RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2024, **14**, 29683

Identification of lead inhibitors for 3CLpro of SARS-CoV-2 target using machine learning based virtual screening, ADMET analysis, molecular docking and molecular dynamics simulations†

Sandeep Poudel Chhetri,[a] Vishal Singh Bhandari,[b] Rajesh Maharjan[a] and Tika Ram Lamichhane *[a]

The SARS-CoV-2 3CLpro is a critical target for COVID-19 therapeutics due to its role in viral replication. We employed a screening pipeline to identify novel inhibitors by combining machine learning classification with similarity checks of approved medications. A voting classifier, integrating three machine learning classifiers, was used to filter a large database (~10 million compounds) for potential inhibitors. This ensemble-based machine learning technique enhances overall performance and robustness compared to individual classifiers. From the screening, three compounds M1, M2 and M3 were selected for further analysis. Absorption, distribution, metabolism, excretion, and toxicity (ADMET) analysis compared these candidates to nirmatrelvir and azvudine. Molecular docking followed by 200 ns MD simulations showed that only M1 (6-[2,4-bis(dimethylamino)-6,8-dihydro-5*H*-pyrido[3,4-d]pyrimidine-7-carbonyl]-1*H*-pyrimidine-2,4-dione) remained stable. For azvudine and M1, the estimated median lethal doses are 1000 and 550 mg kg^{-1}, respectively, with maximum tolerated doses of 0.289 and 0.614 log mg per kg per day. The predicted inhibitory activity of M1 is 7.35, similar to that of nirmatrelvir. The binding free energy based on Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) of M1 is -18.86 ± 4.38 kcal mol^{-1}, indicating strong binding interactions. These findings suggest that M1 merits further investigation as a potential SARS-CoV-2 treatment.

Received 20th June 2024
Accepted 4th September 2024

DOI: 10.1039/d4ra04502e

rsc.li/rsc-advances

1 Introduction

The SARS coronavirus (SARS-CoV) causes severe acute respiratory syndrome (SARS).[1] As of January 21, 2024, there were 774 395 593 confirmed cases of SARS-CoV-2 infection, resulting in 7 023 271 deaths.[2] SARS-CoV-2, an enveloped positive-sense single-stranded RNA virus,[3] belongs to the genus Betacoronavirus. Viral proteases, crucial for replication, are well-validated targets for treating hepatitis C and HIV.[4] The primary protease, 3CLpro (also known as Mpro or Nsp5),[5] cleaves polyproteins at 11 sites, essential for viral protein maturation.[6] Inhibiting 3CLpro halts viral replication by preventing the production of necessary enzymes like RNA-dependent RNA polymerase.[7] Human proteases lack 3CLpro's cleavage specificity, making these inhibitors safe for human use.[8] Known oral 3CLpro inhibitors[9] are shown in ESI Fig. S1.†

The COVID-19 pandemic has increased the demand for new antiviral drugs. Traditional high-throughput screening (HTS) of 1 to 2 million compounds is expensive and operationally challenging.[10,11] Artificial Intelligence (AI) can accelerate drug discovery by evaluating vast data, predicting drug efficacy, and reducing the time and resources needed for clinical trials, enhancing the chances of developing effective treatments. Drug discovery has been revolutionized over the last ten years by AI models.[12–14]

As discussed in ref. 15, we used machine learning combined with similarity analysis, ADMET analysis, molecular docking and MD simulation in our study. We used a voting classifier to screen a large database (~10 million compounds) for potential inhibitors. Selected compounds were compared to known 3CLpro inhibitors and analyzed for ADMET properties. Stability was assessed using molecular docking and molecular dynamics simulations.[16] Fig. 1 illustrates our study's workflow.

2 Materials and methods

2.1 Data collection and curation

The OpenCADD platform, an open-source tool for chem-informatics, was employed to obtain compound data and develop machine learning models.[17] Simplified Molecular Input Line Entry System (SMILES) for 903 inhibitors of 3CLpro, along with their respective Half-maximal inhibitory concentration

[a]*Central Department of Physics, Tribhuvan University, Kathmandu 44600, Nepal. E-mail: tika.lamichhane@cdp.tu.edu.np*

[b]*Central Department of Chemistry, Tribhuvan University, Kathmandu 44600, Nepal*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4ra04502e

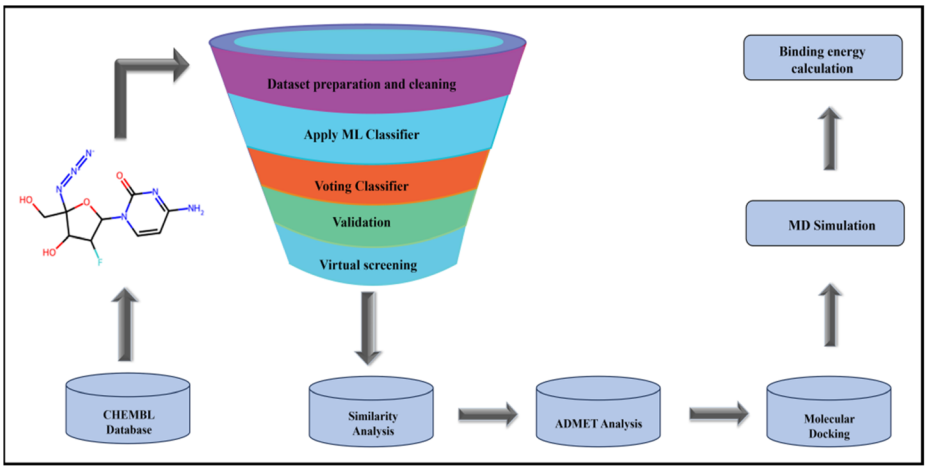

Fig. 1 Schematic workflow of the study.

(IC50) values, were retrieved from the Chemical European Molecular Biology Laboratory (CHEMBL) database.[18] After downloading the data, we filtered out SMILES entries lacking IC50 values, retained only bioactivity entries measured in nanomolar (nM), and removed duplicate molecules, resulting in 744 data points. Due to the varied scales of IC50 values, they were converted into corresponding negative logarithms, known as pIC50 values. Pfizer's rule, also known as Lipinski's Rule of Five (RO5), was utilized at this stage to filter the data according to drug-likeness.[19,20] Meeting most of the Ro5 parameters does not ensure that a compound will become a drug; it merely indicates drug-likeness and assists in eliminating weaker compounds during the preclinical phase. Our models were trained using the 659 data points that remained after the RO5 filter was applied. The spider plots of the compounds in the dataset that are either inside or outside RO5 domain are displayed in Fig. 2.

2.2 Model building and evaluation

Molecular fingerprints[21] encode structural data into numerical vectors or fixed-length bit-strings, which enable fast similarity comparisons crucial for virtual screening,[22] structure–activity relationship studies, and chemical space maps creation.[23] In our work, molecular fingerprints derived from SMILES were computed using RDKit[24] and used as inputs for machine learning models. The dataset was split into 332 active and 327 inactive compounds based on a pIC50 cut-off value of 6.2. We built twenty machine learning classifiers using Morgan3 fingerprints for quantitative structure–activity relationship (QSAR) classification,[25] selecting the top three classifiers based on various learning methods and evaluation metrics. The classifiers were built using Scikit-learn and Lightgbm.[26,27] The hyperparameters of the top three classifiers were fine-tuned and combined to form a voting classifier, enhancing overall performance and robustness compared to individual classifiers.[28] Similar approach was used for QSAR regression.

To assess classifiers, metrics including accuracy, precision, sensitivity, specificity, and AUC (Area Under Curve) were computed based on the confusion matrix.[29] Regressors were evaluated based on mean absolute error (MAE), root-mean-squared error (RMSE), and R^2 – score.[30]

2.3 Ligand and similarity based virtual screening

We employed the voting classifier for ligand-based virtual screening[31,32] of the eMolecules databases[33] to screen for active compounds, selecting molecules with a predicted probability

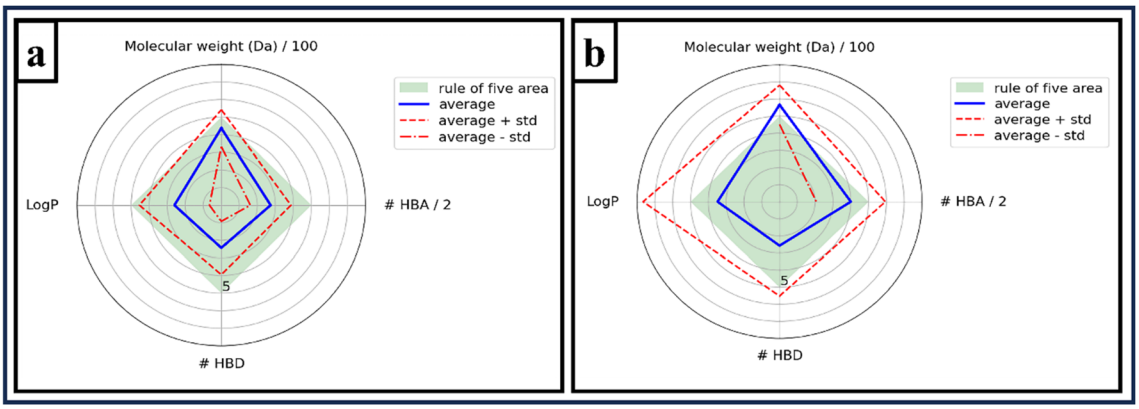

Fig. 2 Physio-chemical radar plots of the compounds in the dataset (a) inside RO5 domain or (b) outside RO5 domain.

exceeding 90% as potential active inhibitors. The database was filtered before screening to remove entries with invalid SMILES, Pan Assay Interference Molecules (PAINS), and those not meeting Lipinski's Rule of Five (RO5) criteria, using RDKit.

Similarity-based virtual screening measures the similarity between database structures and reference structures, based on the principle that similar structures likely have similar bioactivities.[34–36] For 3CLpro inhibitors, the chemical similarity between potential active inhibitors and known inhibitors was calculated using Molecular ACCess System (MACCS) and Morgan2 fingerprints, with Tanimoto and Dice similarity indices ensuring consistent comparisons.[37] Three potential compounds, consistently ranking in the top five during analysis, were selected for further assessment. Similarity maps of these candidates, created using Morgan2 fingerprints, visualized their similarity to known inhibitors.[38]

2.4 ADMET analysis of potential inhibitors

Assessing the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties is a crucial yet complex part of the drug discovery process, as these factors contribute to a significant portion of clinical failures.[39,40] In this study, we conducted a preliminary ADMET analysis using the SwissADME platform[41] for ADMET profiling and the ProTox-II tool[42] for toxicity predictions of candidate compounds relative to known inhibitors. Additionally, the maximum tolerated dose (MTD) for humans was estimated using the pkCSM tool.[43] Although these tools offer useful preliminary insights, their results are speculative and should be interpreted carefully.

2.5 Molecular docking

Molecular docking was used to determine how drugs attach to and interact with a protein. The crystal structure of 3CLpro (PDB ID: 5R82) complexed with an Z219104216 inhibitor was retrieved from the Protein Data Bank (PDB)[44,45] and refined using I-TASSER.[46] AutoDock4 was used to perform molecular docking.[47] To validate the docking parameters, the native ligand was redocked in the same binding pocket and the root mean square deviation (RMSD) between the initial pose and the docked pose was calculated using PyMOL.[48] Proteins and ligands were prepared using AutoDockTools by removing water molecules, adding Kollmann's charges, integrating polar hydrogens, and converting to protein data bank with partial charge and atom type (PDBQT) format. A cubic grid box (50 Å sides) centered at coordinates 10.364, 1.549, and 20.182 was used for site-specific docking. Docking parameters included a grid spacing of 0.375 Å, a population size of 300; 2 500 000 energy evaluations; and 100 docking runs using the Lamarckian Genetic Algorithm.[49–51] Protein–ligand interactions for the best-scored poses were analyzed with Protein–Ligand Interaction Profiler (PLIP).[52]

2.6 Molecular dynamics simulation

Molecular dynamics (MD) simulations were conducted for three candidate compounds to refine binding affinities, stability, and interactions. Using GROMACS[53] with the CHARMM36 force-field,[54] the highest-scoring protein–ligand complex from docking was simulated. Ligands were parameterized *via* SwissParam,[55] and the system was neutralized with 0.15 mol L^{-1} concentration of Cl^- and Na^+ ions,[56] solvated in a dodecahedron box of SPC water.[57] Energy minimization used 50 000 steps of the steepest descent method, followed by equilibration for 100 ps at 300 K in an NVT ensemble with a V-rescale thermostat.[58] Further equilibration for 100 ps at 1 bar and 300 K used an NPT ensemble with isotropic Berendsen pressure coupling. An unrestrained 200 ns MD simulation was then run with a 2 fs timestep, using a Parrinello-Rahman barostat and V-rescale thermostat.[59]

Stability was assessed by analyzing the root mean square deviation (RMSD), root mean square fluctuation (RMSF), protein solvent accessible surface area (SASA), radius of gyration (Rg), number of hydrogen bonds (H-bonds), and Dictionary of Secondary Structure in Proteins (DSSP).[60] Ligand–protein binding free energies were calculated using gmx_MMPBSA and gmx_MMPBSA_ana, following the MM-PBSA approach, over the final 20 ns of equilibrated trajectories.[61]

3 Results and discussions

3.1 Model building and database screening

The performance of 20 classifiers (CLF) is summarized in ESI Table S1.† We selected Nu-Support Vector Classifier (NuSVC), ExtraTreesClassifier (ET), and Light Gradient Boosting Machine (LGBM) Classifier to construct a Voting Classifier (VC). For NuSVC, parameters were set to nu = '0.2', kernel = 'rbf', and

Table 1 Evaluation of three individual classifiers and voting classifier

Classifier	Accuracy	Precision	Sensitivity	Specificity	AUC
NuSVC	0.89	0.94	0.86	0.93	0.96
ET	0.90	0.96	0.86	0.95	0.95
LGBM	0.88	0.91	0.86	0.90	0.95
VC	0.88	0.91	0.86	0.90	0.96

Table 2 Five-fold cross validation of individual classifiers and voting classifier

Classifier	Accuracy	Precision	Sensitivity	Specificity	AUC
NuSVC	0.87 (±0.04)	0.87 (±0.06)	0.88 (±0.03)	0.87 (±0.06)	0.94 (±0.01)
ET	0.88 (±0.04)	0.89 (±0.07)	0.87 (±0.04)	0.89 (±0.06)	0.94 (±0.03)
LGBM	0.85 (±0.04)	0.84 (±0.06)	0.87 (±0.04)	0.83 (±0.06)	0.92 (±0.03)
VC	0.87 (±0.04)	0.86 (±0.06)	0.87 (±0.05)	0.86 (±0.05)	0.94 (±0.03)

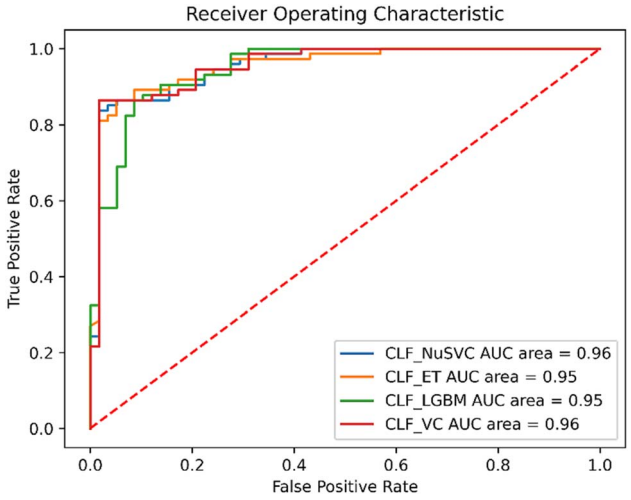

Fig. 3 ROC curve of the individual classifiers and voting classifier.

gamma = 'scale'; for ET, n_estimators = '1000', criterion = 'gini', and max_features = 'sqrt'; for LGBM, n_estimators = '200', learning_rate = '0.2', max_depth = '4', and num_leaves = '50'; all other parameters were left at their default values. The VC employed a 'soft' voting mechanism. The confusion matrices of the individual classifiers and the VC are presented in ESI Fig. S2,† with their evaluations detailed in Table 1.

Table 2 presents the five-fold cross-validation results for individual classifiers and the voting classifier using a 20% random data selection.

Fig. 3 illustrates the ROC curves for these classifiers. With AUC scores of 0.96, 0.95, 0.95, and 0.96, all classifiers demonstrated strong classification performance. The voting classifier was chosen for screening the eMolecules database due to its superior robustness, identifying 39 molecules with prediction probabilities above 90% as potential active inhibitors.

3.2 Similarity measures analysis

We assessed the chemical similarity between 39 potential active inhibitors and known inhibitors of 3CLpro-azvudine, ensitrelvir, nirmatrelvir, and simnotrelvir using MACCS and Morgan2 fingerprints. Tanimoto and Dice similarity metrics were computed for both MACCS and Morgan2 fingerprints. Table 3 displays the top five compounds with the highest similarity to each reference, considering Tanimoto and Dice similarities for both MACCS and Morgan fingerprints.

For further analysis, we selected three structures-M1 (PubChem CID 56879830), M2 (PubChem CID 70722105), and M3 (PubChem CID 72893585)-based on their higher frequency of occurrence and higher similarity indexes among the similar compounds.

Fig. 4 illustrates the similarity map generated for these three compounds with four known inhibitors using the Morgan2 fingerprint.

3.3 ADMET analysis and drug-likeness

ESI Table S2† presents the physicochemical properties, pharmacokinetics, and drug-likeness of the molecules. ADMET analysis with reference to oral 3CLpro inhibitors azvudine and nirmatrelvir in phase 4 trials,[62–64] shows all candidate compounds meet Lipinski's rule of five, suggesting favorable drug-likeness with good absorption and permeability.[65] Solubility analysis indicates that M2 and nirmatrelvir are soluble, while azvudine, M1, and M3 are highly soluble.

ESI Fig. S3† presents the bioavailability radar diagram comparing candidate compounds with reference molecules across various physicochemical properties: lipophilicity, size, polarity, solubility, flexibility, and saturation. The pink region indicates the ideal drug-likeness zone, while the red hexagon represents drug-likeness profile of molecules. A bioavailability

Table 3 Similarity checking between azvudine, ensitrelvir, nirmatrelvir and simnotrelvir with top-ranked molecules using Morgan2 and MACCS fingerprints

Known inhibitors	Top 5 molecules	Tanimoto_morgan	Dice_morgan	Top 5 molecules	Tanimoto_maccs	Dice_maccs
Azvudine	5	0.147287	0.256757	37	0.578313	0.732824
	26	0.144330	0.252252	34	0.556818	0.715328
	15	0.136364	0.240000	35	0.547619	0.707692
	19	0.135922	0.239316	31	0.534884	0.696970
	14	0.134615	0.237288	38	0.530120	0.692913
Ensitrelvir	20	0.186667	0.314607	30	0.653846	0.790698
	16	0.174497	0.297143	37	0.636364	0.777778
	11	0.169935	0.290503	35	0.623377	0.768000
	27	0.168919	0.289017	13	0.623377	0.768000
	8	0.166667	0.285714	24	0.618421	0.764228
Simnotrelvir	34	0.154930	0.268293	4	0.535211	0.697248
	14	0.133803	0.236025	17	0.532468	0.694915
	5	0.130178	0.230366	14	0.531646	0.694215
	17	0.120805	0.215569	34	0.524390	0.688000
	38	0.118881	0.212500	3	0.520548	0.684685
Nirmatrelvir	5	0.148148	0.258065	34	0.587500	0.740157
	34	0.135714	0.238994	30	0.550000	0.709677
	38	0.123188	0.219355	31	0.544304	0.704918
	31	0.117241	0.209877	33	0.525641	0.689076
	30	0.116438	0.208589	38	0.519481	0.683761

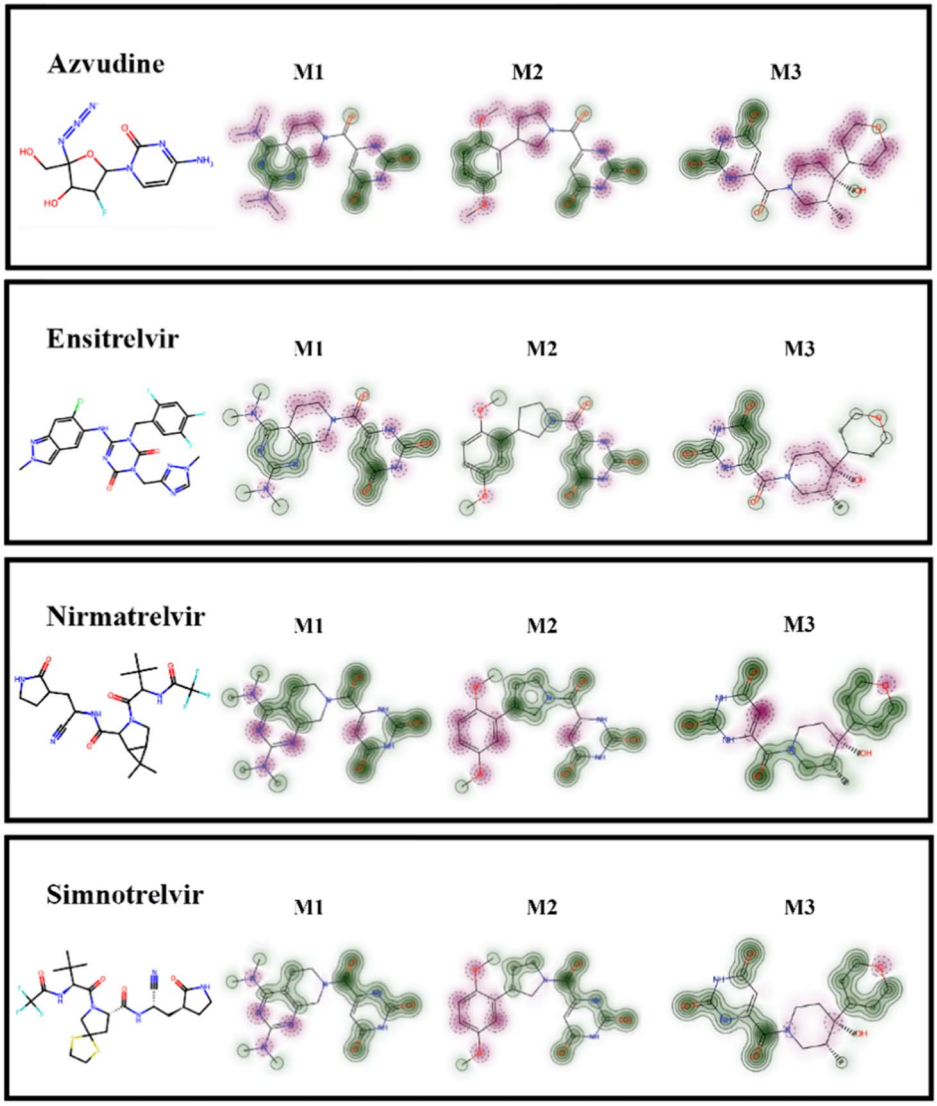

Fig. 4 Similarity maps between azvudine, ensitrelvir, nirmatrelvir, and simnotrelvir as references and candidates M1, M2, and M3 using Morgan2 fingerprint. Coloring method: green: positive difference, gray: no change in similarity, and pink: negative difference.

score of 0.55 suggests favorable pharmacokinetic characteristics. The log Kp values of the candidate compounds suggest good skin permeability, falling within the range of −9.7 to −3.5.[39]

The Brain Or IntestinaL EstimateD permeation method (BOILED-Egg) was used to predict molecular permeability, estimating the potential for passive human gastrointestinal absorption (HIA) and blood–brain barrier (BBB) penetration.[66] ESI Fig. S4† presents a boiled-egg graph comparing known inhibitors with potential inhibitors.

The yolk portion represents the physicochemical space indicating molecules most likely to penetrate the brain, while the white part denotes molecules with a high probability of gastrointestinal (GI) absorption. Molecules predicted to have low human gastrointestinal absorption (HIA) and blood–brain barrier (BBB) penetration are depicted in the gray zone. Blue

Table 4 Oral toxicity assessment of the candidate compounds with azvudine as reference drug

Chemical compound	Predicted LD50 (mg kg^{-1})	Predicted toxicity class	Prediction accuracy (%)	Average similarity (%)
Azvudine	1000	4	67.38	59.91
M1	550	4	54.26	46.19
M2	500	4	54.26	48.36
M3	200	3	67.38	55.68

points indicate molecules that are substrates of P-glycoprotein (P-gp) and actively effluxed, while red points represent non-substrates. Fig. S4† shows that azvudine, M2 and M3 are not P-gp substrates, whereas M1 and nirmatrelvir are P-gp substrates.[67] All candidate compounds and known inhibitors, except for azvudine are predicted to exhibit favorable absorption characteristics and are not expected to penetrate the BBB.

Table 4 displays the oral toxicity assessment of the candidate compounds using azvudine as the reference drug.

Compared to azvudine, all candidates showed lower LD50 values,[68] suggesting potentially higher toxicity. M1, M2, and azvudine are predicted to be in class IV, while M3 may fall into class III based on toxicity classification criteria. Additionally, The MTDs of azvudine, M1, M2, and M3 are predicted as 0.289, 0.614, 0.615, and 0.542 (log mg per kg per day), respectively.

3.4 Prediction of pIC50 values

The performance of 20 regressors (RG) is summarized in ESI Table S3.† To construct the Voting Regressor (VR), we chose the Random Forest (RF), Hist Gradient Boosting (HGB), and Light Gradient Boosting Machine (LGBM) regressors. The RF regressor had n_estimators set to '200', criterion set to 'squared_error', max_features set to 'sqrt' and min_samples_split to '2'; the HGB regressor had max_iter set to '200' and learning_rate to '0.1'; the LGBM regressor had n_estimators set to '200' and learning_rate to '0.1'; all other parameters were left at their default values. These three regressors were combined to build the voting regressor. Evaluation metrics for the voting regressor in training, testing, and 5-fold CV are presented in Table 5.

Experimental and predicted pIC50 values are compared in ESI Fig. S5.† Predicted pIC50 values for M1, M2, and M3 were 7.35, 7.59, and 7.71 which are comparable to activity of nirmatrelvir (7.70).

3.5 Molecular docking analysis

Molecular docking was used to generate the 3CLpro protein–ligand complexes of candidate compounds. The RMSD between the initial pose and the re-docked pose of the native ligand was found to be 0.426 Å (Fig. S6†). These validated parameters were used for the docking of 3CLpro and the candidate compounds. The protein–ligand interactions of candidate compounds are shown in Fig. 5.

The binding energy (with each contributing factor) of candidate compounds with 3CLpro for best docking pose is shown in ESI Table S4.† The binding energies of M1 (6-[2,4-bis(dimethylamino)-6,8-dihydro-5*H*-pyrido[3,4-d]pyrimidine-7-carbonyl]-1*H* pyrimidine-2,4-dione), M2 (6-[3-(2,5-dimethoxyphenyl)pyrrolidine-1-carbonyl]-1*H*-pyrimidine-2,4-dione) and M3 ([(3*R*,4*R*)-4-hydroxy-3-methyl-4-(oxan-4-yl)piperidine-1-carbonyl]-1*H*-pyrimidine-2,4-dione) are −8.64 kcal mol^{-1}, −8.22 kcal mol^{-1} and −8.00 kcal mol^{-1} respectively which suggests a good binding affinity with target protein.

Table 5 Evaluation metrics of voting regressor in training, testing and 5-fold CV

Statistical metrics	Training	Testing	5-fold CV
R^2	0.97	0.71	0.73
MAE	0.13	0.45	0.41
RMSE	0.18	0.62	0.57

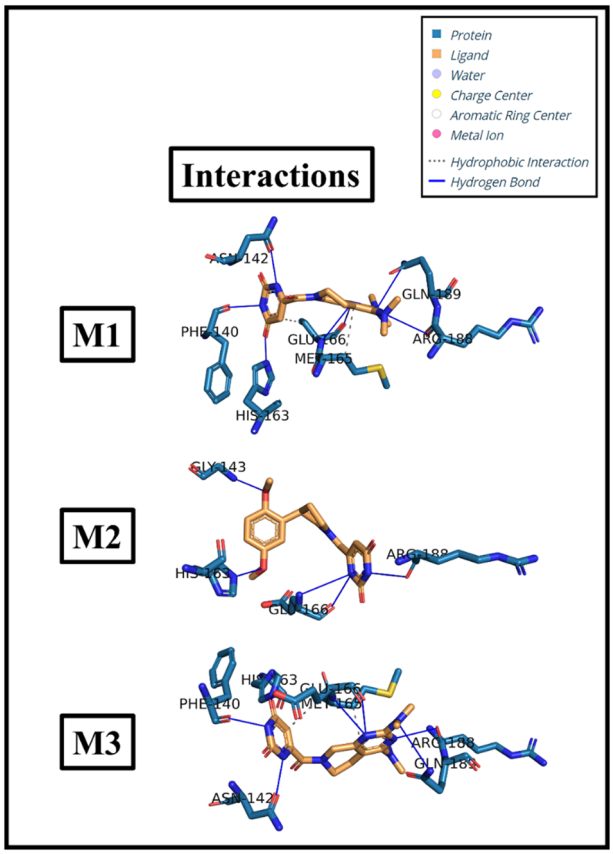

Fig. 5 Protein–ligand interactions of 3CLpro and candidate compounds.

The interactions between the active residues of 3CLpro and the best docked pose of candidate compounds are shown in ESI Table S5.†

3.6 Molecular dynamics simulation analysis

We performed MD simulations for the complexes of the three candidate compounds and target protein to verify the outcomes of our virtual screening using machine learning and docking. Through trajectory analysis, only M1 was found to be stable during MD simulation among the three candidate compounds. From RMSD data we found that all our systems reached stability after 180 ns (Fig. 6a), so we defined the productive phase of our simulations as the time between 180 and 200 ns for all the runs. The RMSD, Rg and, RMSF plots of MD simulation for apo and M1-complex are shown in Fig. 6.

The stability of the ligand and protein in a complex was studied using RMSD analysis. The average RMSD of protein backbone in apo and M1 binding forms is 2.02 ± 0.21 Å and 1.91 ± 0.31 Å, respectively. The RMSD value of the protein

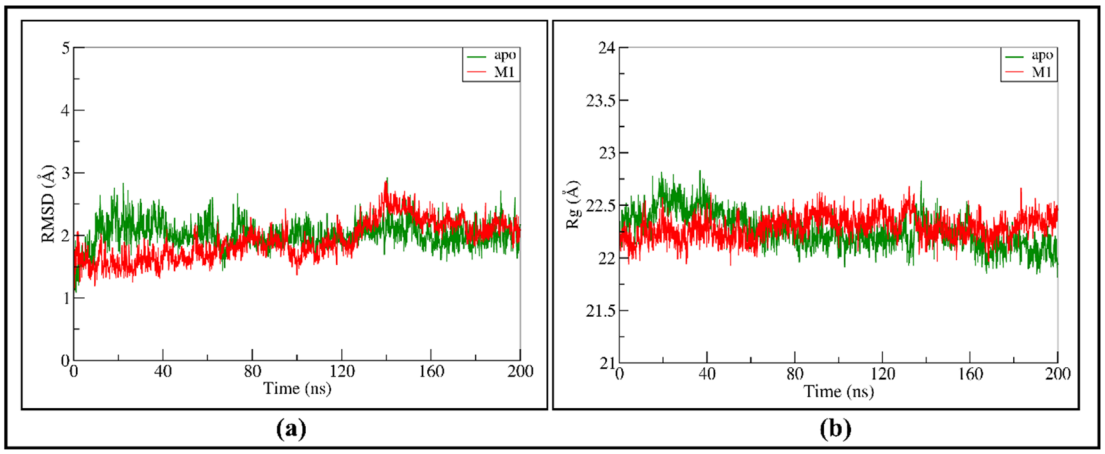

Fig. 6 (a) RMSD and (b) Rg plots for apo and M1 binding forms of 3CLpro during 200 ns MD simulation.

backbone was less than 3 Å, indicating a minor change for globular proteins. These results demonstrate the stability of apo and ligand binding forms.

Next, we examined the Rg, which is a reliable indicator of protein folding. The average value of Rg throughout the simulation for apo and M1 binding forms is 22.26 ± 0.17 Å and 21.29 ± 0.12 Å respectively, which shows the overall stable protein folding in the complex without any significant expansion or condensation.

The average RMSF values for apo and ligand binding forms are 1.21 ± 0.59 Å and 1.09 ± 0.61 Å respectively, with the majority of residues showing similar RMSF values, while some regions – like SER1 (6.51 Å), GLY2 (4.38 Å), SER301 (3.03 Å), THR304 (3.27 Å), and GLN306 (3.07 Å) – showed larger fluctuations (Fig. S7†). These residues are not critical because they are found in the inactive regions of protein. On the other hand, key residues in the active site, like HIS41, SER144, CYS145, GLU166, and HIS172, showed reduced fluctuations with RMSF values below 1.1 Å, indicating that the formed hydrogen bonds stabilize the ligand complexation with protein 3CLpro.

Furthermore, we used the DSSP module installed in GROMACS to examine the stability of their secondary structure.[69,70] During our simulation, the M1 and apo binding forms both kept a stable secondary structure on a global scale (Fig. S8†).

The GROMACS Hbond module[71] with default parameters and the HbMap2Grace program[72] were utilized to assess the hydrogen bond pattern, while the SurfinMD program[73] was employed to evaluate the molecular surface area. The hydrogen bond data indicates that there were notable interactions between the M1 and the active residues (Fig. 7). M1 displayed hydrogen bonding with the SER144 complex for nearly the whole simulation period.

Additionally, we calculated the atomic contacts between M1 and SARS-CoV-2 Mpro (Fig. 8). The contact surface area disclosed interactions with key residues in the active site.

By using MM-PBSA calculations, the post-MD free energy of M1 in complex with 3CLpro has been examined. Van der Waals energy (VDWAALS), electrostatic energy (EEL), polar solvation energy (EPB), and nonpolar solvation energy (ENPOLAR) are the main contributors to the total binding free energy. Fig. 9a shows

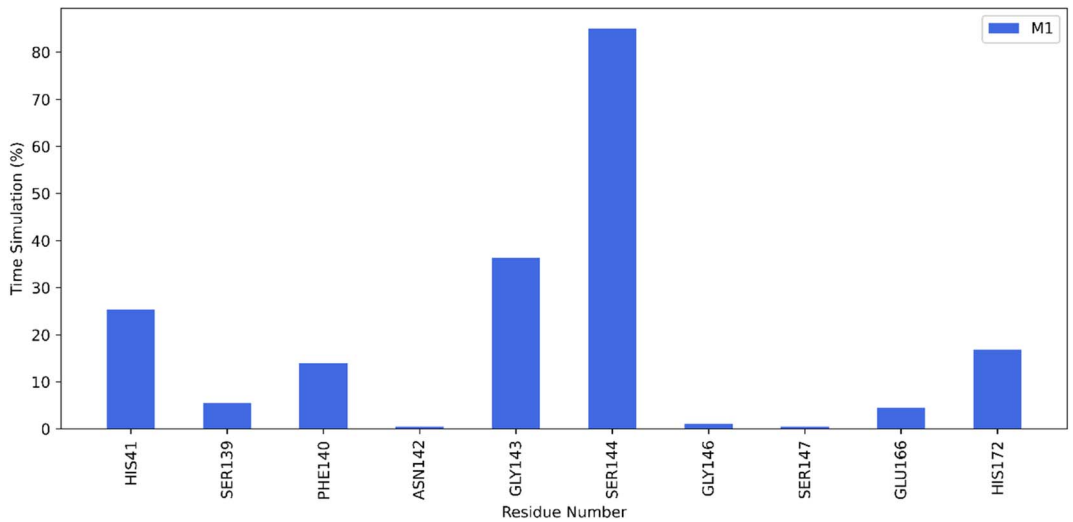

Fig. 7 Hydrogen bond stability in 3CLpro-M1 complex for the productive phase.

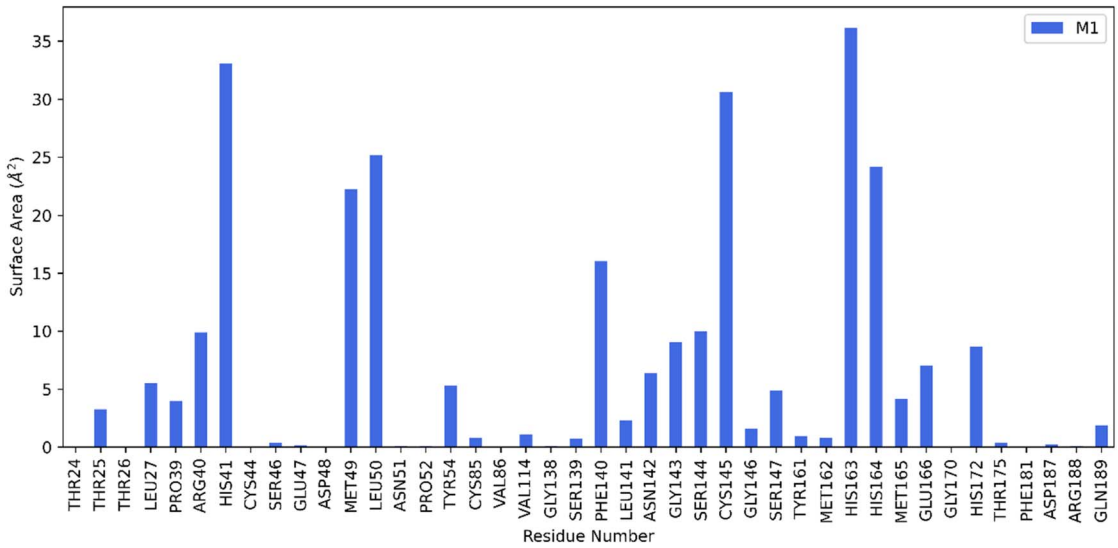

Fig. 8 Surface molecular area of 3CLpro-M1 complex for the productive phase.

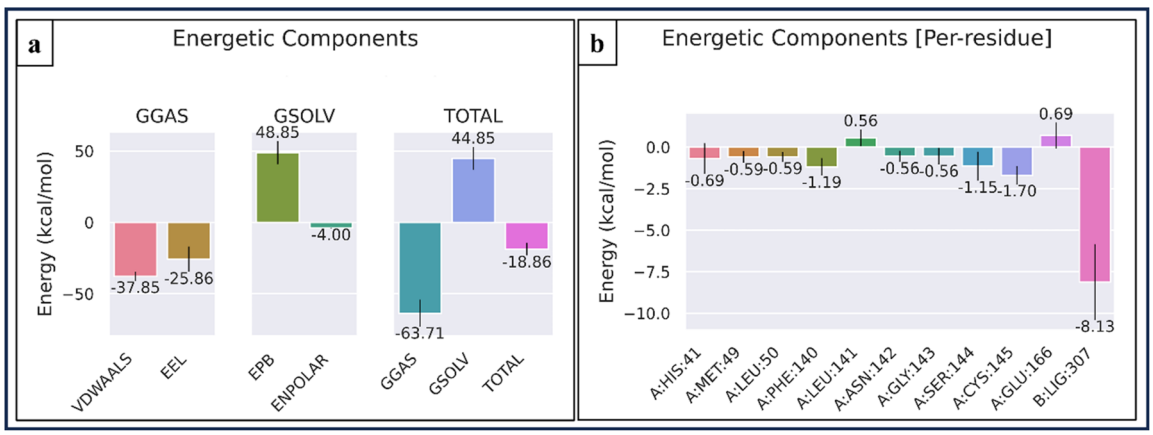

Fig. 9 MM-PBSA results of M1 in complex with 3CLpro during last 20 ns MD simulations: (a) binding free energy contribution by different interactions, (b) binding free energy contributions by active residues and ligand.

the overall binding free energy contributors of M1 in complex with 3CLpro over the last 200 frames. The major contributors to the total MM-PBSA free energy of -18.86 ± 4.38 kcal mol^{-1}, expressed as average $\pm$ SD, are electrostatic energy (-25.86 ± 8.88 kcal mol^{-1}) and vdW energy (-37.85 ± 3.24 kcal mol^{-1}), as shown in ESI Table S6.†

Fig. 9b displays the binding free energies that are contributed by the active residues of 3CLpro and M1. The decomposition analysis indicated that M1 has a strong binding affinity. The ligand engages with critical residues in 3CLpro, notably forming a significant interaction with the CYS145–HIS41 catalytic dyad, which is essential for the enzyme's functionality.[74] Of the total MM-PBSA free energy, M1 contributes -8.13 ± 2.28 kcal mol^{-1}. The lowest binding free energies of -1.70 ± 0.54 kcal mol^{-1} and -1.19 ± 0.52 kcal mol^{-1} are displayed by CYS145 and PHE140 respectively, out of all the residues (ESI Table S7†). Additionally, ESI Fig. S9† displays the heatmap of the binding free energy contribution by active residues and ligand.

4 Conclusion

Drug development is costly and time-consuming. We utilized a workflow integrating ligand-based virtual screening with similarity assessments of approved drugs to identify potential 3CLpro inhibitors. Using three machine learning classifiers, we created a voting classifier to predict activity probabilities, analyzing approximately 10 million molecules. We selected three compounds M1, M2 and M3 for further investigation. ADMET analysis, with azvudine and nirmatrelvir as references, and 200 ns MD simulations identified M1 (6-[2,4-bis(dimethylamino)-6,8-dihydro-5*H*-pyrido[3,4-d]pyrimidine-7-carbonyl]-1*H*-pyrimidine-2,4-dione) as stable. Predicted LD50 values for M1 and azvudine were 550 and 1000 mg kg^{-1}, respectively. The pIC50 value for M1 was approximately 7.35, similar to

nirmatrelvir. MM-PBSA calculations showed a binding energy of -18.86 ± 4.38 kcal mol^{-1} for the M1-3CLpro complex. Our study suggests that M1 warrants further investigation as a potential SARS-CoV-2 therapeutic, potentially improving drug discovery efficiency and conserving resources.

Data availability

The data supporting the findings of this study are available within the article and its ESI.†

Author contributions

Sandeep Poudel Chhetri: experiment design, data generation, analyzed data, and drafted the manuscript. Vishal Singh Bhandari: technical support and revised the manuscript. Rajesh Maharjan: technical support, data generation and revised the manuscript. Tika Ram Lamichhane: critical feedback, graphical and statistical analysis, and revised the manuscript.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1 V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, D. G. Mulders, B. L. Haagmans, B. Van Der Veer, S. Van Den Brink, L. Wijsman, G. Goderski, J.-L. Romette, J. Ellis, M. Zambon, M. Peiris, H. Goossens, C. Reusken, M. P. Koopmans and C. Drosten, *Eurosurveillance*, 2020, **25**, DOI: 10.2807/1560-7917.ES.2020.25.3.2000045.
2 *COVID-19 cases | WHO COVID-19 dashboard*, https://data.who.int/dashboards/covid19/cases, (accessed January 21, 2024).
3 B. Hu, H. Guo, P. Zhou and Z.-L. Shi, *Nat. Rev. Microbiol.*, 2021, **19**, 141–154.
4 A. A. Agbowuro, W. M. Huston, A. B. Gamble and J. D. A. Tyndall, *Med. Res. Rev.*, 2018, **38**, 1295–1331.
5 K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters and R. Hilgenfeld, *Science*, 2003, **300**, 1763–1767.
6 Y. Unoh, S. Uehara, K. Nakahara, H. Nobori, Y. Yamatsu, S. Yamamoto, Y. Maruyama, Y. Taoda, K. Kasamatsu, T. Suto, K. Kouki, A. Nakahashi, S. Kawashima, T. Sanaki, S. Toba, K. Uemura, T. Mizutare, S. Ando, M. Sasaki, Y. Orba, H. Sawa, A. Sato, T. Sato, T. Kato and Y. Tachibana, *J. Med. Chem.*, 2022, **65**, 6499–6512.
7 S. Ullrich and C. Nitsche, *Bioorg. Med. Chem. Lett.*, 2020, **30**, 127377.
8 L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox and R. Hilgenfeld, *Science*, 2020, **368**, 409–412.
9 G. Li, R. Hilgenfeld, R. Whitley and E. De Clercq, *Nat. Rev. Drug Discovery*, 2023, **22**, 449–475.
10 A. Lavecchia and C. Giovanni, *CMC*, 2013, **20**, 2839–2860.
11 D. E. Gloriam, *Nature*, 2019, **566**, 193–194.
12 F. Zhong, J. Xing, X. Li, X. Liu, Z. Fu, Z. Xiong, D. Lu, X. Wu, J. Zhao, X. Tan, F. Li, X. Luo, Z. Li, K. Chen, M. Zheng and H. Jiang, *Sci. China: Life Sci.*, 2018, **61**, 1191–1204.
13 Y. Duan, J. S. Edwards and Y. K. Dwivedi, *J. Inf. Manag.*, 2019, **48**, 63–71.
14 A. Lavecchia, *Drug Discovery Today*, 2019, **24**, 2017–2032.
15 A. Salimi, J. H. Lim, J. H. Jang and J. Y. Lee, *Sci. Rep.*, 2022, **12**, 18825.
16 R. Maharjan, K. Gyawali, A. Acharya, M. Khanal, M. P. Ghimire and T. R. Lamichhane, *Mol. Simul.*, 2024, **50**, 717–728.
17 D. Sydow, A. Morger, M. Driller and A. Volkamer, *J. Cheminf.*, 2019, **11**, 29.
18 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 and A. R. Leach, *Nucleic Acids Res.*, 2019, **47**, D930–D940.
19 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2012, **64**, 4–17.
20 B. C. Doak, B. Over, F. Giordanetto and J. Kihlberg, *Chem. Biol.*, 2014, **21**, 1115–1142.
21 D. Bajusz, A. Rácz and K. Héberger, in *Comprehensive Medicinal Chemistry III*, Elsevier, 2017, pp. 329–378.
22 P. Willett, *Drug Discovery Today*, 2006, **11**, 1046–1053.
23 M. Awale, R. Visini, D. Probst, J. Arús-Pous and J.-L. Reymond, *CHIMIA*, 2017, **71**, 661.
24 G. Landrum, *Rdkit: Open-source cheminformatics software. (version 2023.9.4)*, 2016.
25 S. Kwon, H. Bae, J. Jo and S. Yoon, *BMC Bioinf.*, 2019, **20**, 521.
26 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 and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
27 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, vol. 30.
28 T. G. Dietterich, in *Multiple Classifier Systems*, Springer, Berlin, Heidelberg, 2000, pp. 1–15.
29 A. Luque, A. Carrasco, A. Martín and A. de las Heras, *Pattern Recognit.*, 2019, **91**, 216–231.
30 D. Chicco, M. J. Warrens and G. Jurman, *PeerJ Comput. Sci.*, 2021, **7**, e623.
31 M. T. J. Quimque, K. I. R. Notarte, R. A. T. Fernandez, M. A. O. Mendoza, R. A. D. Liman, J. A. K. Lim, L. A. E. Pilapil, J. K. H. Ong, A. M. Pastrana, A. Khan, D.-Q. Wei and A. P. G. Macabeo, *J. Biomol. Struct. Dyn.*, 2021, **39**, 4316–4333.
32 A. N. Lima, E. A. Philot, G. H. G. Trossini, L. P. B. Scott, V. G. Maltarollo and K. M. Honorio, *Expert Opin. Drug Discovery*, 2016, **11**, 225–239.
33 *eMolecules*, https://search.emolecules.com/, (accessed January 4, 2024).

34 R. P. Sheridan and S. K. Kearsley, *Drug Discovery Today*, 2002, **7**, 903–911.
35 A. G. Maldonado, J. P. Doucet, M. Petitjean and B.-T. Fan, *Mol. Divers.*, 2006, **10**, 39–79.
36 G. Cheng, M. Lajiness and M. A. Johnson, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 909–915.
37 D. Bajusz, A. Rácz and K. Héberger, *J. Cheminf.*, 2015, **7**, 20.
38 S. Riniker and G. A. Landrum, *J. Cheminf.*, 2013, **5**, 43.
39 J. Bojarska, M. Remko, M. Breza, I. D. Madura, K. Kaczmarek, J. Zabrocki and W. M. Wolf, *Molecules*, 2020, **25**, 1135.
40 I. Kola and J. Landis, *Nat. Rev. Drug Discovery*, 2004, **3**, 711–716.
41 A. Daina, O. Michielin and V. Zoete, *Sci. Rep.*, 2017, **7**, 42717.
42 P. Banerjee, A. O. Eckert, A. K. Schrey and R. Preissner, *Nucleic Acids Res.*, 2018, **46**, W257–W263.
43 D. E. V. Pires, T. L. Blundell and D. B. Ascher, *J. Med. Chem.*, 2015, **58**, 4066–4072.
44 P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman and S. K. Burley, *Nucleic Acids Res.*, 2017, **45**, D271–D281.
45 A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C. D. Owen, E. Resnick, C. Strain-Damerell, A. Aimon, P. Ábrányi-Balogh, J. Brandão-Neto, A. Carbery, G. Davison, A. Dias, T. D. Downes, L. Dunnett, M. Fairhead, J. D. Firth, S. P. Jones, A. Keeley, G. M. Keserü, H. F. Klein, M. P. Martin, M. E. M. Noble, P. O'Brien, A. Powell, R. N. Reddi, R. Skyner, M. Snee, M. J. Waring, C. Wild, N. London, F. von Delft and M. A. Walsh, *Nat. Commun.*, 2020, **11**, 5047.
46 J. Yang and Y. Zhang, *Nucleic Acids Res.*, 2015, **43**, W174–W181.
47 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785–2791.
48 W. L. DeLano, *CCP4 Newslett. Protein Cryst.*, 2002, **40**, 82.
49 G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *J. Comput. Chem.*, 1998, **19**, 1639–1662.
50 M. Khanal, A. Acharya, R. Maharjan, K. Gyawali, R. Adhikari, D. D. Mulmi, T. R. Lamichhane and H. P. Lamichhane, *PLoS One*, 2024, **19**, e0307501.
51 A. Acharya, M. Khanal, R. Maharjan, K. Gyawali, B. R. Luitel, R. Adhikari, D. D. Mulmi, T. R. Lamichhane, H. P. Lamichhane, A. Acharya, M. Khanal, R. Maharjan, K. Gyawali, B. R. Luitel, R. Adhikari, D. D. Mulmi, T. R. Lamichhane and H. P. Lamichhane, *AIMSBPOA*, 2024, **11**, 142–165.
52 M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt and M. Schroeder, *Nucleic Acids Res.*, 2021, **49**, W530–W534.
53 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.
54 J. Huang and A. D. MacKerell Jr, *J. Comput. Chem.*, 2013, **34**, 2135–2145.
55 V. Zoete, M. A. Cuendet, A. Grosdidier and O. Michielin, *J. Comput. Chem.*, 2011, **32**, 2359–2368.
56 T. R. Lamichhane and M. P. Ghimire, *Heliyon*, 2021, **7**, e08220.
57 P. Mark and L. Nilsson, *J. Phys. Chem. A*, 2001, **105**, 9954–9960.
58 G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.
59 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
60 R. C. Silva, H. F. Freitas, J. M. Campos, N. M. Kimani, C. H. T. P. Silva, R. S. Borges, S. S. R. Pita and C. B. R. Santos, *Int. J. Mol. Sci.*, 2021, **22**, 11739.
61 M. S. Valdés-Tresanco, M. E. Valdés-Tresanco, P. A. Valiente and E. Moreno, *J. Chem. Theory Comput.*, 2021, **17**, 6281–6291.
62 B. Yu and J. Chang, *Sig. Transduct. Target Ther.*, 2020, **5**, 1–2.
63 D. R. Owen, C. M. N. Allerton, A. S. Anderson, L. Aschenbrenner, M. Avery, S. Berritt, B. Boras, R. D. Cardin, A. Carlo, K. J. Coffman, A. Dantonio, L. Di, H. Eng, R. Ferre, K. S. Gajiwala, S. A. Gibson, S. E. Greasley, B. L. Hurst, E. P. Kadar, A. S. Kalgutkar, J. C. Lee, J. Lee, W. Liu, S. W. Mason, S. Noell, J. J. Novak, R. S. Obach, K. Ogilvie, N. C. Patel, M. Pettersson, D. K. Rai, M. R. Reese, M. F. Sammons, J. G. Sathish, R. S. P. Singh, C. M. Steppan, A. E. Stewart, J. B. Tuttle, L. Updyke, P. R. Verhoest, L. Wei, Q. Yang and Y. Zhu, *Science*, 2021, **374**, 1586–1593.
64 *Study Details | A Study of Efficacy and Safety of Azvudine vs. Nirmatrelvir-Ritonavir in the Treatment of COVID-19 Infection | ClinicalTrials.gov*, **https://clinicaltrials.gov/study/NCT05697055**, (accessed March 3, 2024).
65 J. S. Delaney, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1000–1005.
66 A. Daina and V. Zoete, *ChemMedChem*, 2016, **11**, 1117–1121.
67 C. Chen, M.-H. Lee, C.-F. Weng and M. K. Leong, *Molecules*, 2018, **23**, 1820.
68 P. O. Lohohola, B. M. Mbala, S.-M. N. Bambi, D. T. Mawete, A. Matondo and J. G. M. Mvondo, *Int. J. Trop. Dis. Health*, 2021, **42**, 1–12.
69 W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten and G. Vriend, *Nucleic Acids Res.*, 2015, **43**, D364–D368.
70 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.
71 D. van der Spoel, P. J. van Maaren, P. Larsson and N. Tîmneanu, *J. Phys. Chem. B*, 2006, **110**, 4393–4398.
72 D. E. B. Gomes, A. W. Silva, R. D. Linis, P. G. Pascutti and T. A. Soares, *HbMap2Grace*, **https://lmdm.biof.ufrj.br/software/hbmap2grace/index.html-2002**.
73 D. E. B. Gomes, G. L. S. C. Sousa, A. W. S. D. Silva and P. G. Pascutti, *SurfinMD*, **https://lmdm.biof.ufrj.br/software/surfinmd/index.html-2012**.
74 J. C. Ferreira, S. Fadl, A. J. Villanueva and W. M. Rabeh, *Front. Chem.*, 2021, **9**, 692168.