

RESEARCH

Open Access



# SMILES-based QSAR virtual screening to identify potential therapeutics for COVID-19 by targeting 3CL<sup>pro</sup> and RdRp viral proteins

Faezeh Bazzi-Allahri<sup>1</sup>, Fereshteh Shiri<sup>1\*</sup>, Shahin Ahmadi<sup>2</sup>, Alla P. Toropova<sup>3</sup> and Andrey A. Toropov<sup>3</sup>

## Abstract

The COVID-19 pandemic has prompted the medical systems of many countries to develop effective treatments to combat the high rate of infection and death caused by the disease. Within the array of proteins found in SARS-CoV-2, the 3 chymotrypsin-like protease (3CL<sup>pro</sup>) holds significance as it plays a crucial role in cleaving polyprotein peptides into distinct functional nonstructural proteins. Meanwhile, RNA-dependent RNA polymerase (RdRp) takes center stage as the key enzyme tasked with replicating the viral genomic RNA within host cells. These proteins, 3CL<sup>pro</sup> and RdRp, are deemed optimal subjects for QSAR modeling due to their pivotal functions in the viral lifecycle. In this study, SMILES-based QSAR classification models were developed for a dataset of 2377 compounds that were defined as either active or inactive against 3CL<sup>pro</sup> and RdRp. Pharmacophore (PH4) and QSAR modeling were used for the virtual screening on 60.2 million compounds including ZINC, ChEMBL, Molport, and MCULE databases to identify new potent inhibitors against 3CL<sup>pro</sup> and RdRp. Then, a filter was established based on typical molecular characteristics to identify drug-like molecules. The molecular docking was also performed to evaluate the binding affinity of 156 AND 51 potential inhibitors to 3CL<sup>pro</sup> and RdRp, respectively. Among the 15 hits identified based on molecular docking scores, M3, N2, and N4 were identified as promising inhibitors due to their good synthetic accessibility scores (3.07, 3.11, and 3.29 out of 10 for M3, N2, and N4 respectively). These compounds contain amine functional groups, which are known for their crucial role in the binding interactions between drugs and their targets. Consequently, these hits have been chosen for further biological assay studies to validate their activity. They may represent novel 3CL<sup>pro</sup> and RdRp inhibitors possessing drug-like properties suitable for COVID-19 therapy.

**Keywords** SMILES-based QSAR classification, 3CL<sup>pro</sup> and RdRp, Molecular docking, Virtual screening, COVID-19

## Introduction

COVID-19, caused by the SARS-CoV-2 virus, has affected over 768 million people worldwide and resulted in more than 6 million deaths [1]. The pandemic has

highlighted the urgent need for effective therapeutics to combat emerging viral diseases, given its devastating impact on global health and the economy. While the rapid development of vaccines mitigated the severity of the outbreak, a continued demand persists for antiviral treatments to manage infections, reduce transmission, and address new viral variants. As a result, drug discovery remains a central focus of ongoing research efforts aimed at combating both current and future viral outbreaks.

One promising approach in drug discovery is targeting specific viral proteins essential for viral replication and

\*Correspondence:

Fereshteh Shiri

fereshteh.shiri@gmail.com; Fereshteh.shiri@uoaz.ac.ir

<sup>1</sup> Department of Chemistry, University of Zabol, Zabol, Iran

<sup>2</sup> Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

<sup>3</sup> Istituto Di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milan, Italy



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

conserved across different strains of coronaviruses [2]. Notable targets include the 3-chymotrypsin-like protease (3CL<sup>Pro</sup>) and RNA-dependent RNA polymerase (RdRp), both of which play crucial roles in the replication cycle of SARS-CoV-2 [3, 4]. The 3-chymotrypsin-like protease (3CLpro) is a crucial enzyme in the SARS-CoV-2 life cycle, responsible for cleaving the viral polyprotein into functional components essential for replication. Its unique catalytic Cys-His dyad and conserved active site, capable of accommodating multiple substrates, make it an ideal target for antiviral drug development [5]. Similarly, the RNA-dependent RNA polymerase (RdRp), another key viral enzyme, facilitates viral RNA synthesis and is highly conserved across coronaviruses, making it a pivotal target for therapeutics [6]. Together, 3CLpro and RdRp play essential roles in viral replication, positioning them as prime targets for developing COVID-19 treatments. Targeting these proteins has shown promise in identifying therapeutic candidates, but conventional drug discovery methods face significant limitations, including high costs, time inefficiency, and low success rates [7, 8]. Consequently, more rapid and cost-effective approaches are needed to identify potential antiviral compounds while ensuring their efficacy and safety [2].

Several anti-RNA polymerase drugs currently available, such as Ribavirin [9], Galidesivir [10], Remdesivir [11], and Tenofovir [12], have been approved for treating various viral infections. These drugs are now being evaluated for their effectiveness against SARS-CoV-2 RNA-dependent RNA polymerase (RdRp). Regarding the 3CLpro target, numerous studies and ongoing clinical trials have highlighted drugs like Lopinavir [13], Darunavir [14], Ritonavir [15], Ganovo [16], and Cobicistat [17]. Among these, the combination of Ritonavir/Lopinavir (LPV) is frequently tested in clinical trials for COVID-19 treatment. While there is some evidence suggesting LPV's potential efficacy, its significant side effects are a major concern [18, 19]. Moreover, these findings underscore the importance of RdRp and 3CLpro as crucial targets for drug development against SARS-CoV-2, with inhibiting their activity emerging as a promising therapeutic strategy.

To address these challenges, quantitative structure–activity relationship (QSAR) machine learning models have emerged as a promising tool for accelerating drug discovery. QSAR models can predict the activity of compounds against specific targets based on their molecular properties, enabling the rapid screening of large compound libraries. These models have been successfully applied in identifying potential therapeutics for various diseases, including cancer and Alzheimer's [20–22].

In QSAR classification, machine learning algorithms classify compounds based on their chemical structure

and predicted activity. The goal is to build a model that can accurately predict a compound's activity against a target by analyzing its structural features, such as molecular weight and hydrophobicity. The model is trained on a dataset of compounds with known activity, which is divided into training and test sets. The performance of the model is evaluated using metrics like sensitivity, specificity, and accuracy, ultimately leading to a binary classification (active or inactive) for each compound in the test set [23, 24].

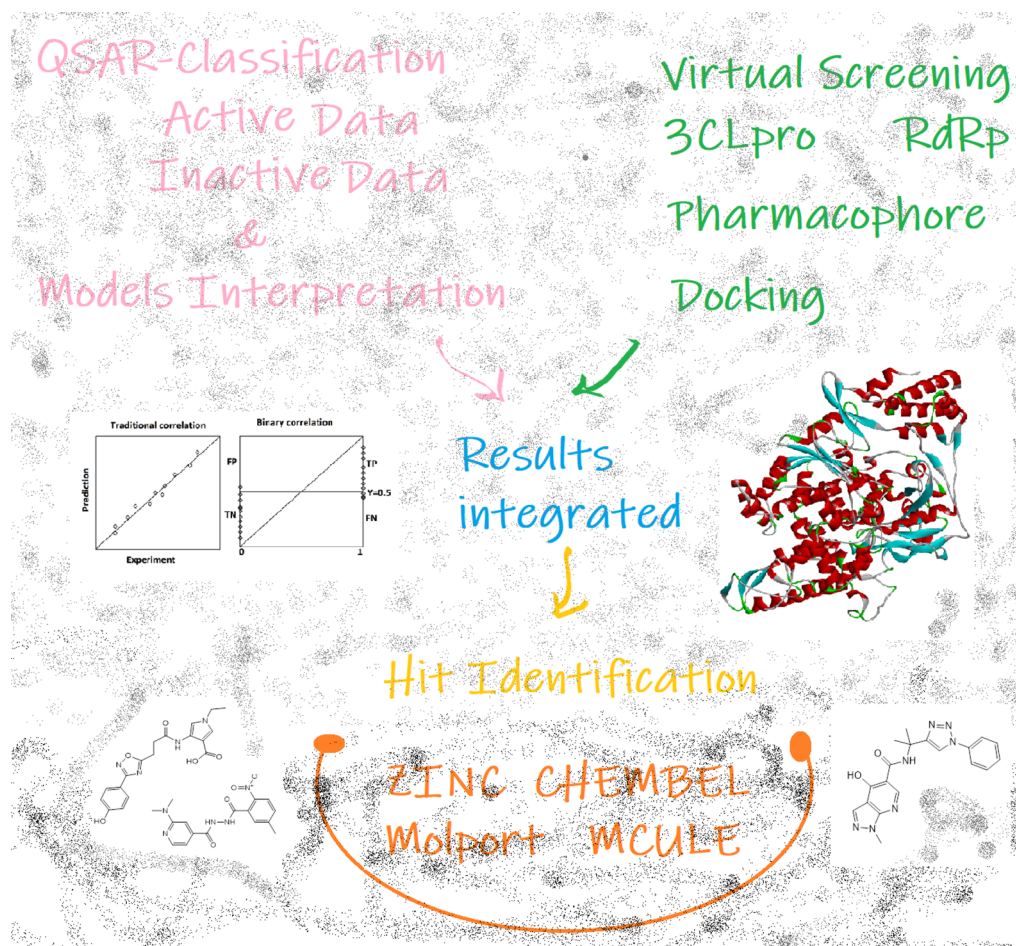
A particularly advantageous approach is using CORAL (Consensus Modeling for Assessing Chemical Toxicity) and SMILES-based QSAR models. These models can handle large and diverse chemical datasets efficiently, thanks to their use of the Simplified Molecular Input Line Entry System (SMILES) notation. This molecular representation captures key structural features while reducing computational costs, making these models well-suited for virtual screening and prioritizing compounds for experimental testing. The CORAL model, which integrates multiple QSAR models and descriptors to generate consensus predictions, further enhances accuracy and robustness [25–27].

In this study, we leverage a dataset compiled by Ivanov et al. [28], which includes compounds tested for activity against the viral targets 3CL<sup>Pro</sup> and RdRp. By applying QSAR machine learning models, we aim to identify additional compounds with the potential to serve as therapeutics for COVID-19 and related viral infections. An overview of the steps of the present study is shown in Fig. 1.

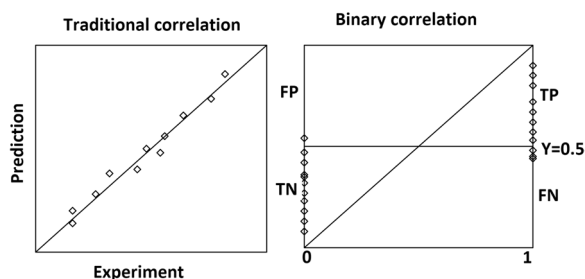
## Materials and methods

### Data collection

2377 molecules were selected from an article published by Julian Ivanov and colleagues in 2020 to investigate their inhibitory potency on the COVID-19 virus proteases [28]. The Simplified Molecular Input Line Entry System (SMILES) strings and IC<sub>50</sub> values used for CORAL input were directly retrieved from the supplementary files of this article. Initially, the CORAL software checks for duplicated chemicals, incorrect SMILES, and inconsistencies in activity data notation based on SMILES. The SMILES format of the compounds was presented in Table S1 and S2. These molecules include 1168 molecules for 3CL<sup>Pro</sup>, consisting of 468 active and 700 inactive molecules, and 1209 molecules for RdRp, consisting of 464 active and 745 inactive molecules. Compounds with an IC<sub>50</sub> of  $\leq 10\mu\text{M}$  were classified as active compounds, and compounds with an IC<sub>50</sub> of  $\geq 10\mu\text{M}$  were classified as inactive compounds. In this study, “semi-correlation” were constructed for



**Fig. 1** Process of molecular modeling in the present study



**Fig. 2** Visualization of the general concepts of the traditional correlation and semi-correlation

ten distinct divisions, as opposed to traditional correlations (Fig. 2)[29, 30]. The splits' identities were also calculated and are displayed in Table 1. It's important to mention that there are no pairs of splits with an identity exceeding 35%.

### CORAL method

The chemical elements in the molecular structure were encoded as symbols for cycles and branching using SMILES attributes, and CORAL software was utilized to build models based on this representation [31]. The CORAL software, which can be downloaded for free from <http://www.insilico.eu/coral>, is a computational tool that utilizes Monte Carlo methods to develop regression and classification models based on SMILES descriptors and their corresponding endpoints (i.e. active or inactive). To build the models, the compounds were randomly divided into four datasets: a training set (30%), an invisible training set (30%), a calibration set (20%), and a validation set (20%). To create the QSAR model, data from the training set was utilized [30]. The training set is the foundation for constructing the model or "builder". Monte Carlo optimization is being employed to adjust correlation weights for molecular features derived from SMILES associated with this particular set. The invisible training set is the

**Table 1** Percentages of identity for random splits

	Set	Split1	Split2	Split3	Split4	Split5	Split6	Split7	Split8	Split9	Split10
3CL <sup>pro</sup>											
Split1	T	100	29.1	29.4	31.6	30.6	29.6	27.8	29.6	30.7	27.6
	IT	100	30.9	31.6	31.1	30.9	34.1	31.7	33.5	30.1	28.3
	C	100	16.4	20.9	21.9	17.1	18.6	20.0	16.1	21.4	19.8
	V	100	22.7	19.3	20.1	16.7	21.8	20.1	23.3	19.6	22.3
Split2	T		100	33.3	28.5	26.6	28.9	27.4	29.8	28.2	27.5
	IT		100	29.2	27.5	28.5	32.0	25.0	29.5	30.8	31.3
	C		100	22.5	19.7	19.5	18.1	19.0	18.9	24.6	17.9
	V		100	17.0	17.4	24.2	19.9	20.8	21.8	21.1	16.0
Split3	T			100	32.4	29.4	35.1	30.1	27.1	29.2	29.7
	IT			100	30.0	30.6	31.3	29.2	29.9	28.1	26.8
	C			100	20.3	23.5	17.0	23.4	22.5	16.1	19.5
	V			100	23.3	22.0	25.0	20.4	20.9	18.6	18.8
Split4	T				100	30.3	30.5	30.5	28.5	28.1	31.5
	IT				100	31.0	30.5	27.2	31.4	27.6	27.8
	C				100	21.1	22.2	23.0	18.1	22.8	22.5
	V				100	17.7	19.8	22.0	19.0	21.0	23.2
Split5	T					100	30.7	30.7	31.0	27.7	29.0
	IT					100	26.5	29.3	26.9	30.9	28.4
	C					100	15.7	20.9	21.2	17.7	21.5
	V					100	17.3	18.6	20.4	21.4	21.2
Split6	T						100	35.6	30.6	29.1	29.5
	IT						100	30.0	29.5	27.1	26.1
	C						100	22.4	21.5	20.5	18.8
	V						100	21.6	16.4	19.0	22.0
Split7	T							100	27.7	31.7	30.1
	IT							100	29.4	32.0	29.3
	C							100	19.5	16.5	24.3
	V							100	18.2	21.5	22.4
Split8	T								100	27.1	32.7
	IT								100	27.8	28.3
	C								100	19.3	21.8
	V								100	16.8	20.0
Split9	T									100	31.2
	IT									100	27.9
	C									100	21.2
	V									100	21.1
Split10	T										100
	IT										100
	C										100
	V										100
RdRp											
Split1	T	100	29.0	30.3	27.1	32.2	30.5	26.7	28.2	29.8	29.0
	IT	100	27.1	28.2	33.6	35.1	33.1	30.5	28.0	33.9	28.2
	C	100	18.4	22.8	18.1	19.9	22.6	22.9	20.4	18.7	18.1
	V	100	19.9	16.1	22.2	24.7	16.2	20.2	19.4	22.6	24.7

**Table 1** (continued)

	Set	Split1	Split2	Split3	Split4	Split5	Split6	Split7	Split8	Split9	Split10
Split2	T		100	26.4	28.7	28.0	31.7	30.3	29.3	28.7	30.4
	IT		100	26.8	30.8	28.3	31.5	29.4	29.2	27.2	28.1
	C		100	19.2	19.1	20.0	13.4	18.1	21.8	18.4	20.1
	V		100	18.9	17.2	20.2	23.5	21.8	20.4	20.4	23.7
Split3	T			100	31.2	31.2	31.8	31.3	30.6	31.7	31.3
	IT			100	32.7	29.2	28.9	30.4	24.9	28.1	27.3
	C			100	21.0	20.6	20.8	16.4	18.4	24.5	26.1
	V			100	22.7	18.5	21.3	23.3	22.5	19.9	20.7
Split4	T				100	31.7	29.7	28.1	31.3	31.3	29.8
	IT				100	34.8	32.0	35.4	28.6	30.4	31.1
	C				100	21.1	18.8	19.0	20.3	21.9	23.2
	V				100	20.8	22.1	20.3	21.4	19.8	18.9
Split5	T					100	25.8	31.8	28.0	31.9	34.2
	IT					100	28.7	33.1	26.3	32.0	30.3
	C					100	16.8	19.9	21.5	21.9	25.6
	V					100	19.0	17.6	18.5	24.4	22.4
Split6	T						100	27.9	28.5	27.3	31.4
	IT						100	32.2	27.8	31.4	26.9
	C						100	23.3	24.4	25.9	23.8
	V						100	21.8	20.8	20.9	21.3
Split7	T							100	30.4	28.0	28.7
	IT							100	25.8	30.1	30.1
	C							100	22.7	21.9	25.8
	V							100	21.1	20.8	24.0
Split8	T								100	27.5	29.2
	IT								100	28.4	28.3
	C								100	23.8	23.6
	V								100	19.1	21.2
Split9	T									100	30.4
	IT									100	31.1
	C									100	26.3
	V									100	23.2
Split10	T										100
	IT										100
	C										100
	V										100

T Training set, IT Invisible Training set, C Calibration set, V Validation set

“inspector” of the model. Computing descriptors for SMILES within this set should either confirm or reject the model’s appropriateness for substances not directly engaged in the optimization procedure. The calibration set should identify the onset of overfitting. Based on computational experiments, it’s evident that optimization enhances the correlation between descriptors and an endpoint for both the training and invisible training sets. However, as the optimization progresses through more epochs, there’s a gradual decrease in the

correlation coefficient between descriptors and the endpoint for the calibration set [32]. The validation set was used to test the predictability of the QSAR model. The CORAL software utilized identical algorithms to compute the classification models. The regression models were established on genuine correlations, while the classification models relied on pseudo correlations. The dataset for the classification models was allocated either a value of 1, indicating “active,” or a value of 0, indicating “inactive.”

### Optimal descriptor

Molecular descriptors, which are mathematical entities, encode the chemical and physical properties of molecules. Choosing an appropriate set of descriptors is crucial as it determines the accuracy of predictive models. 2D representations based on molecular graphs contain around 70–80% of the information in QSAR models. Fragment-based QSAR calculates descriptors for molecular fragments generated by predefined rules. 2D QSAR is simpler and less time-consuming than other methods. Useful fragments can be identified and mapped onto molecules to suggest improvements.

CORAL software offers three distinct descriptor of correlation weights (DCW) types: graph-based, SMILES-based, and hybrid [33]. The graph-based and SMILES-based DCWs are calculated solely based on their respective input types, while the hybrid DCW is generated using both graph and SMILES inputs. To compute the optimal descriptors using SMILES, the following formula is utilized:

$${}^{SMILES}DCW = \sum_{k=1}^N CW(S_k) + \sum_{k=1}^{N-1} CW(SS_k) + \sum_{k=1}^{N-2} CW(SSS_k) + \sum_k CW(NOSP_k) + \sum_k CW(HALO_k) + \sum_k CW(BOND_k) + \sum_k CW(HARD_k) \quad (1)$$

Table 2 contains a detailed description of the SMILES attributes invariants used in Eq. (1).

In the context of SMILES fragments, the threshold is a parameter used to distinguish between rare and non-rare fragments. If the number of SMILES containing a particular fragment in the training set is less than the threshold, it is considered rare; otherwise, it is classified as non-rare. During the Monte Carlo optimization process, one epoch refers to a complete cycle of modifying all correlation weights. The correlation weights, which determine the maximum value of the correlation coefficient

between the endpoint (0,1) and DCW (Threshold, Nepoch), are obtained numerically using the Monte Carlo method.

The target function of the optimization is to find the correlation coefficient between the optimal descriptor and an endpoint. This is achieved through the CORAL model, which establishes a linear relationship between a predicted endpoint  $Y$  and a descriptor of correlation weights (DCW). The DCW is represented by the following mathematical equation:

$$Y = C_0 + C_1 \times DCW(T^*, N^*) \quad (2)$$

Using the least squares method, we can obtain two regression coefficients,  $C_0$  and  $C_1$ , which are used to create an optimal DCW model based on the dataset. During the Monte Carlo optimization process, the parameters  $T^*$  and  $N^*$  are used, where  $T^*$  represents the threshold value and  $N^*$  represents the number of epochs. In DCW modeling,  $T$  refers to the threshold number of active attributes. For example, if  $T$  is set to four, any attributes that appear

in less than four molecules are considered inactive, and their correlation weight will be zero [34–36].

### Statistical criteria

For the binary classification model where “active” is represented by 1 and “inactive” by 0, several statistical measures are used to evaluate the model’s performance. These measures include the Matthews correlation coefficient (MCC), sensitivity, specificity, and accuracy, which are calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

**Table 2** The detailed description of SMILES attributes

ID	Definitions
SMILES $S_k$	Represent one element SMILES attributes
$SS_k$	Represent two element SMILES attributes
$SSS_k$	Represent of combinations of three SMILES attributes
$NOSP_k$	Presence of one or more of four chemical elements (nitrogen, oxygen, sulphur and phosphorus)
$HALO_k$	Presence of fluorine, chlorine, bromine, and iodine
$BOND_k$	Presence or absence of three categories of chemical bonds: double, triple and stereo specific
$HARD_k$	gather together BOND, NOSP, and HALO

A confusion matrix is a table used to assess the performance of a classification model. It includes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) predictions made by the model [30].

The performance measures used in evaluating a classification model are important indicators of its quality. One such measure is the Matthews correlation coefficient (MCC), which is commonly used for binary classification problems [37]. The MCC is similar to the traditional correlation coefficient but is designed specifically for this case. The MCC coefficient, applied in machine learning as a balanced measure of the quality of binary classifications, proving beneficial even when the classes vary significantly in size. In practical terms, a model is considered good if the MCC is 1, or generally, the MCC should be greater than 0.6 for it to be effective. Another important measure is sensitivity, which assesses the model's ability to correctly identify positive observations, such as active compounds. Specificity, on the other hand, evaluates the model's ability to accurately identify negative observations, such as inactive compounds. Finally, accuracy is a measure of the overall performance of the model, taking into account its ability to predict both positive and negative observations. These measures are crucial in determining the effectiveness of a classification model and ensuring its ability to make accurate predictions.

### Applicability domain

The OECD QSAR validation principles require QSAR models to be used within their applicability domain (AD), which refers to the space or knowledge used to develop the model and make predictions for new compounds. Regarding CORAL models, the domain of applicability is determined by analyzing the statistical defects of SMILES, which are calculated based on the distribution of available data in the training, validation, calibration, and validation sets [38]. To define the domain of applicability, the defect of the SMILES attribute is measured as the difference between the probability of the attribute in the training set and its probability in the calibration set. The total SMILES-defect is obtained by summing up the defects of all attributes. If the SMILES-defect of a particular SMILES is less than double the average defect of compounds in the training set, it is considered to be within the domain of applicability; otherwise, it falls outside the domain of applicability.

$$Defect_{A_k} = \frac{|P_{TRN}(A_k) - P_{CAL}(A_k)|}{N_{TRN}(A_k) + N_{CAL}(A_k)}, \text{ if } N_{TRN}(A_k) > 0 \text{ Defect}_{A_k} = 1, \text{ if } N_{TRN}(A_k) = 0 \quad (8)$$

The probabilities of attribute A in the training and calibration sets are denoted by  $P_{TRN}(A_k)$  and  $P_{CAL}(A_k)$ , respectively. On the other hand, the frequencies of A in the training and calibration sets are represented by  $N_{TRN}(A_k)$  and  $N_{CAL}(A_k)$ , respectively.

In the SMILES notation, the statistical defect (D) is the total sum of all statistical defects of all attributes.

$$Defect_{Molecule} = \sum_{k=1}^{NA} Defect_{A_k} \quad (8)$$

In CORAL, the number of active SMILES attributes in a compound is denoted by NA. If a compound is outside the domain of applicability, it is considered an outlier. The detection of outliers in CORAL is based on the condition expressed in inequality 9.

$$Defect_{Molecule} > 2 \times \overline{Defect_{TRN}} \quad (9)$$

The mean value of statistical defects calculated for the training dataset is referred to as  $\overline{Defect_{TRN}}$ .

### Virtual screening workflow

Virtual screening is a computational technique used in drug discovery and development to identify potential drug candidates through in-silico (computer-based) screening of large libraries of small molecules. The process involves the use of computer algorithms and molecular modeling techniques to predict the potential binding affinity of small molecules with target proteins, and thus, identify molecules with a high likelihood of being effective drug candidates [39]. The Pharmit webserver was used to conduct in silico virtual screening on four databases (ChEMBL, ZINC, MCULE, and MolPort) in building pharmacophore (PH4) model and the identification of drug-like molecules stages [40]. Pharmit is an online interactive server that allows for the virtual screening of various compound databases using pharmacophore models and molecular shapes. The reference article identified several molecules with in vitro inhibitory effects on 3CL<sup>PRO</sup>, among which 1H-Indole-2-carboxylic acid, 5-fluoro-, 1H-benzotriazol-1-yl ester (A1) with an IC<sub>50</sub> of 0.013 was considered the most active compound. Similarly, 3-[Isopropyl(trans-4-methylcyclohexylcarbonyl)amino]-5 phenylthiophene -2 carboxylic acid (A2) was considered the most active compound with in vitro inhibitory effect on RdRp, with an IC<sub>50</sub> of 0.009 [28]. To perform the structure-based pharmacophore screening, we utilized these two most active compounds A1 and A2 in the Pharmit server. The QSAR models were

**Table 3** the statistical properties of the inhibitor activity classification models of 3CL<sup>Pro</sup> that were generated through the Monte Carlo method optimization for ten random splits

Split	Set	n	Sensitivity	Specificity	Accuracy	MCC
<i>Inhibitory activity</i> = $-1.2754379(\pm 0.0016968) + 0.0549150(\pm 0.0000625) \times DCW(1, 25)$						
1	Training	339	0.9917	0.9916	0.9916	0.9813
	Invisible training	368	0.9865	1	0.9948	0.9891
	Calibration	230	0.9444	0.9778	0.9644	0.9258
	Validation	231	0.8269	0.9701	0.9375	0.8739
<i>Inhibitory activity</i> = $-0.7655690(\pm 0.0011840) + 0.0538805(\pm 0.0000561) \times DCW(1, 22)$						
2	Training	342	0.9835	1	0.9943	0.9874
	Invisible training	358	0.9733	1	0.9893	0.9778
	Calibration	233	0.9231	0.9574	0.944	0.8823
	Validation	235	0.9216	0.9933	0.9643	0.9266
<i>Inhibitory activity</i> = $-0.4847983(\pm 0.0012123) + 0.0509742(\pm 0.0000518) \times DCW(1, 23)$						
3	Training	354	0.9789	1	0.992	0.983
	Invisible training	341	1	0.9907	0.9938	0.9862
	Calibration	238	0.9541	0.9865	0.9728	0.9443
	Validation	235	0.8962	0.9533	0.9297	0.8547
<i>Inhibitory activity</i> = $-0.5277815(\pm 0.0011310) + 0.0588772(\pm 0.0000562) \times DCW(1, 25)$						
4	Training	338	0.9778	1	0.9915	0.982
	Invisible training	339	0.9291	0.9811	0.9617	0.918
	Calibration	254	0.9072	0.9554	0.937	0.8662
	Validation	237	0.9029	0.9254	0.9156	0.8283
<i>Inhibitory activity</i> = $-0.8095880(\pm 0.0012977) + 0.0540050(\pm 0.0000485) \times DCW(1, 22)$						
5	Training	341	0.9416	1	0.9765	0.9519
	Invisible training	351	0.9568	0.9623	0.9601	0.9169
	Calibration	239	0.9375	0.965	0.954	0.9041
	Validation	237	0.9271	0.9149	0.9198	0.8358
<i>Inhibitory activity</i> = $-0.5912367(\pm 0.0010215) + 0.0665357(\pm 0.0000581) \times DCW(1, 22)$						
6	Training	336	0.9638	0.9956	0.9835	0.9659
	Invisible training	336	0.9481	0.9851	0.9702	0.9381
	Calibration	232	0.8969	0.9481	0.9267	0.8491
	Validation	237	0.9082	0.8849	0.8945	0.7862
<i>Inhibitory activity</i> = $-0.8271686(\pm 0.0011930) + 0.0657067(\pm 0.0000743) \times DCW(1, 17)$						
7	Training	<b>344</b>	<b>0.9593</b>	<b>0.991</b>	<b>0.9797</b>	<b>0.9556</b>
	Invisible training	<b>338</b>	<b>0.9403</b>	<b>0.9804</b>	<b>0.9645</b>	<b>0.9257</b>
	Calibration	<b>250</b>	<b>0.9048</b>	<b>0.9517</b>	<b>0.932</b>	<b>0.8601</b>
	Validation	<b>236</b>	<b>0.9151</b>	<b>0.9692</b>	<b>0.9449</b>	<b>0.889</b>
<i>Inhibitory activity</i> = $-0.5998361(\pm 0.0012035) + 0.0646588(\pm 0.0000541) \times DCW(1, 22)$						
8	Training	363	0.9625	0.9951	0.9807	0.9611
	Invisible training	348	0.9638	0.9905	0.9799	0.958
	Calibration	242	0.8974	0.9329	0.9215	0.8224
	Validation	215	0.9457	0.9431	0.9442	0.8865
<i>Inhibitory activity</i> = $-0.6242973(\pm 0.0012742) + 0.0663009(\pm 0.0000696) \times DCW(1, 22)$						
9	Training	338	0.9242	0.9854	0.9615	0.9193
	Invisible training	336	0.9398	0.9901	0.9702	0.938
	Calibration	246	0.9439	0.9424	0.9431	0.8846
	Validation	248	0.8854	0.9211	0.9073	0.805

selected to conduct virtual screening. Pharmit filters compounds during hit screening based on drug-like properties including number of rotatable bonds, molecular weight,

logP, topological polar surface area, number of HBAs, number of HBDs, and number of aromatic groups. The specified ranges for these properties are: MW ≤ 500, rotatable



**Table 3** (continued)

Split	Set	n	Sensitivity	Specificity	Accuracy	MCC
<i>Inhibitory activity</i> = $-0.7964906(\pm 0.0012467) + 0.0613058(\pm 0.0000611) \times DCW(1, 22)$						
10	Training	341	0.938	0.9906	0.9707	0.9378
	Invisible training	338	0.9771	0.9807	0.9793	0.9565
	Calibration	235	0.8641	0.9621	0.9191	0.8368
	Validation	254	0.7714	0.9597	0.8819	0.7587

bonds  $\leq 10$ ,  $\log P \leq 5$ ,  $PSA \leq 140 \text{ \AA}^2$ , aromatic groups  $\leq 5$ ,  $2 \leq HBA \leq 7$ , and  $2 \leq HBD \leq 7$ . The binding modes of inhibitors, along with the critical molecular interactions inside 3CL<sup>pro</sup> and RdRp active sites with PDB ID 6Y2F and 6NUR, were investigated using the Smina [41] molecular docking package. A purification process was performed by removing all heteroatoms and solvent molecules from the structure. Polar hydrogens were then added to the PDB files. All ligands for docking were sketched using Discovery Studio 2020, and assigned gasteiger charges and energy optimization of ligands using the steepest descent algorithm carried out by Open Babel [42]. The details of the molecular docking algorithm in Smina have been explained in previous studies [36, 43]. Visualization and interaction analyses were performed using the Discovery Studio 2020 viewer.

## Results and discussion

### QSAR models

For the 3CL<sup>pro</sup> enzyme, a collection of 1168 molecules were used, with 468 molecules known to be active against the enzyme and 700 molecules known to be inactive. These molecules were split into four different groups: a training set of 339 molecules, an invisible training set of 368 molecules, a calibration set of 230 molecules, and a validation set of 231 molecules. Similarly, for the RdRp enzyme, a group of 1209 molecules were used, with 464 known to be active and 745 known to be inactive. These molecules were also divided into four groups: a training set of 358 molecules, an invisible training set of 386 molecules, a calibration set of 225 molecules, and a validation set of 240 molecules. To begin with, the process of parameter optimization was carried out in order to determine the most suitable values for the threshold and number of epochs ( $N_{\text{epoch}}$ ). This was done to ensure that the model would produce accurate predictions while minimizing the risk of overfitting.

To analyze the data set for proteases 3CL<sup>pro</sup> and RdRp, it was divided into ten separate parts referred to as “splits” (split 1, split 2, and so on up to split 10). The analysis was conducted using a  $N_{\text{epoch}}$  value of 30 and threshold values ranging from 1 to 3. For all ten splits of both proteases, the Monte Carlo optimization used a preferred threshold value ( $T^*$ ) of 1 and a preferred number of epochs ( $N^*$ ) of 3. The models for each of the ten splits (splits 1 through 10) were

obtained using this methodology. Tables 3 and 4 displays the statistical properties of the binary classifications for 3CL<sup>pro</sup> and RdRp, with the activity being determined using the following formula:

$$Class = \begin{cases} Inhibitory\ activity \leq 10\mu M, class = 1(active) \\ Inhibitory\ activity > 10\mu M, class = 0(inactive) \end{cases} \quad (10)$$

The majority of the models in the study achieved an accuracy, sensitivity, and specificity greater than 90% for the training, invisible training, calibration, and validation sets of 3CL<sup>pro</sup> and RdRp across splits 1–10, indicating their ability to predict the activity of the viral enzymes.

It's important to highlight that the classification model relies on unique semi-correlations, making it unsuitable for using of some traditional criteria. For traditional correlation, an  $r^2$  value around 0.4 suggests a weak regression model. But when it comes to semi-correlation with a similar  $r^2$  value (0.4), having specificity, sensitivity, accuracy, and MCC at that level could be seen as good or even outstanding. This means most items in the chart are accurately classified, usually with just one false positive and one false negative. The CORAL software operates under the fundamental assumption that a well-performing model on a calibration set should also perform well on an external validation set. In this case, the statistical parameters of the models for the two proteases are deemed good and acceptable. Furthermore, for split #7 (bold in the Table 3) of 3CL<sup>pro</sup> and split #4 (bold in Table 4) of RdRp, the MCC values for the validation sets are among the top-performing models, with 0.9478 and 0.8890, respectively.

By running Monte Carlo optimization multiple times, one can obtain three groups of SMILES attributes: (i) attributes or promoters that only have positive correlation weights, which promote the activity of compounds; (ii) attributes or promoters that only have negative correlation weights, which promote the inactivity of compounds; and (iii) attributes or promoters that have both positive and negative correlation weights in multiple Monte Carlo runs, and their role is not yet understood. This method allows for a mechanistic interpretation of the model. Tables 5 and 6 provides a list of potential

**Table 4** The statistical properties of the inhibitor activity classification models of RdRp that were generated through the Monte Carlo method optimization for ten random splits

Split	Set	n	Sensitivity	Specificity	Accuracy	MCC
<i>Inhibitory activity</i> = $-1.6278890(\pm 0.0011429) + 0.0552155(\pm 0.0000444) \times DCW(1, 25)$						
1	Training	358	0.9917	0.9916	0.9916	0.9813
	Invisible training	386	0.9865	1	0.9948	0.9891
	Calibration	225	0.9444	0.9778	0.9644	0.9258
	Validation	240	0.8269	0.9701	0.9375	0.8739
<i>Inhibitory activity</i> = $-1.6484312(\pm 0.0011655) + 0.0498625(\pm 0.0000398) \times DCW(1, 16)$						
2	Training	352	0.9835	1	0.9943	0.9874
	Invisible training	373	0.9733	1	0.9893	0.9778
	Calibration	232	0.9231	0.9574	0.944	0.8823
	Validation	252	0.9216	0.9933	0.9643	0.9266
<i>Inhibitory activity</i> = $-1.7380462(\pm 0.0012938) + 0.0418836(\pm 0.0000323) \times DCW(1, 18)$						
3	Training	374	0.9789	1	0.992	0.983
	Invisible training	322	1	0.9907	0.9938	0.9862
	Calibration	257	0.9541	0.9865	0.9728	0.9443
	Validation	256	0.8962	0.9533	0.9297	0.8547
<i>Inhibitory activity</i> = $-1.7574849(\pm 0.0012422) + 0.0632141(\pm 0.0000445) \times DCW(1, 18)$						
4	Training	<b>351</b>	<b>0.9778</b>	<b>1</b>	<b>0.9915</b>	<b>0.982</b>
	Invisible training	<b>393</b>	<b>0.9868</b>	<b>1</b>	<b>0.9949</b>	<b>0.9893</b>
	Calibration	<b>228</b>	<b>0.95</b>	<b>0.9595</b>	<b>0.9561</b>	<b>0.9044</b>
	Validation	<b>237</b>	<b>0.9592</b>	<b>0.9856</b>	<b>0.9747</b>	<b>0.9478</b>
<i>Inhibitory activity</i> = $-1.7539257(\pm 0.0013973) + 0.0536652(\pm 0.0000380) \times DCW(1, 18)$						
5	Training	369	0.9655	0.9955	0.9837	0.966
	Invisible training	349	0.9699	0.9954	0.9857	0.9697
	Calibration	228	0.9405	0.9792	0.9649	0.9244
	Validation	263	0.9706	0.9689	0.9696	0.9363
<i>Inhibitory activity</i> = $-1.5928815(\pm 0.0014258) + 0.0539876(\pm 0.0000436) \times DCW(1, 19)$						
6	Training	336	0.9728	1	0.9881	0.976
	Invisible training	370	0.9856	0.9957	0.9919	0.9827
	Calibration	261	0.9694	0.9816	0.977	0.951
	Validation	242	0.95	0.9506	0.9504	0.89
<i>Inhibitory activity</i> = $-1.7698491(\pm 0.0012758) + 0.0505012(\pm 0.0000404) \times DCW(1, 21)$						
7	Training	360	0.9857	0.9955	0.9917	0.9825
	Invisible training	369	0.9756	0.9959	0.9892	0.9756
	Calibration	255	0.9906	0.9933	0.9922	0.9839
	Validation	225	0.9158	0.9846	0.9556	0.9095
<i>Inhibitory activity</i> = $-1.7604432(\pm 0.0013978) + 0.0492024(\pm 0.0000376) \times DCW(1, 21)$						
8	Training	345	0.9867	0.9897	0.9884	0.9764
	Invisible training	313	0.9752	1	0.9904	0.9799
	Calibration	264	0.9684	0.9941	0.9948	0.9671
	Validation	287	0.9388	0.963	0.9547	0.8996
<i>Inhibitory activity</i> = $-1.8181568(\pm 0.0013362) + 0.0322936(\pm 0.0000265) \times DCW(1, 14)$						
9	Training	346	0.985	0.9953	0.9913	0.9817
	Invisible training	369	0.9632	0.9957	0.9837	0.9651
	Calibration	257	0.9574	0.9877	0.9767	0.9496
	Validation	237	0.901	0.9929	0.9536	0.9068

promoters that be linked to both increased and decreased activities for both proteases. The findings from Table 5 revealed that certain structural features, such as aliphatic

oxygen and double bond, nitrogen and double bond, aliphatic oxygen and double bond with branching, two successive aliphatic carbon, aliphatic carbon and double

**Table 4** (continued)

Split	Set	n	Sensitivity	Specificity	Accuracy	MCC
<i>Inhibitory activity</i> = $-1.5764651(\pm 0.0011165) + 0.0590559(\pm 0.0000446) \times DCW(1, 24)$						
10	Training	346	0.9925	0.9906	0.9913	0.9817
	Invisible training	369	0.9632	1	0.9864	0.9711
	Calibration	257	0.9468	0.9877	0.9728	0.9412
	Validation	237	0.901	0.9706	0.9409	0.8795

bond with branching, double bond, presence of at least three rings, carbon and double bond with branching, aliphatic carbon and aliphatic nitrogen with branching, aliphatic carbon and aliphatic oxygen with branching, molecule containing nitrogen and oxygen with at least one ring and branching, were identified as significant factors for increased inhibitory activity against 3CL<sup>Pro</sup>. On the other hand, decreased activity against 3CL<sup>Pro</sup> was associated with certain structural features, such as nitrogen and oxygen with double bond, oxygen and double bond, three successive aliphatic carbon, aliphatic carbon and double bond with at least three rings, hydrogen and stereo specific bonds.

Table 6 presented our findings on the influential features that affect inhibitory activity against RdRp. We observed that negative charge, nitrogen and double bond, aliphatic carbon with branching, double bond, nitrogen and oxygen, double bond, aliphatic carbon and aliphatic nitrogen, aliphatic carbon and branching, presence of at least two rings, the highest number of sulfur equal to zero, oxygen and double bond, and aromatic carbon in the first ring had a positive impact on inhibitory activity. On the other hand, aliphatic carbon with branching and aliphatic oxygen, aliphatic oxygen and two branching, aliphatic carbon and branching, aliphatic nitrogen with branching and aliphatic carbon, branching with aliphatic carbon and stereo specific bonds, successive two aliphatic carbon and aliphatic nitrogen, aliphatic carbon and double bond, aromatic nitrogen in the second ring, the highest number of oxygens equal to one, two successive aliphatic carbons and ring, carbon and double bond with branching, and presence of three successive aliphatic carbon were found to decrease inhibitory activity against RdRp.

#### Virtual screening analysis

The Pharmit server was utilized to propose pharmacophore characteristics by emphasizing on the key residues involved in active site interactions of 3CL<sup>Pro</sup> and RdRp with compound A1 and A2 as the most active compounds, respectively (Fig. 3). Among the databases supported by Pharmit, we opted to screen the ZINC, ChEMBL32, MCULE, and MolPort databases due to

the availability of purchasable compounds for virtual screening. The QSAR models were selected for predicting the activity (active or inactive) of compounds. Then, the filter was established using typical molecular characteristics for recognizing drug-like molecules. These features comprise molecular weight, log P (a measurement of lipophilicity), topological polar surface area (a sign of the compound's ability to penetrate cell membranes), the number of rotatable bonds, the number of aromatic groups, the number of hydrogen bond acceptors, and the number of hydrogen bond donors. OpenBabel is utilized to precompute these features. Smina was used in the third screening to confirm that the search for the binding site of 3CL<sup>Pro</sup> and RdRp had produced all hits. The compounds were then sorted based on their docking score values, and only those with scores higher than compounds A1 with 3CL<sup>Pro</sup> ( $-7.22$  kcal/mol) and A2 with RdRp ( $-7.84$  kcal/mol) were included in the list. Figure 3 shows the PH4 models and interaction patterns of two complexes: one between 3CL<sup>Pro</sup> and A1, and the other between RdRp and A2.

The results of molecular docking indicate that the A1 molecule forms hydrogen bonds with three different amino acids, namely ARG188, GLN192, and THR190. Additionally, this molecule exhibits hydrophobic interactions with amino acids PRO168, MET165, and HIS41. The amino acids GLN189 and VAL186 are also found to be involved in van der Waals interactions with the A1 molecule. The findings of molecular docking analysis between RdRp protease with molecule A2 indicated that GLN724 amino acid forms a hydrogen bond with the carbonyl group's oxygen. Additionally, LEU708 and ARG721 were observed to engage in hydrophobic interactions with the rings and branches, while THR710, GLY712, HIS725, and TYR732 were found to participate in van der Waals interactions. The results are consistent with the pharmacophoric models.

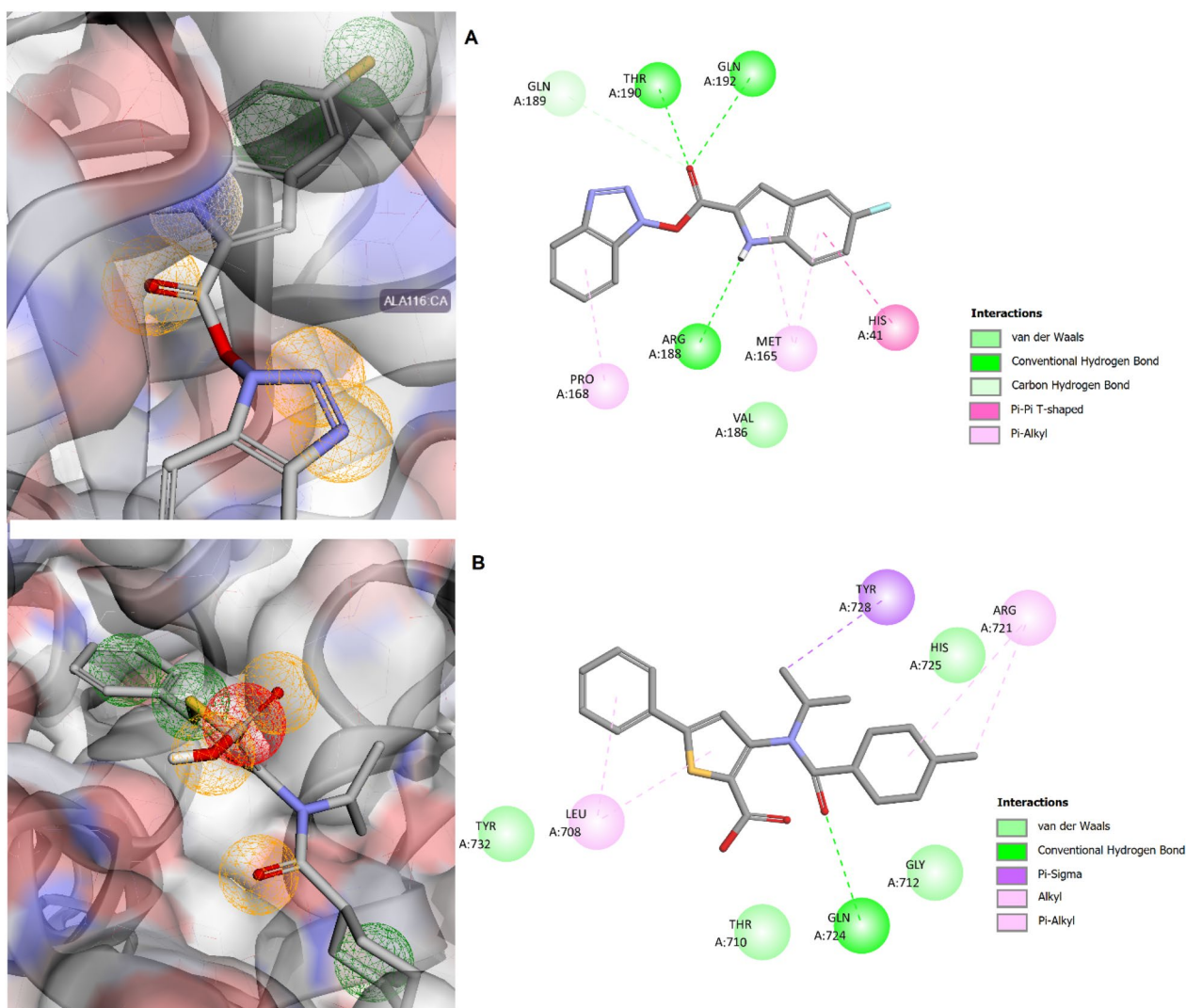
Figure 4 presents the hits obtained from virtual screening procedures across all four databases for each protein. During the final stage of screening, a certain number of compounds were subjected to docking simulations using different proteins, with 156 compounds docked into 3CL<sup>Pro</sup> and 51 compounds into RdRp. The structures and

**Table 5** A collection of SMILES attributes that can be interpreted as promoters of activity or inactivity of compounds against 3CL<sup>Pro</sup>

SMILES attribute	Interpretation
Promoter of activity increase	
<chem>O...=.....</chem>	Presence of aliphatic oxygen and double bond
<chem>+++ +N--B2= =</chem>	Presence of nitrogen and double bond
<chem>=...O ... (...</chem>	Presence of aliphatic oxygen and double bond and branching
<chem>C...C.....</chem>	Presence of two successive aliphatic carbon
<chem>C...=... (...</chem>	Presence of aliphatic carbon and double bond with branching
<chem>BOND10000000</chem>	Presence of double bond in the structure of compound
<chem>C...3.....</chem>	Presence of at least three rings
<chem>C... (... =...</chem>	Presence of carbon and double bond and branching
<chem>C...N ... (...</chem>	Presence of aliphatic carbon and aliphatic nitrogen and branching
<chem>C...O ... (...</chem>	Presence of aliphatic carbon and aliphatic oxygen and branching
<chem>C...2...=...</chem>	Presence of at least two rings and double bond
<chem>NOSP11000000</chem>	Molecule contains nitrogen and oxygen
<chem>1... (......</chem>	Presence of at least one ring with branching
Promoter of activity decrease	
<chem>+++ +N--O== =</chem>	Presence of nitrogen and oxygen and double bond
<chem>+++ +O--B2= =</chem>	Presence of oxygen and double bond
<chem>C...C...C...</chem>	Presence of three successive aliphatic carbon
<chem>C...=...3...</chem>	Presence of aliphatic carbon and double bond and at least three rings
<chem>H...@.....</chem>	Presence of hydrogen and stereo specific bonds

**Table 6** A collection of SMILES attributes that can be interpreted as promoters of activity or inactivity of compounds against RdRp

SMILES attribute	Interpretation
Promoter of activity increase	
<chem>-.....</chem>	Presence of negative charge
<chem>+++ +N--B2= =</chem>	Presence of nitrogen and double bond
<chem>C... (...</chem>	Presence of aliphatic carbon with branching
<chem>BOND10000000</chem>	Presence of double bond
<chem>+++ +N--O== =</chem>	Presence of nitrogen and oxygen and double bond
<chem>N...C ...</chem>	Presence of aliphatic carbon and aliphatic nitrogen
<chem>1...C... (...</chem>	Presence of aliphatic carbon and branching
<chem>C...2...</chem>	Presence of at least two rings
<chem>Smax.0...</chem>	The highest number of sulfur equal to zero
<chem>+++ +O--B2= =</chem>	Presence of oxygen and double bond
<chem>1...C... (......</chem>	Presence of aromatic carbon at first ring
Promoter of activity decrease	
<chem>C...(...O</chem>	Presence of aliphatic carbon with branching and aliphatic oxygen
<chem>O... (... (</chem>	Presence of aliphatic oxygen and two branching
<chem>C... [... C...</chem>	Presence of aliphatic carbon and branching
<chem>N... (...C...</chem>	Presence of aliphatic nitrogen with branching and aliphatic carbon
<chem>[...C...@...]</chem>	Presence of branching with aliphatic carbon and stereo specific bonds
<chem>N...C...C...</chem>	Presence of successive two aliphatic carbon and aliphatic nitrogen
<chem>C...=...</chem>	Presence of aliphatic carbon and double bond
<chem>n...2...</chem>	Presence of aromatic nitrogen in second ring
<chem>Omax.1</chem>	The highest number of oxygens equal to one
<chem>C...C...1...</chem>	Presence of two successive aliphatic carbons and ring
<chem>C... (... =...</chem>	Presence of carbon and double bond with branching
<chem>C...C...C...</chem>	Presence of three successive aliphatic carbon

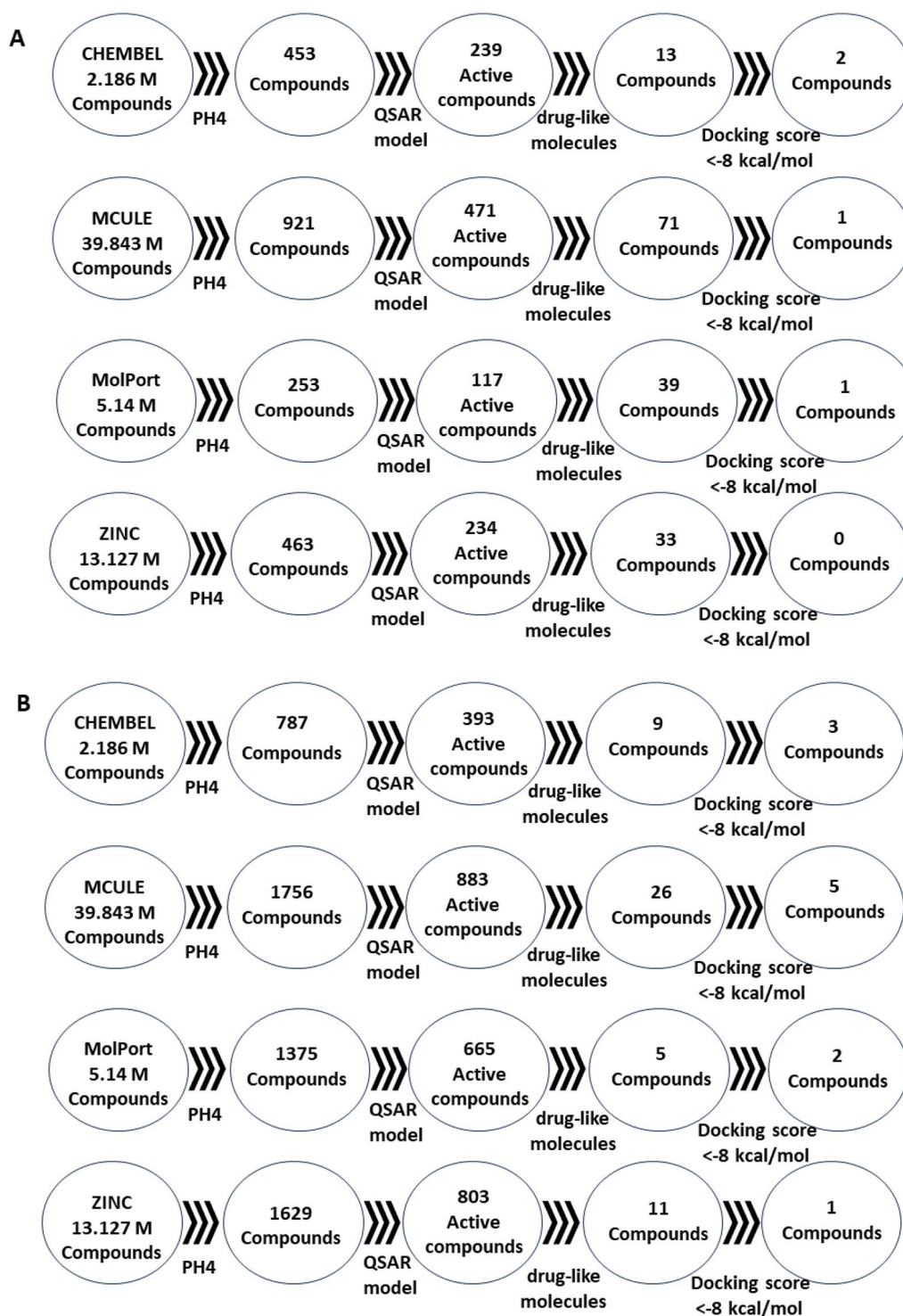


**Fig. 3** The interaction patterns and superpositions of the PH4 models for 3CL<sup>Pro</sup> and RdRp with PDB ID (A) 6Y2F with A1 and (B) 6NUR with A2. PH4 features are represented using spheres colored green for hydrophobic, orange for hydrogen bond acceptor, white for hydrogen bond donor, and red for negative ion features

molecular docking minimized affinity values of these hits were shown in Table 7.

Specifically, the hit structures M3 for 3CL<sup>Pro</sup>, attributed to the presence of rings and oxygen with a double bond, were found to enhance the compound's activity in the QSAR model. Additionally, N2 and N4 for RdRp, characterized by the presence of activity-promoting factors such as nitrogen and a ring in the chain, and the absence of sulfur, are highly compatible with the QSAR models. As a result, they were studied in greater detail. Upon analyzing the interaction between M3 and 3CL<sup>Pro</sup>, it was observed that there were significant hydrogen

bonds formed between the oxygen and nitrogen of M3 with THR25, LEU141, GLY143, SER144, and CYS145. Additionally, two hydrophobic interactions occurred between the two rings of M3 and MET49 and CYS145 (Fig. 5A). According to Fig. 5B, the N2 molecule bound to RdRp forms three hydrogen bonds from N and O atoms with ASN713, GLN724, and TYR732 amino acid residues. Furthermore, the three rings of the molecule engage in hydrophobic interactions with specific amino acid residues such as VAL128, HIS133, LEU207, LEU240, GLY712, and LEU708. The docking results of N5 with RdRp show that there is a hydrogen bond

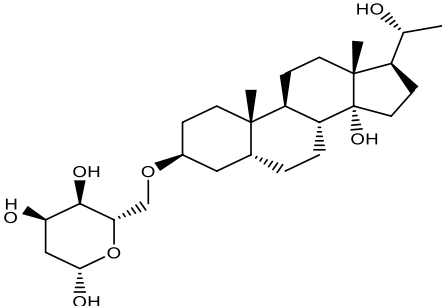
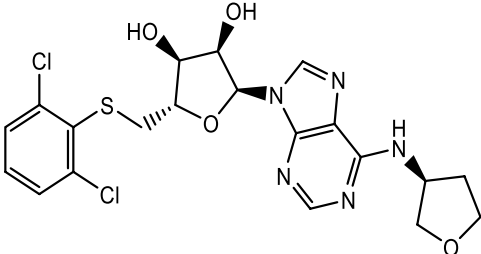
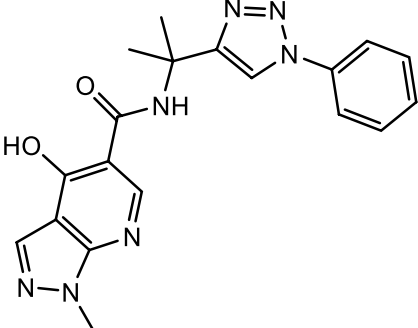
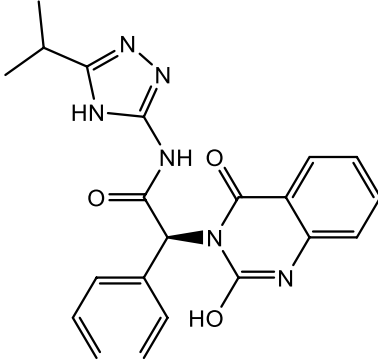


**Fig. 4** Result diagram screening of inhibitors for (A) 3CL<sup>Pro</sup> and (B) RdRp from the databases

between N5 and TYR732. Hydrophobic interactions were observed between N5 and VAL128, TYR129, HIS133, LEU240, LEU708, and TYR728. Furthermore, ARG132 and ASP465 were involved in attractive charge and salt

bridges with the charged part of the molecule (Fig. 5C). The M3, N2, and N4 compounds contain at least an amine functional group, which is known to play a crucial role in drug-target binding interactions. Amine groups

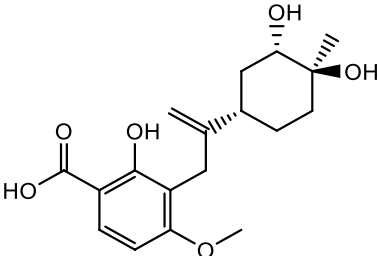
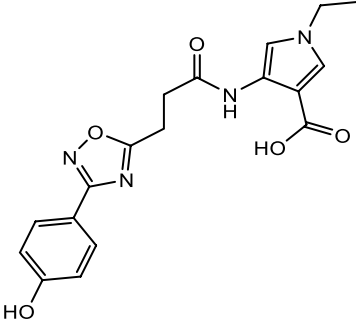
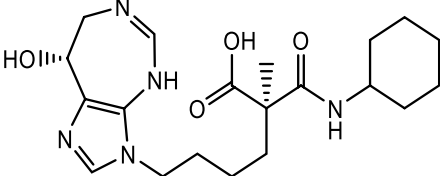
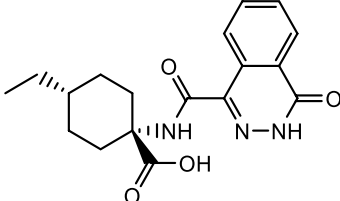
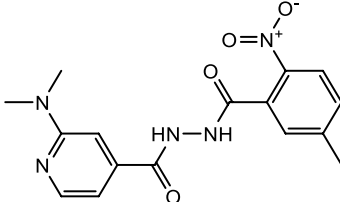
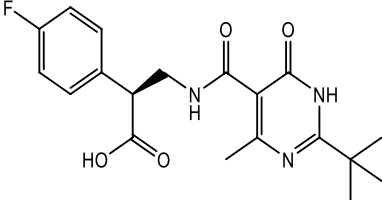
**Table 7** Hits retrieved from the virtual screening alongside their minimized affinity values

No	Hit name in database	Structure	Minimized affinity (kcal/mol)
3CL <sup>Pro</sup> M1	CHEMBL253085471 PubChem-76327933		- 8.62
M2	CHEMBL25609538 PubChem-46876171		- 8.09
M3	MCULE-4083440218		- 8.37
M4	MolPort-044-566-517 MCULE-5384906049		- 8.05

can pass through cell membranes when in their non-ionized form, and when ionized, they exhibit favorable solubility in water. Furthermore, the fact that the three compounds are highly soluble in water makes them ideal

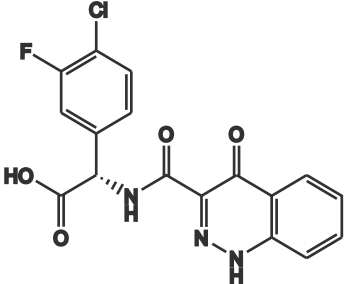
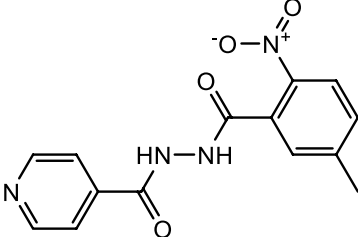
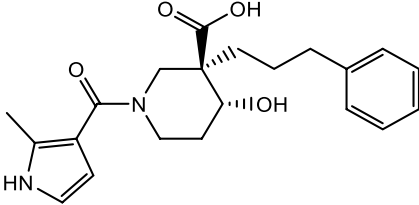
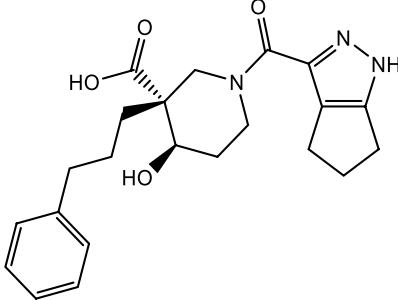
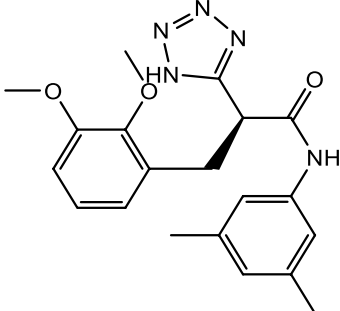
for various drug production activities. It is worth noting that solubility is a critical feature that affects absorption, especially for discovery projects that aim for oral administration. Additionally, a drug intended for injection must

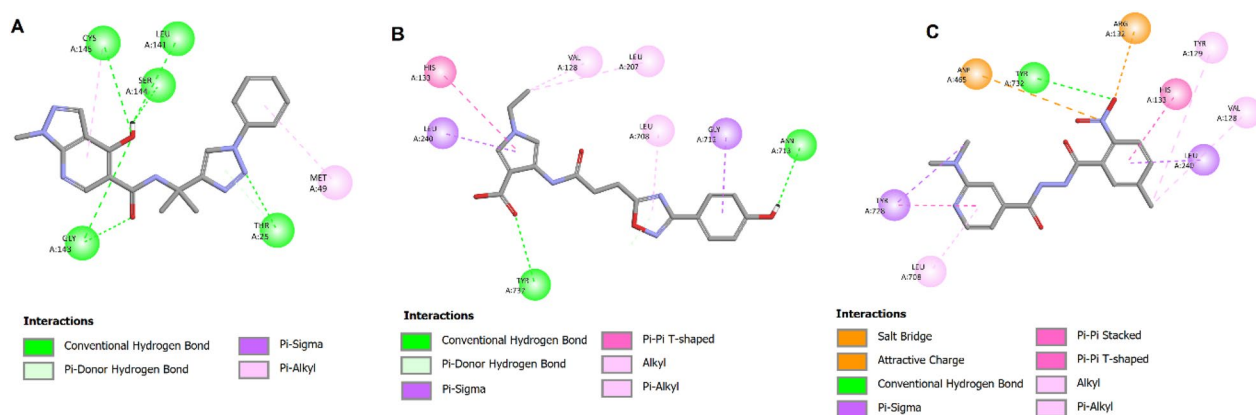
**Table 7** (continued)

No	Hit name in database	Structure	Minimized affinity (kcal/mol)
RdRp N1	CHEMBL25149284 PubChem-44364935		-8.25
N2	CHEMBL254636989 PubChem-136375410 PubChem-136375449		-8.22
N3	CHEMBL25289639 PubChem-10644984		-8.22
N4	MCULE-7755982919 CSC057656310 PubChem-120203617		-8.13
N5	MCULE-3802558326		-8.12
N6	MCULE-1230725924 CSC057726018 PubChem-137030145		-8.1



**Table 7** (continued)

No	Hit name in database	Structure	Minimized affinity (kcal/mol)
N7	MCULE-2082745487		- 8.07
N8	MCULE-7352972814		- 8.03
N9	MolPort-051-456-150		- 8.42
N10	MolPort-051-445-187		- 8.14
N11	ZINC000005488750 5488750		- 8.55



**Fig. 5** The interaction patterns models for (A) 6Y2F with M3, (B) 6NUR with N2 and (C) 6NUR with N5

**Table 8** The projected ADMET characteristics for the identified hits

Name	Bioavailability score	GI absorption	log Kp (cm/s)	logS	Mutagenic	Tumorigenic	Reproductive effective	Irritant
M1	0.55	High	-7.84	-4.12	None	None	None	None
M2	0.55	Low	-7.03	-5.87	None	None	None	None
M3	0.55	High	-7.29	-3.8	None	None	None	None
M4	0.55	High	-6.92	-4.89	None	None	Low	None
N1	0.55	High	-6.37	-4.7	None	None	None	None
N2	0.55	High	-7.37	-4.02	None	None	None	None
N3	0.55	High	-7.94	-3.48	None	None	None	None
N4	0.56	High	-6.7	-4.39	None	None	None	None
N5	0.55	High	-7.01	-4.1	None	High	High	None
N6	0.55	High	-6.9	-4.38	None	None	None	None
N7	0.55	High	-6.06	-5.6	None	None	None	None
N8	0.56	High	-7.07	-3.55	None	High	High	None
N9	0.56	High	-6.54	-4.46	None	None	None	None
N10	0.55	High	-6.62	-4.87	None	None	None	None
N11	0.55	High	-6.48	-4.83	None	None	None	None

have excellent water solubility to ensure that a small medicinal dose contains enough of the active substance. The three compounds possess a significant number of nitrogens with non-bonding electrons, which suggests that they exhibit hydrophilic properties. As a result, they are logical choices for inhibiting 3CL<sup>PRO</sup> and RdRp and can be considered as potential candidates.

#### ADMET properties

ADMET stands for Absorption, Distribution, Metabolism, Excretion, and Toxicity. These are critical factors that affect the pharmacokinetics (how drugs move through the body) and pharmacodynamics (how drugs interact with the body) of a drug. The SwissADME server [44] and the DataWarrior [45] software were used to

compute various measures, including the bioavailability score, gastrointestinal absorption, logKp for skin permeation, water solubility and toxicity. The Abbott bioavailability score is used to assess drug-likeness, with a score of 0.55 indicating that the best-predicted hits from virtual screening passed the rule-of-five. Skin permeability, logKp was also calculated and fell within the standard range of -1 to -8 for 95% of drugs [46]. The range of water solubility typically reported in drug discovery and development is  $-6.5 < \log S < 0.5$ . There are various factors that regulate bioavailability, but ultimately, gastrointestinal absorption is the key determinant [47]. Among the 15 hits identified based on molecular docking scores, M3, N2, and N4 were identified as promising inhibitors due to their good synthetic accessibility scores (3.07,

3.11, and 3.29 out of 10 for M3, N2, and N4 respectively). The reported hits showed high gastrointestinal absorption, and toxicity risk was assessed for mutagenic, tumorigenic, irritant, and reproductive effects. Reproductive toxicity may cause alterations to the male and female reproductive systems, while irritant toxicity can cause reversible damage to the skin or other organs. A majority of the virtual screening hits exhibited satisfactory molecular properties, as shown in Table 8.

## Conclusion

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, poses a significant threat to global health. To support the development of effective treatments, we combined ligand-based and structure-based drug design approaches to identify potent inhibitors against SARS-CoV-2. SMILES-based classification models were used to create predictive two-dimensional QSAR models. Our comprehensive virtual screening workflow included PH4 analysis, QSAR modeling, evaluation of drug-like properties, molecular docking, and ADMET testing. The ease of using CORAL software to generate QSAR models proved advantageous for rapidly screening potential compounds, as it reduces the chemical space to a manageable size for further analysis and development. This approach identified 15 potential inhibitors, with M3, N2, and N4 emerging as the most promising candidates due to their favorable synthetic accessibility scores (3.07, 3.11, and 3.29, respectively) and the presence of amine functional groups, which are crucial for effective binding interactions with drug targets. These compounds have been selected for further biological assays to validate their efficacy. This study provides valuable insights into the development of novel inhibitors with potential therapeutic applications for COVID-19. To confirm our computational findings, experimental evaluation of the identified compounds against RdRp and 3CLpro activity using standard enzyme-based or cell-based assays is recommended.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13065-024-01302-3>.

Additional file 1.

## Author contributions

Faezeh Bazzi-Allahri: Visualization, Methodology, Formal analysis. Fereshteh Shir: Writing—review & editing, Writing—original draft, Supervision, Funding acquisition, Conceptualization. Shahin Ahmad: review & editing, Validation, Software. Alla P. Toropova: review & editing, software. Andrey A. Toropov: review & editing, software.

## Funding

This work was funded by the University of Zabol [Grant code: IR-UOZ-GR-0144].

## Availability of data and materials

Data is provided within the manuscript and supplementary information files.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 15 April 2024 Accepted: 18 September 2024

Published online: 03 October 2024

## References

- Li J, Wang J, Wang H. Emerging landscape of preclinical models for studying COVID-19 neurologic diseases. *ACS Pharmacol Transl Sci*. 2023;6(10):1323–39.
- Farha MA, Brown ED. Drug repurposing for antimicrobial discovery. *Nat Microbiol*. 2019;4(4):565–77.
- Dai W, Jochmans D, Xie H, Yang H, Li J, Su H, Chang D, Wang J, Peng J, Zhu L. Design, synthesis, and biological evaluation of peptidomimetic aldehydes as broad-spectrum inhibitors against enterovirus and SARS-CoV-2. *J Med Chem*. 2021;65(4):2794–808.
- Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, Wang Q, Xu Y, Li M, Li X. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*. 2020;10(5):766–88.
- Kneller DW, Phillips G, O'Neill HM, Jedrzejczak R, Stols L, Langan P, Joachimiak A, Coates L, Kovalevsky A. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. *Nat Commun*. 2020;11(1):3202.
- Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, Wang T, Sun Q, Ming Z, Zhang L. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*. 2020;368(6492):779–82.
- Xue H, Li J, Xie H, Wang Y. Review of drug repositioning approaches and resources. *Int J Biol Sci*. 2018;14(10):1232.
- Leelananda SP, Lindert S. Computational methods in drug discovery. *Beilstein J Org Chem*. 2016;12(1):2694–718.
- Hung I. Lopinavir/ritonavir, ribavirin and IFN-beta combination for nCoV treatment. *NCT04276688*. 2020.
- USNLo M. A Study to Evaluate the Safety, Pharmacokinetics and Antiviral Effects of Galidesivir in Yellow Fever or COVID-19. *ClinicalTrials.gov*. 2020.
- Wang Y, Zhang D, Du G, Du R, Zhao J, Jin Y, Fu S, Gao L, Cheng Z, Lu Q. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet*. 2020;395(10236):1569–78.
- Polo R, Hernán M. Randomized clinical trial for the prevention of SARS-CoV-2 infection (COVID-19) in healthcare personnel (EPICOS). *ClinicalTrials.gov*; 2020.
- Wang Z, Xu X. scRNA-seq profiling of human testes reveals the presence of the ACE2 receptor, a target for SARS-CoV-2 infection in spermatogonia, Leydig and Sertoli cells. *Cells*. 2020;9(4):920.
- De Meyer S, Bojkova D, Cinatl J, Van Damme E, Buyck C, Van Loock M, Woodfall B, Ciesek S. Lack of antiviral activity of darunavir against SARS-CoV-2. *Int J Infect Dis*. 2020;97:7–10.
- Ye X-T, Luo Y-L, Xia S-C, Sun Q-F, Ding J-G, Zhou Y, Chen W, Wang X-F, Zhang W-W, Du W-J. Clinical efficacy of lopinavir/ritonavir in the treatment of Coronavirus disease 2019. *Eur Rev Med Pharmacol Sci*. 2020;24(6):3390–6.
- Chen YW, Yiu CPB, Wong K-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*. 2020;9:129.

17. Nabirotkhin S, Peluffo AE, Bouaziz J, Cohen D. Focusing on the unfolded protein response and autophagy related pathways to reposition common approved drugs against COVID-19. Basel: MDPI AG; 2020.
18. Verdugo-Paiva F, Izcovich A, Ragusa M, Rada G. C.-L.-OW Group, Lopinavir/ritonavir for the treatment of COVID-19: a living systematic review protocol. medRxiv. 2020;9:399.
19. Tobaiqy M, Qashqary M, Al-Dahery S, Mujallad A, Hershman AA, Kamal MA, Helmi N. Therapeutic management of patients with COVID-19: a systematic review. *Infect Prevent Pract.* 2020;2(3):100061.
20. Vatanserver S, Schlessinger A, Wacker D, Kaniskan HÜ, Jin J, Zhou MM, Zhang B. Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: state-of-the-arts and future directions. *Med Res Rev.* 2021;41(3):1427–73.
21. Prachayasittikul V, Worachartcheewan A, Toropova A, Toropov A, Schadu-angrat N, Prachayasittikul V, Nantasenamat C. Large-scale classification of P-glycoprotein inhibitors using SMILES-based descriptors. *SAR QSAR Environ Res.* 2017;28(1):1–16.
22. Tsou LK, Yeh S-H, Ueng S-H, Chang C-P, Song J-S, Wu M-H, Chang H-F, Chen S-R, Shih C, Chen C-T. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Sci Rep.* 2020;10(1):16771.
23. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in cheminformatics and drug discovery. *Drug Discov Today.* 2018;23(8):1538–46.
24. Shiri F, Pirhadi S, Rahmani A. Identification of new potential HIV-1 reverse transcriptase inhibitors by QSAR modeling and structure-based virtual screening. *J Recept Signal Transduct.* 2018;38(1):37–47.
25. Cappelli CI, Toropov AA, Toropova AP, Benfenati E. Ecosystem ecology: Models for acute toxicity of pesticides towards *Daphnia magna*. *Environ Toxicol Pharmacol.* 2020;80:103459.
26. Lotfi S, Ahmadi S, Zohrabi P. QSAR modeling of toxicities of ionic liquids toward *Staphylococcus aureus* using SMILES and graph invariants. *Struct Chem.* 2020;31:2257–70.
27. Manganelli S, Benfenati E, Manganaro A, Kulkarni S, Barton-Maclaren TS, Honma M. New quantitative structure–activity relationship models improve predictability of Ames mutagenicity for aromatic azo compounds. *Toxicol Sci.* 2016;153(2):316–26.
28. Ivanov J, Polshakov D, Kato-Weinstein J, Zhou Q, Li Y, Granet R, Garner L, Deng Y, Liu C, Albaiu D. Quantitative structure–activity relationship machine learning models and their applications for identifying viral 3CLpro- and RdRp-targeting compounds as potential therapeutics for COVID-19 and related viral infections. *ACS Omega.* 2020;5(42):27344–58.
29. Toropov A, Toropova A, Benfenati E, Gini G, Leszczynska D, Leszczynski J. CORAL: classification model for predictions of anti-sarcoma activity. *Curr Top Med Chem.* 2012;12(24):2741–4.
30. Toropov AA, Toropova AP, Rasulev BF, Benfenati E, Gini G, Leszczynska D, Leszczynski J. CORAL: binary classifications (active/inactive) for liver-related adverse effects of drugs. *Curr Drug Saf.* 2012;7(4):257–61.
31. Toropova AP, Toropov AA. CORAL: binary classifications (active/inactive) for drug-induced liver injury. *Toxicol Lett.* 2017;268:51–7.
32. Toropova AP, Toropov AA. QSPR and nano-QSPR: what is the difference? *J Mol Struct.* 2019;1182:141–9.
33. Ahmadi S. Mathematical modeling of cytotoxicity of metal oxide nanoparticles using the index of ideality correlation criteria. *Chemosphere.* 2020;242:125192.
34. Lotfi S, Ahmadi S, Kumar P. A hybrid descriptor based QSPR model to predict the thermal decomposition temperature of imidazolium ionic liquids using Monte Carlo approach. *J Mol Liq.* 2021;338:116465.
35. Kumar A, Kumar P. Cytotoxicity of quantum dots: Use of quasiSMILES in development of reliable models with index of ideality of correlation and the consensus modelling. *J Hazard Mater.* 2021;402:123777.
36. Soleymani N, Ahmadi S, Shiri F, Almasirad A. QSAR and molecular docking studies of isatin and indole derivatives as SARS 3CLpro inhibitors. *BMC Chem.* 2023;17(1):32.
37. Toropova AP, Toropov AA, Benfenati E. Semi-correlations as a tool to model for skin sensitization. *Food Chem Toxicol.* 2021;157:112580.
38. Javidfar M, Ahmadi S. QSAR modelling of larvicidal phytocompounds against *Aedes aegypti* using index of ideality of correlation. *SAR QSAR Environ Res.* 2020;31(10):717–39.
39. Ghasemi JB, Shiri F, Pirhadi S, Heidari Z. Discovery of new potential antimalarial compounds using virtual screening of ZINC database. *Comb Chem High Throughput Screen.* 2015;18(2):227–34.
40. Sunseri J, Koes DR. Pharmit: interactive exploration of chemical space. *Nucleic Acids Res.* 2016;44(W1):W442–8.
41. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model.* 2013;53(8):1893–904.
42. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open babel: an open chemical toolbox. *J Cheminform.* 2011;3:1–14.
43. Hashemizadeh M, Shiri F, Shahraki S, Razmara Z. A multidisciplinary study for investigating the interaction of an iron complex with bovine liver catalase. *Appl Organomet Chem.* 2022;36(11): e6881.
44. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep.* 2017;7(1):42717.
45. Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J Chem Inf Model.* 2015;55(2):460–73.
46. Ntie-Kang F. An in silico evaluation of the ADMET profile of the StreptotomeDB database. *Springerplus.* 2013;2:1–11.
47. Newby D, Freitas AA, Ghafourian T. Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *Eur J Med Chem.* 2015;90:751–65.

## Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.