Peng Zhan:** Supervision, Conceptualization. **Van V. Vu:** Conceptualization, Visualization. **Duong Tuan**

Quang: Funding acquisition, Supervision. **Son Tung Ngo:** Supervision, Writing – original draft, Writing – review & editing, Investigation, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by Vietnam National Foundation for Science & Technology Development (NAFOSTED) grant #02/2020/ĐX.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemphys.2022.111709.

References

[1] C.M. Fauquet, D. Fargette, International Committee on Taxonomy of Viruses and the 3,142 unassigned species, Virology 2 (2005) 64.
[2] D. Schoeman, B.C. Fielding, Coronavirus envelope protein: current knowledge, Virology 16 (2019) 69.
[3] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M.A. Müller, C. Drosten, S. Pöhlmann, SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor, Cell 181 (2020) 1–10.
[4] E. de Wit, N. van Doremalen, D. Falzarano, V.J. Munster, SARS and MERS: recent insights into emerging coronaviruses, Nat. Rev. Microbiol. 14 (2016) 523–534.
[5] A.R. Fehr, R. Channappanavar, S. Perlman, Middle East Respiratory Syndrome: Emergence of a Pathogenic Human Coronavirus, Annu. Rev. Med. 68 (2017) 387–399.
[6] WHO, Coronavirus disease 2019 (COVID-19) Situation Report - 52, 2020.
[7] C.L. Huang, Y.M. Wang, X.W. Li, L.L. Ren, J.P. Zhao, Y. Hu, L. Zhang, G.H. Fan, J.Y. Xu, X.Y. Gu, Z.S. Cheng, T. Yu, J.A. Xia, Y. Wei, W.J. Wu, X.L. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J.G. Xie, G.F. Wang, R.M. Jiang, Z.C. Gao, Q. Jin, J.W. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (2020) 497-506.
[8] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (2020) 470–473.
[9] C. Yu Wai, Y. Chin-Pang, W. Kwok-Yin, Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like Protease (3CLpro) Structure: Virtual Screening Reveals Velpatasvir, Ledipasvir, and Other Drug Repurposing Candidates, F1000Res 9 (2020) 129.
[10] J.F.W. Chan, S.F. Yuan, K.H. Kok, K.K.W. To, H. Chu, J. Yang, F.F. Xing, J.L. Liu, C. C.Y. Yip, R.W.S. Poon, H.W. Tsoi, S.K.F. Lo, K.H. Chan, V.K.M. Poon, W.M. Chan, J. D. Ip, J.P. Cai, V.C.C. Cheng, H.L. Chen, C.K.M. Hui, K.Y. Yuen, A Familial Cluster of Pneumonia Associated with the 2019 Novel Coronavirus Indicating Person-to-Person Transmission: a Study of a Family Cluster, Lancet 395 (2020) 514-523.
[11] N. van Doremalen, T. Bushmaker, D.H. Morris, M.G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J.L. Harcourt, N.J. Thornburg, S.I. Gerber, J.O. Lloyd-Smith, E. de Wit, V.J. Munster, Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1, N Engl J Med 382 (2020) 1564–1567.
[12] FDA Approves First Treatment for COVID-19, FDA, 2020.
[13] J. Cohen, K. Kupferschmidt, The 'very, very bad look' of remdesivir, the first FDA-approved COVID-19 drug, Sci. News, 2020.
[14] M.L. Holshue, C. DeBolt, S. Lindquist, K.H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, A. Cohn, L. Fox, A. Patel, S.I. Gerber, L. Kim, S. Tong, X. Lu, S. Lindstrom, M.A. Pallansch, W.C. Weldon, H.M. Biggs, T.M. Uyeki, S.K. Pillai, First Case of 2019 Novel Coronavirus in the United States, N. Engl. J. Med. 382 (2020) 929-936.
[15] J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M.-d. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, H.C. Lane, Remdesivir for the Treatment of Covid-19 — Final Report, N. Engl. J. Med. (2020).
[16] Coronavirus (COVID-19) Update: FDA Authorizes First Oral Antiviral for Treatment of COVID-19, U.S. Food and Drug Administration, 2021.
[17] South African SARS-CoV-2 Variant Alarms Scientists, TheScientist, TheScientist, 2021.
[18] H. Tu, M.R. Avenarius, L. Kubatko, M. Hunt, X. Pan, P. Ru, J. Garee, K. Thomas, P. Mohler, P. Pancholi, D. Jones, Distinct Patterns of Emergence of SARS-CoV-2 Spike Variants including N501Y in Clinical Samples in Columbus Ohio, bioRxiv (2021) 2021.2001.2012.426407.
[19] P. Wang, M.S. Nair, L. Liu, S. Iketani, Y. Luo, Y. Guo, M. Wang, J. Yu, B. Zhang, P. D. Kwong, B.S. Graham, J.R. Mascola, J.Y. Chang, M.T. Yin, M. Sobieszczyk, C. A. Kyratsous, L. Shapiro, Z. Sheng, Y. Huang, D.D. Ho, Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7, Nature 593 (2021) 130–135.
[20] M. Hoffmann, P. Arora, R. Groß, A. Seidel, B.F. Hörnich, A.S. Hahn, N. Krüger, L. Graichen, H. Hofmann-Winkler, A. Kempf, M.S. Winkler, S. Schulz, H.-M. Jäck, B. Jahrsdörfer, H. Schrezenmeier, M. Müller, A. Kleger, J. Münch, S. Pöhlmann, SARS-CoV-2 Variants B.1.351 and P.1 Escape from Neutralizing Antibodies, Cell 184 (2021) 2384–2393.
[21] Z. Alex, A. Vladimir, Z. Alexander, Z. Bogdan, T. Victor, B. Dmitry S., P. Daniil, S. Rim, F. Andrey, O. Philipp, Y. Yilin, P. Olga, V. Quentin, A. Alex, I. Yan, Potential COVID-2019 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches, 2020.
[22] B.T. Freitas, I.A. Durie, J. Murray, J.E. Longo, H.C. Miller, D. Crich, R.J. Hogan, R. A. Tripp, S.D. Pegan, Characterization and Noncovalent Inhibition of the Deubiquitinase and deISGylase Activity of SARS-CoV-2 Papain-Like Protease, ACS Infect. Dis. 6 (2020) 2099–2109.
[23] K. Anand, G.J. Palm, J.R. Mesters, S.G. Siddell, J. Ziebuhr, R. Hilgenfeld, Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain, EMBO J 21 (2002) 3213–3224.
[24] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L.W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, H. Yang, Structure of Mpro from SARS-CoV-2 and Discovery of its Inhibitors, Nature 582 (2020) 289–293.
[25] W. Dai, B. Zhang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, F. Liu, C. Li, Y. Li, F. Bai, H. Wang, X. Cheng, X. Cen, S. Hu, X. Yang, J. Wang, X. Liu, G. Xiao, H. Jiang, Z. Rao, L.-K. Zhang, Y. Xu, H. Yang, H. Liu, Structure-based Design of Antiviral Drug Candidates Targeting the SARS-CoV-2 Main Protease, Science 368 (2020) 1331–1335.
[26] S.T. Ngo, N.M. Tam, M.Q. Pham, T.H. Nguyen, Benchmark of Popular Free Energy Approaches Revealing the Inhibitors Binding to SARS-CoV-2 Mpro, J. Chem. Inf. Model. 61 (2021) 2302–2312.
[27] N.M. Tam, P.C. Nam, D.T. Quang, N.T. Tung, V.V. Vu, S.T. Ngo, Binding of Inhibitors to the Monomeric and Dimeric SARS-CoV-2 Mpro, RSC Adv 11 (2021) 2926–2934.
[28] A.D. Rathnayake, J. Zheng, Y. Kim, K.D. Perera, S. Mackin, D.K. Meyerholz, M. M. Kashipathy, K.P. Battaile, S. Lovell, S. Perlman, W.C. Groutas, K.-O. Chang, 3C-like protease inhibitors block coronavirus replication in vitro and improve survival in MERS-CoV–infected mice, Sci. Transl. Med. 12 (2020) eabc5332.
[29] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors, Science 368 (2020) 409–412.
[30] H.-X. Su, S. Yao, W.-F. Zhao, M.-J. Li, J. Liu, W.-J. Shang, H. Xie, C.-Q. Ke, H.-C. Hu, M.-N. Gao, K.-Q. Yu, H. Liu, J.-S. Shen, W. Tang, L.-K. Zhang, G.-F. Xiao, L. Ni, D.-W. Wang, J.-P. Zuo, H.-L. Jiang, F. Bai, Y. Wu, Y. Ye, Y.-C. Xu, Anti-SARS-CoV-2 activities in vitro of Shuanghuanglian preparations and bioactive ingredients, Acta Pharmacol. Sin. 41 (2020) 1167–1177.
[31] C. Ma, M.D. Sacco, B. Hurst, J.A. Townsend, Y. Hu, T. Szeto, X. Zhang, B. Tarbet, M.T. Marty, Y. Chen, J. Wang, Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease, Cell Res 30 (2020) 678–692.
[32] W. Vuong, M.B. Khan, C. Fischer, E. Arutyunova, T. Lamer, J. Shields, H.A. Saffran, R.T. McKay, M.J. van Belkum, M.A. Joyce, H.S. Young, D.L. Tyrrell, J.C. Vederas, M.J. Lemieux, Feline coronavirus drug inhibits the main protease of SARS-CoV-2 and blocks virus replication, Nat. Commun. 11 (2020) 4282.
[33] S.T. Ngo, B.K. Mai, P. Derreumaux, V.V. Vu, Adequate prediction for inhibitor affinity of Aβ40 protofibril using the linear interaction energy method, RSC Adv 9 (2019) 12455–12461.
[34] S.T. Ngo, N. Hung Minh, H. Le Thi Thuy, Q. Pham Minh, T. Vi Khanh, T. Nguyen Thanh, V. Van, Assessing Potential Inhibitors for SARS-CoV-2 Main Protease from Available Drugs using Free Energy Perturbation Simulations, RSC Adv 10 (2020) 40284-40290.
[35] Z. Li, X. Li, Y.-Y. Huang, Y. Wu, R. Liu, L. Zhou, Y. Lin, D. Wu, L. Zhang, H. Liu, X. Xu, K. Yu, Y. Zhang, J. Cui, C.-G. Zhan, X. Wang, H.-B. Luo, Identify Potent SARS-CoV-2 Main Protease Inhibitors via Accelerated Free Energy Perturbation-Based Virtual Screening of Existing Drugs, Proc. Natl. Acad. Sci. U.S.A. 117 (2020) 27381–27387.
[36] S.T. Ngo, N. Quynh Anh Pham, L. Thi Le, D.-H. Pham, V.V. Vu, Computational Determination of Potential Inhibitors of SARS-CoV-2 Main Protease, J. Chem. Inf. Model. 60 (2020) 5771–5780.
[37] M.Q. Pham, K.B. Vu, T.N. Han Pham, L.T. Thuy Huong, L.H. Tran, N.T. Tung, V. V. Vu, T.H. Nguyen, S.T. Ngo, Rapid prediction of possible inhibitors for SARS-CoV-2 main protease using docking and FPL simulations, RSC Adv 10 (2020) 31991–31996.
[38] A.-S. Abd Al-Aziz A., A. Ibrahim, Y. Arpita, P. Raymond A., Computational Design of Potent Inhibitors for SARS-CoV-2's Main Protease, 2020.
[39] A. Francés-Monerris, C. Hognon, T. Miclot, C. García-Iriepa, I. Iriepa, A. Terenzi, S. Grandemange, G. Barone, M. Marazzi, A. Monari, Molecular Basis of SARS-CoV-2 Infection and Rational Design of Potential Antiviral Agents: Modeling and Simulation Approaches, J. Proteome Res. 19 (2020) 4291–4315.
[40] K. Gao, D.D. Nguyen, J. Chen, R. Wang, G.-W. Wei, Repositioning of 8565 Existing Drugs for COVID-19, J. Phys. Chem. Lett 11 (2020) 5373–5382.
[41] J.H. Van Drie, Computer-aided drug design: the next 20 years, J Comput Aided Mol Des 21 (2007) 591–601.

[42] G.R. Marshall, Computer-Aided Drug Design, Ann. Rev. Pharmacol. Toxicol. 27 (1987) 193–213.
[43] T.N. Doman, S.L. McGovern, B.J. Witherbee, T.P. Kasten, R. Kurumbail, W. C. Stallings, D.T. Connolly, B.K. Shoichet, Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B, J. Med. Chem 45 (2002) 2213–2221.
[44] R. Vijayakrishnan, Structure-based drug design and modern medicine, J. Postgrad. Med. 55 (2009) 301–304.
[45] G. Sliwoski, S. Kothiwale, J. Meiler, E.W. Lowe, Computational Methods in Drug Discovery, Pharmacol. Rev. 66 (2014) 334–395.
[46] W. Yu, A.D. MacKerell, Computer-Aided Drug Design Methods, in: P. Sass (Ed.), Antibiotics: Methods and Protocols, Springer, New York, New York, NY, 2017, pp. 85–106.
[47] U. Ryde, P. Soderhjelm, Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods, Chem. Rev. 116 (2016) 5520–5566.
[48] D.K. Gehlhaar, G. Verkhivker, P.A. Rejto, D.B. Fogel, L.J. Fogel, S.T. Freer, Docking Conformationally Flexible Small Molecules into a Protein Binding Site through Evolutionary Programming, in: M.D. John R, R. Robert G, F. David B (Eds.), Proceedings of the Fourth International Conference on Evolutionary Programming: 1-3 March 1995; San Diego, MIT Press1995.
[49] J. Yang, J. Chen, QSAR Analysis of Purine-Type and Propafenone-Type Substrates of P-Glycoprotein Targeting β-Amyloid Clearance 2013.
[50] S.T. Ngo, H.M. Hung, M.T. Nguyen, Fast and Accurate Determination of the Relative Binding Affinities of Small Compounds to HIV-1 Protease using Non-Equilibrium Work, J. Comput. Chem. 37 (2016) 2734–2742.
[51] J. Aqvist, C. Medina, J.-E. Samuelsson, A New Method for Predicting Binding Affinity in Computer-Aidded Drug Design, Protein Eng. 7 (1994) 385–391.
[52] D.K. Jones-Hertzog, W.L. Jorgensen, Binding Affinities for Sulfonamide Inhibitors with Human Thrombin Using Monte Carlo Simulations with a Linear Response Method, J. Med. Chem. 40 (1997) 1539–1549.
[53] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models, Acc. Chem. Res. 33 (2000) 889–897.
[54] B. Kuhn, P.A. Kollman, Binding of a siverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models, J. Med. Chem. 43 (2000) 3786–3791.
[55] W. Wang, P.A. Kollman, Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance, Proc. Natl. Acad. Sci. USA 98 (2001) 14937–14942.
[56] S.T. Ngo, T.H. Nguyen, N.T. Tung, P.C. Nam, K.B. Vu, V.V. Vu, Oversampling Free Energy Perturbation Simulation in Determination of the Ligand-Binding Free Energy, J. Comput. Chem n/a (2019).
[57] W. Jiang, B. Roux, Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations, J. Chem. Theory Comput. 6 (2010) 2559–2565.
[58] Y. Meng, D. Sabri Dashti, A.E. Roitberg, Computing Alchemical Free Energy Differences with Hamiltonian Replica Exchange Molecular Dynamics (H-REMD) Simulations, J. Chem. Theory Comput. 7 (2011) 2721–2727.
[59] W. Jiang, J. Thirman, S. Jo, B. Roux, Reduced Free Energy Perturbation/ Hamiltonian Replica Exchange Molecular Dynamics Method with Unbiased Alchemical Thermodynamic Axis, J. Phys. Chem. B 122 (2018) 9435–9442.
[60] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, R.K. Tekade, Artificial Intelligence in Drug Discovery and Development, Drug Discov. Today 26 (2020) 80–93.
[61] A.N. Ramesh, C. Kambhampati, J.R.T. Monson, P.J. Drew, Artificial intelligence in medicine, Ann R Coll Surg Engl 86 (2004) 334–338.
[62] M.J. Lamberti, M. Wilkinson, B.A. Donzanti, G.E. Wohlhieter, S. Parikh, R. G. Wilkins, K. Getz, A Study on the Application and Use of Artificial Intelligence to Support Drug Development, Clin Ther 41 (2019) 1414–1426.
[63] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J.F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A.R. Leach, ChEMBL: towards direct deposition of bioassay data, Nucleic Acids Res 47 (2018) D930–D940.
[64] B. Andi, D. Kumaran, D.F. Kreitler, A.S. Soares, W. Shi, J. Jakoncic, M.R. Fuchs, J. Keereetaweep, J. Shanklin, S. McSweeney, Hepatitis C Virus NSP3/NSP4A Inhibitors as Promising Lead Compounds for the Design of New Covalent Inhibitors for SARS-CoV-2 3CLpro/Mpro Protease, 2020.
[65] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016) 785–794.
[66] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional Networks on Graphs for Learning Molecular Fingerprints, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.2015.
[67] J. Bergstra, D. Yamins, D. Cox, Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, Proceedings of the 30th International Conference on Machine Learning 28 (2013) 115-123.
[68] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[69] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, Z. Wu, Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More, O'Reilly Media2019.
[70] O. Trott, A.J. Olson, Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading, J. Comput. Chem. 31 (2010) 455–461.
[71] T.N.H. Pham, T.H. Nguyen, N.M. Tam, T.Y. Vu, N.T. Pham, N.T. Huy, B.K. Mai, N. T. Tung, M.Q. Pham, V.V. Vu, S.T. Ngo, Improving Ligand-Ranking of AutoDock Vina by Changing the Empirical Parameters, J. Comput. Chem. 43 (2021).
[72] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A. J. Olson, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, J. Comput. Chem. 30 (2009) 2785–2791.
[73] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindahl, GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers, SoftwareX 1–2 (2015) 19–25.
[74] A.E. Aliev, M. Kulke, H.S. Khaneja, V. Chudasama, T.D. Sheppard, R.M. Lanigan, Motional Timescale Predictions by Molecular Dynamics Simulations: Case Study using Proline and Hydroxyproline Sidechain Dynamics, Proteins: Struct., Funct., Bioinf. 82 (2014) 195–215.
[75] S.T. Ngo, T.H. Nguyen, N.T. Tung, B.K. Mai, Insights into the Binding and Covalent Inhibition Mechanism of PF-07321332 to SARS-CoV-2 Mpro, RSC Adv 12 (2022) 3729–3737.
[76] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of Simple Potential Functions for Simulating Liquid Water, J. Chem. Phys. 79 (1983) 926–935.
[77] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, Development and Testing of a General Amber Force Field, J. Comput. Chem. 25 (2004) 1157–1174.
[78] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E.C. Cheatham, III, V.W.D. , T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R. Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, Y. D.M., a.K. P.A., AMBER 18, University of California, San Francisco (2018).
[79] A.W. Sousa da Silva, W.F. Vranken, ACPYPE - AnteChamber PYthon Parser interfacE, BMC Research Notes 5 (2012) 1–8.
[80] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089–10092.
[81] H. Gutiérrez-de-Terán, J. Åqvist, Linear Interaction Energy: Method and Applications in Drug Design, in: R. Baron (Ed.), Computational Drug Discovery and Design, Springer, New York, 2012, pp. 305–323.
[82] R.W. Zwanzig, High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases, J. Chem. Phys. 22 (1954) 1420–1426.
[83] F.M. Ytreberg, Absolute FKBP Binding Affinities Obtained via Non-Equilibrium Unbinding Simulations, J. Chem. Phys. 130 (2009), 164906.
[84] C.H. Bennett, Efficient estimation of free energy differences from Monte Carlo data, J. Comput. Phys. 22 (1976) 245–268.
[85] B. Efron, Bootstrap Methods: Another Kook at the Jackknife, Ann. Stat. 7 (1979) 1–26.
[86] P. Schrödinger LLC, Schrödinger Release 2020-4: Maestro, 2020.
[87] G. Subramanian, B. Ramsundar, V. Pande, R.A. Denny, Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches, J. Chem. Inf. Model. 56 (2016) 1936–1949.
[88] J. Li, C. Lin, X. Zhou, F. Zhong, P. Zeng, Y. Yang, Y. Zhang, B. Yu, X. Fan, P.J. McCormick, R. Fu, Y. Fu, H. Jiang, J. Zhang, Structural basis of main proteases of coronavirus bound to drug candidate PF-07321332, bioRxiv (2021) 2021.2011.2005.467529.
[89] J. Gera, T. Szögi, Z. Bozsó, L. Fülöp, E.E. Barrera, A.M. Rodriguez, L. Méndez, C.M. L. Delpiccolo, E.G. Mata, F. Cioffi, K. Broersen, G. Paragi, R.D. Enriz, Searching for Improved Mimetic Peptides Inhibitors Preventing Conformational Transition of Amyloid-β42 Monomer, Bioorg. Chem. 81 (2018) 211–221.
[90] C.-H. Zhang, E.A. Stone, M. Deshmukh, J.A. Ippolito, M.M. Ghahremanpour, J. Tirado-Rives, K.A. Spasov, S. Zhang, Y. Takeo, S.N. Kudalkar, Z. Liang, F. Isaacs, B. Lindenbach, S.J. Miller, K.S. Anderson, W.L. Jorgensen, Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations, ACS Cent Sci 7 (2021) 467–475.
[91] S. Decherchi, A. Cavalli, Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation, Chem. Rev. 120 (2020) 12788–12833.
[92] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, D.A. Case, Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices, J. Am. Chem. Soc. 120 (1998) 9401–9409.