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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Searching for potential inhibitors of SARS-COV-2 main protease using supervised learning and perturbation calculations

Trung Hai Nguyen, Nguyen Minh Tam, Mai Van Tuan, Peng Zhan, Van V. Vu, Duong Tuan Quang, Son Tung Ngo

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 ± 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
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 polypeptides 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-adenosylmethionine-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 Glu189 [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 (IC50). The experimental ligand-binding free energy was calculated as \( \Delta G_{\text{EXP}} = RT \ln IC_{50} \), which was based on the approximation that IC50 equals the inhibition constant (Ki) 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 \( \Delta G_{\text{EXP}} \) 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 \( \rho \) 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 7JYJC [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
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_conv 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 × 24 × 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. 1C. Water molecules were parameterized via the TIP3P water model [76]. In addition, the general Amber force field [77] was utilized to parameterize

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.
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 ($\Delta G_{\text{bind}}$) 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 $u$) and the ligand in solution (unbound state - noted as $l$). $\Delta G_{\text{LIE}}$ can be determined by:

$$\Delta G_{\text{LIE}} = \alpha (\langle V_{\text{vdW}}^l \rangle - \langle V_{\text{vdW}}^u \rangle) + \beta (\langle V_{\text{cou}}^l \rangle - \langle V_{\text{cou}}^u \rangle) + \gamma$$

(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. $\lambda$-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, $\Delta G_{\lambda=0\rightarrow1} = -k_b T \ln (e^{\Delta G_{\lambda=0\rightarrow1}/k_b T})$, can be calculated via several values of the coupling parameter $\lambda$ over MD simulations. The ligand changing from bound to unbound states over $\lambda$-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 ($\Delta G_{\text{Comp}}^{\text{Comp}}$) and individual ligand ($\Delta G_{\text{individual}}^{\text{Comp}}$), as follows:

$$\Delta G_{\text{FEP}} = \Delta G_{\text{Comp}}^{\text{Comp}} - \Delta G_{\text{individual}}^{\text{Comp}}$$

(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 $z$ (acceptor (A)-hydrogen (H)-donor (D)) was larger than 135$^\circ$ 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, Pearson’s R of 0.748 ± 0.044 and Spearman’s $\rho$ of 0.765 ± 0.048. 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 ($\Delta G_{\text{XGB}}$) 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 $-7.47 \pm 0.46$ kcal mol$^{-1}$. The top 100 compounds with $\Delta G_{\text{XGB}}$ 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 $\Delta G_{\text{XGB}} = -9.57$ kcal mol$^{-1}$. The mean value of $\Delta G_{\text{XGB}}$ is $-9.72 \pm 0.01$ kcal mol$^{-1}$. The predicted binding free energy is probably underestimated since the ML model was trained and tested on $k$, which was approximated by $k_{\text{CGD}}$. 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

| Model        | RMSE (kcal mol$^{-1}$) | Pearson’s R | Spearman’s $\rho$ |
|--------------|------------------------|-------------|-------------------|
| 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     |

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.
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, $\hat{\rho} = 67\%$, 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 $\hat{\rho}$ 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 $\text{RMSE} = 1.85 \pm 0.15$ such that the $\hat{\rho}$ value was $62.5 \pm 7.8\%$. Interestingly, the corresponding values of the modified AutoDock Vina are $1.80 \pm 0.14$ Å and $65.0 \pm 7.6\%$, 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.

### 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 $\hat{\rho} = 65.0 \pm 7.6\%$, 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

| N° | Compound name | $\Delta G_{\text{GB}}$ | $\Delta G_{\text{vina}}$ | $\Delta G_{\text{FEP}}$ | $\Delta G_{\text{LEK}}$ | $\Delta G_{\text{HB}}$ | $\Delta G_{\text{vina,HB}}$ | $\Delta G_{\text{FEP,HB}}$ | $\Delta G_{\text{EXD}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | CHEMBL315050 | $-10.02$ | $-15.7$ | 1.76 | $-16.37$ | $-16.71$ | $-16.06$ | $-16.06$ | $-12.12$ |
| 2 | CHEMBL4300604 | $-11.98$ | $-19.7$ | 3.14 | $-20.78$ | $-20.89$ | $-20.55$ | $-20.55$ | $-12.12$ |
| 3 | CHEMBL3945443 | $-13.98$ | $-23.7$ | 7.58 | $-23.43$ | $-23.67$ | $-23.33$ | $-23.33$ | $-12.12$ |
| 4 | CHEMBL3788802 | $-15.98$ | $-26.7$ | 9.32 | $-27.08$ | $-27.12$ | $-26.78$ | $-26.78$ | $-12.12$ |
| 5 | CHEMBL4170638 | $-17.98$ | $-30.7$ | 11.58 | $-31.04$ | $-31.08$ | $-30.74$ | $-30.74$ | $-12.12$ |
| 6 | CHEMBL4111845 | $-19.98$ | $-34.7$ | 13.84 | $-35.72$ | $-35.76$ | $-35.42$ | $-35.42$ | $-12.12$ |
| 7 | CHEMBL580289 | $-22.98$ | $-40.7$ | 16.10 | $-40.96$ | $-40.99$ | $-40.65$ | $-40.65$ | $-12.12$ |
| 8 | CHEMBL3640046 | $-24.98$ | $-46.7$ | 18.36 | $-46.92$ | $-46.95$ | $-46.61$ | $-46.61$ | $-12.12$ |
| 9 | CHEMBL538763 | $-26.98$ | $-52.7$ | 20.62 | $-52.97$ | $-52.99$ | $-52.65$ | $-52.65$ | $-12.12$ |
| 10 | CHEMBL76909 | $-28.98$ | $-58.7$ | 22.88 | $-58.93$ | $-58.95$ | $-58.61$ | $-58.61$ | $-12.12$ |
| 11 | CHEMBL4095929 | $-31.98$ | $-64.7$ | 25.14 | $-64.98$ | $-64.99$ | $-64.65$ | $-64.65$ | $-12.12$ |
| 12 | CHEMBL3640394 | $-34.98$ | $-70.7$ | 27.40 | $-70.98$ | $-70.99$ | $-70.65$ | $-70.65$ | $-12.12$ |
| 13 | CHEMBL2657195 | $-37.98$ | $-76.7$ | 29.66 | $-76.97$ | $-76.98$ | $-76.64$ | $-76.64$ | $-12.12$ |
| 14 | CHEMBL416434 | $-40.98$ | $-82.7$ | 31.92 | $-82.97$ | $-82.98$ | $-82.64$ | $-82.64$ | $-12.12$ |
| 15 | CHEMBL1471687 | $-43.98$ | $-88.7$ | 34.18 | $-88.97$ | $-88.98$ | $-88.64$ | $-88.64$ | $-12.12$ |
| 16 | CHEMBL2657908 | $-46.98$ | $-94.7$ | 36.44 | $-94.97$ | $-94.98$ | $-94.64$ | $-94.64$ | $-12.12$ |
| 17 | CHEMBL3673817 | $-49.98$ | $-100.7$ | 38.70 | $-100.97$ | $-100.98$ | $-100.64$ | $-100.64$ | $-12.12$ |
| 18 | CHEMBL4101002 | $-52.98$ | $-106.7$ | 40.96 | $-106.97$ | $-106.98$ | $-106.64$ | $-106.64$ | $-12.12$ |
| 19 | CHEMBL20260 | $-55.98$ | $-112.7$ | 43.22 | $-112.97$ | $-112.98$ | $-112.64$ | $-112.64$ | $-12.12$ |
| 20 | Nirmatrelvir | $-58.98$ | $-118.7$ | 45.48 | $-118.97$ | $-118.98$ | $-118.64$ | $-118.64$ | $-12.12$ |

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

$\Delta G_{\text{EXD}}$ value obtained based on IC50 value, in which term was approximately equal to inhibition constant $K$, and contribution of covariant binding energy is assumed to be small [88]. Unit is kcal mol \(^{-1}\).
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-S and His41-Nε (d(S\text{Cys145} - \text{NεHis41})) and His41-Nδ vs. Asp187-Oδ (d(\text{NδHis41} - \text{OδAsp187})) 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(\text{NδHis41} - \text{OδAsp187}) and (d(\text{S\text{Cys145} - NεHis41})) 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(\text{NδHis41} - \text{OδAsp187})), (d(\text{S\text{Cys145} - NεHis41})) 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(\text{NδHis41} - \text{OδAsp187})), (d(\text{S\text{Cys145} - NεHis41})) 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\text{LIE} = 0.73 ± 0.09 [26]. The binding free energy ΔG\text{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 ± 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 α = 0.288, β = −0.049, and γ = −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\text{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
value is \(-13.87 \pm 0.14\) kcal mol\(^{-1}\). The magnitude of \(\Delta G_{\text{ML}}\) is larger than that of \(\Delta G_{\text{ML}}\) by an amount of 4.14 kcal mol\(^{-1}\), which is consistent with the previous work that the \(\Delta G_{\text{ML}}\) is smaller than \(\Delta G_{\text{EXP}}\) by a value of 3.89 kcal mol\(^{-1}\). Note that the overestimation of \(\Delta G_{\text{ML}}\) compared with \(\Delta G_{\text{ML}}\) was observed because the ML model was trained using \(\Delta G_{\text{ML}}\) as an approximation to \(\Delta G_{\text{ML}}\) in calculating the experimental binding free energy. Moreover, the obtained \(\Delta G_{\text{ML}}\) are in good agreement with the docking simulation, which is \(\Delta G_{\text{Dock}} = -14.89 \pm 0.10\) kcal mol\(^{-1}\). The RMSE between \(\Delta G_{\text{Dock}}\) and \(\Delta G_{\text{ML}}\) is 1.88 kcal mol\(^{-1}\). Furthermore, 19% of the ligands formed a strong binding free energy to the protease, in which the calculated \(\Delta G_{\text{ML}}\) was lower than \(-15.00\) kcal mol\(^{-1}\) (Table 2). Interestingly, the top 19 inhibitors formed a lower \(\Delta G_{\text{ML}}\) than that of the positive control nirmatrelvir, whose \(\Delta G_{\text{ML}}\) is \(-14.57 \pm 0.06\) 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, \(R_{\text{FEP}} = 0.85 \pm 0.06\) [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 \(\Delta G_{\text{ML}}\) smaller than \(-15.00\) kcal mol\(^{-1}\).

The obtained \(\Delta G_{\text{FEP}}\) was mentioned in Table 2. In particular, the value fell in the range from \(-3.56 \pm 1.93\) to \(-31.65 \pm 0.92\) kcal mol\(^{-1}\) with a mean of \(-14.56 \pm 1.74\) kcal mol\(^{-1}\). This may suggest that nine compounds having \(\Delta G_{\text{FEP}} < -15.00\) 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 \(-9.18 \pm 2.48\) \((\alpha\text{-ketamidie 13b})\), \(-7.73 \pm 1.77\) (ritonavir), and \(-14.56 \pm 2.65\) (proscillaridin) kcal mol\(^{-1}\) [26,36]. Furthermore, the binding free energy of PF-07321332 (nirmatrelvir) to SARS-CoV-2 Mpro is \(-14.35 \pm 0.04\) kcal mol\(^{-1}\), which was calculated as a positive control. The compound CHEMBL3945443 can thus form a covalent bond to Cys145-S of SARS-CoV-2 Mpro. Furthermore, over all considered systems, the mean electrostatic free energy difference \(\Delta G_{\text{E}}\) is \(-2.66 \pm 1.91\) kcal mol\(^{-1}\), while the average vdW free energy difference \(\Delta G_{\text{vdW}}\) is \(-11.89 \pm 0.54\) 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].

### 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 \(\Delta G_{\text{XGB}}\) 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 \(\Delta G_{\text{ML}}\) 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 \pm 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 \(\Delta G_{\text{LIE}} < -15.00\) 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 \(\Delta G_{\text{FEP}} < -15.00\) kcal mol\(^{-1}\), which is lower than that of nirmatrelvir, \(\Delta G_{\text{FEP}} = -14.35\) 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 \(d(N_{\text{O41}} - O_{\text{A191}})\) and \(d(S_{\text{Cys145}} - N_{\text{H41}})\). The catalytic triad Cys145 - His41 - Asp187 was thus disturbed.

---

**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**
