RESEARCH

Open Access

QSAR and molecular docking studies of isatin and indole derivatives as SARS 3CL^{pro} inhibitors

Niousha Soleymani[1], Shahin Ahmadi[2*], Fereshteh Shiri[3*] and Ali Almasirad[1]

Abstract

The 3C-like protease (3CL^{pro}), known as the main protease of SARS-COV, plays a vital role in the viral replication cycle and is a critical target for the development of SARS inhibitor. Comparative sequence analysis has shown that the 3CL^{pro} of two coronaviruses, SARS-CoV-2 and SARS-CoV, show high structural similarity, and several common features are shared among the substrates of 3CL^{pro} in different coronaviruses. The goal of this study is the development of validated QSAR models by CORAL software and Monte Carlo optimization to predict the inhibitory activity of 81 isatin and indole-based compounds against SARS CoV 3CL^{pro}. The models were built using a newer objective function optimization of this software, known as the index of ideality correlation (IIC), which provides favorable results. The entire set of molecules was randomly divided into four sets including: active training, passive training, calibration and validation sets. The optimal descriptors were selected from the hybrid model by combining SMILES and hydrogen suppressed graph (HSG) based on the objective function. According to the model interpretation results, eight synthesized compounds were extracted and introduced from the ChEMBL database as good SARS CoV 3CL^{pro} inhibitor. Also, the activity of the introduced molecules further was supported by docking studies using 3CL^{pro} of both SARS-COV-1 and SARS-COV-2. Based on the results of ADMET and OPE study, compounds CHEMBL4458417 and CHEMBL4565907 both containing an indole scaffold with the positive values of drug-likeness and the highest drug-score can be introduced as selected leads.

Keywords QSAR, Molecular docking, Isatin derivatives, Indole derivatives, SARS CoV 3CL^{pro} inhibitor, Index of ideality of correlation

Introduction

In the end of February 2003, a novel human coronavirus was detected as the causative agent of the first major pandemic of the twenty-first century, severe acute respiratory syndrome (SARS). The first case of "atypical pneumonia" was declared in China and quickly and unexpectedly spread to 29 countries, especially in Asia and North America, alarming the World Health Organization (WHO). Within several months of the outbreak in 2003, the WHO reported that it had caused 916 deaths out of 8422 cases worldwide (10–15% case fatality rate) [1]. In early 2003, a new human coronavirus known as SARS

*Correspondence:
Shahin Ahmadi
Ahmadi.chemometrics@gmail.com
Fereshteh Shiri
Fereshteh.shiri@gmail.com
[1] Department of Medicinal Chemistry, Faculty of Pharmacy, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran
[2] Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran
[3] Department of Chemistry, University of Zabol, Zabol, Iran

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

coronavirus (SARS CoV) was recognized as the causative agent of SARS [2].

COVID-19 is the active pandemic which was first reported in late 2019 in Wuhan, China. In February 2020, SARS-COV-2 was announced as the causative agent. As of October 24th 2021, 243 million cases and over 4.9 million deaths have been reported. The 3C-like protease ($3CL^{pro}$) enzyme or major protease (M^{pro}), is essential for the process of viral replication and infection, thereby making it an ideal target for antiviral therapy [1]. The coronavirus $3CL^{pro}$ is a cysteine protease consisting of about 300 amino acids and containing three domains. Domains I (amino acids 8 to 99) and II (amino acids 100 to 183) consist of beta barrels that simulate the chymotrypsin and 3C proteinases. The binding site is located between the mentioned domains, and about 16 residues join domains I and II to residues 200 to 300 as the C-terminal domain III. The proteolytic activity of $3CL^{pro}$ has been performed by this third five helices domain [3]. The 3CLpro enzymes show a highly conserved structure among known coronavirus species, and several common characteristics are shared among different coronavirus 3CLpro substrates [4]. Comparative sequence analysis has shown that the 3CLpros of the three coronaviruses of SARS-CoV-2, SARS-CoV, and MERS-CoV are very similar in structure and conservatism [5]. These findings indicate that 3CLpro could be used as a homologous target for the development of anti-coronavirus drugs that can inhibit the proliferation of various coronaviruses [4].

Based on various studies, a combination of nucleoside analogues such as ribavirin can be used for the treatment of SARS along with corticosteroids such as methylprednisolone and hydrocortisone [6–9]. Since the beginning of the COVID-19 pandemic different options for the treatment of this disease have been used including monoclonal antibodies, protease inhibitors, corticosteroids, convalescent plasma and so on. However, the definitive efficacy of these drugs has not been proven.

Previous research has revealed that isatin and its derivatives have a broad range of anti-bacterial and antiviral activities such as anti-HIV [10, 11], anti-rhinovirus [12] and against mycobacterium tuberculosis [13]. The derivatized isatin scaffold may be a good candidate for the SARS CoV $3CL^{pro}$ inhibitor because both proteases (human SARS CoV and rhinovirus) are cysteine proteases and are structurally similar in the active site [14].

In 2005, Chen et al. investigate that N-substituted isatin derivatives with anti-rhinovirus activity may also have anti-SARS activity. Therefore, based on these compounds, they synthesized new isatin derivatives and evaluated their inhibition activities against SARS CoV $3CL^{pro}$. The IC_{50} values showed that the mentioned isatin derivatives could inhibit SARS CoV $3CL^{pro}$ in the low micro molar range (0.95–17.50 μM) [15]. Using the results of the previous study, Zhou et al. designed and synthesized a series of N-substituted 5-carboxamide-isatin compounds and evaluated their activities. They introduced some compounds as SARS CoV $3CL^{pro}$ inhibitors which the most potent compound showed an IC_{50} of 0.37 μM [2]. In 2014 Liu et al. in order to improve the inhibitory activity of isatin derivatives against SARS CoV $3CL^{pro}$, investigated a replacement of the carboxamide group using a series of substituted sulfonamide groups in isatin. Optimization of 5-sulfonyl isatin derivatives led to the discovery of a new compound with the strongest potency ($IC_{50} = 1.04$ μM) [16].

Quantitative structure–activity relationship (QSAR) is one of the critical computational techniques for ligand-based drug design, which can statistically show the correlation between the structural and bioactive properties of compounds [17]. Molecular docking is a computational technique for predicting the optimal interaction of two molecules that creates a binding model, typically a small ligand with a protein receptor [18], most commonly used in drug discovery [19]. CORAL is a new software for developing the reliable and predictive QSAR/QSPR models based on SMILES or quasi-SMILES of materials and Monte Carlo optimization [17, 20].

The main goal of this study is to create the simple and reliable QSAR models by CORAL software to predict the inhibitory activity of 81 isatin and indole-based compounds against SARS CoV $3CL^{pro}$. In addition, the effect of using the index of ideality correlation (IIC) as the objective function for modeling in CORAL software has been investigated [21]. Moreover, the results from Monte Carlo optimization-based QSAR modeling with the further addition of molecular docking studies applied for pharmacologically important endpoints. SMILES notation-based optimal descriptors, defined as molecular fragments, identified as main contributors to the increase/decrease of biological activity, which are used further to search compounds from the ChEMBL database with targeted activity based on computer calculation, are presented. Here, molecular docking was applied as an additional method to validate the calculated activity of proposed compounds as novel SARS CoV $3CL^{pro}$ inhibitors.

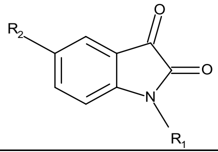
Data and methods

Dataset

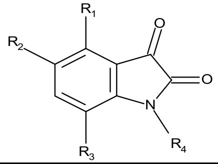
In this study 81 isatin and indole-based SARS 3CLpro inhibitors were gathered from literature [2, 15, 16, 22–25]. The number isatin based compounds were 41 and the rest were indole-based compounds. The IC_{50} (μM) values for inhibitors were converted into their

pIC_{50} (− $logIC_{50}$). Table 1 shows the structure of the molecules along with their pIC_{50} (range between 4.08 and 7.77). BIOVIA Draw 2020 was used to draw the molecular structures of the compounds and convert them into SMILES symbols. The dataset divided the active training (≈25%), passive training (≈20%), calibration (≈20%), and validation (≈35%) sets randomly. To construct the QSAR models based on Monte Carlo

Table 1 Molecular structures of isatin and indole derivatives along with their pIC_{50}



No.	R_1	R_2	IC_{50} (μM)	pIC_{50}	Ref.
1	n-C_4H_9	I	66	4.18	[2]
2	β-$C_{10}H_7CH_2$	I	1.1	5.96	[2]
3	CH_3	$CONH_2$	71	4.15	[2]
4	$CH_3CH_2CH_2$	$CONH_2$	25	4.60	[2]
5	n-C_4H_9	$CONH_2$	19	4.72	[2]
6	$PhCH_2$	$CONH_2$	12.5	4.90	[2]



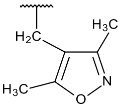
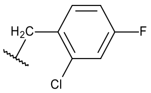
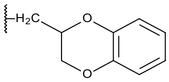
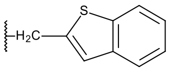
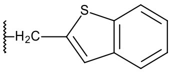
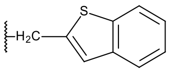
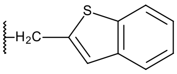
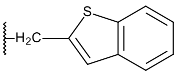
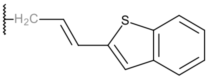
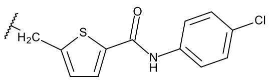
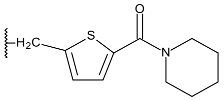
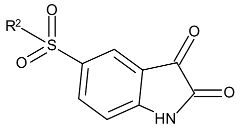
No.	R_1	R_2	R_3	R_4	IC_{50} (μM)	pIC_{50}	Ref.
7	H	CN	H		7.2	5.14	[15]
8	H	I	H		9.4	5.03	[15]
9	H	I	H		13.5	4.87	[15]
10	H	H	H		13.11	4.88	[15]
11	H	H	NO_2		2	5.7	[15]
12	H	H	Br		0.98	6.01	[15]
13	H	F	H		4.82	5.32	[15]
14	Cl	H	H		11.2	4.95	[15]
15	H	I	H		23.5	4.63	[15]

Table 1 (continued)

16	H	I	H		12.57	4.90	[15]
17	H	I	H		17.5	4.76	[15]



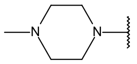
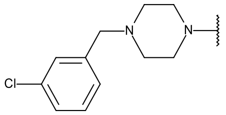
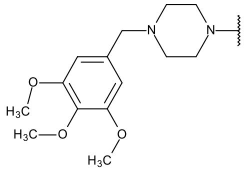
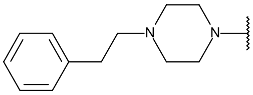
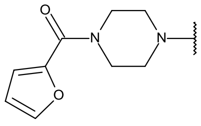
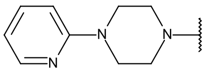
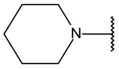
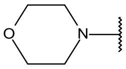
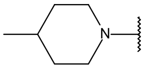
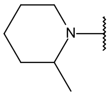
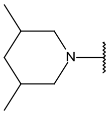
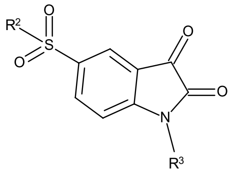
No.	R^2	IC_{50} (μM)	pIC_{50}	Ref.
18		76.74	4.11	[16]
19		31.71	4.5	[16]
20		32.08	4.5	[16]
21		34.91	4.46	[16]
22		10.07	5	[16]
23		51.33	4.3	[16]
24		4.45	5.35	[16]
25		12.66	4.90	[16]
26		1.18	5.93	[16]
27		2.25	5.65	[16]

Table 1 (continued)

28		4.3	5.37	[16]



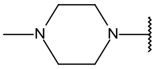
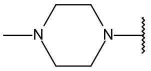
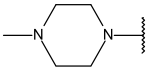
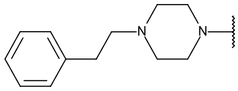
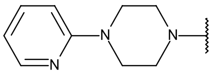
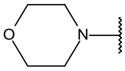
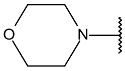
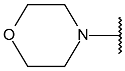
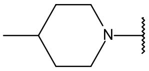
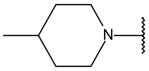
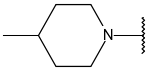
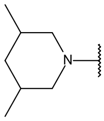
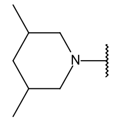
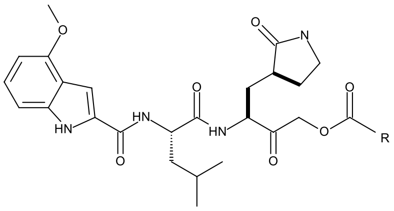
No.	R^2	R^3	IC_{50} (μM)	pIC_{50}	Ref.
29		CH_3	11.83	4.93	[16]
30		$PhCH_2$	67.2	4.17	[16]
31		β-$C_{10}H_7CH_2$	82.91	4.08	[16]
32		β-$C_{10}H_7CH_2$	13.86	4.86	[16]
33		β-$C_{10}H_7CH_2$	5.52	5.26	[16]
34		CH_3	9.91	5	[16]
35		$PhCH_2$	13.86	4.86	[16]
36		β-$C_{10}H_7CH_2$	39.87	4.4	[16]
37		$PhCH_2$	1.04	5.98	[16]
38		β-$C_{10}H_7CH_2$	1.69	5.77	[16]
39		CH_3	17.82	4.75	[16]
40		$PhCH_2$	2.82	5.55	[16]

Table 1 (continued)

41		β-$C_{10}H_7CH_2$	4.7	5.33	[16]



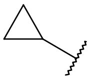
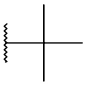
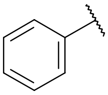
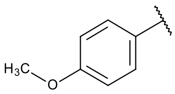
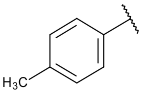
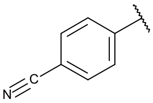
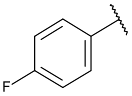
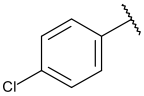
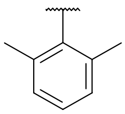
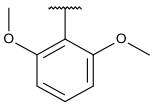
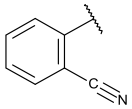
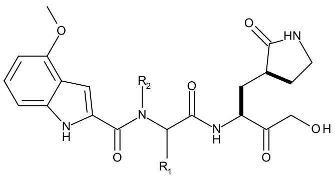
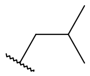
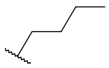
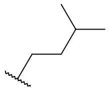
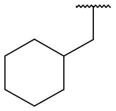
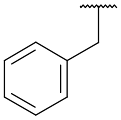
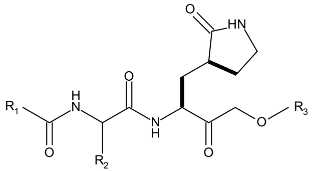
No.	R	IC_{50} (μM)	pIC_{50}	Ref.
42	CH_3	0.22	6.66	[22]
43		0.18	6.74	[22]
44		0.23	6.64	[22]
45		0.09	7.07	[22]
46		0.08	7.1	[22]
47		0.09	7.06	[22]
48		0.05	7.28	[22]
49		0.08	7.09	[22]
50		0.1	7.01	[22]
51		0.07	7.13	[22]

Table 1 (continued)

52		0.21	6.69	[22]
53		0.02	7.77	[22]



No.	R_1	R_2	IC_{50} (μM)	pIC_{50}	Ref.
54		CH_3	0.08	7.08	[22]
55		H	0.02	7.7	[22]
56		H	0.03	7.47	[22]
57		H	0.04	7.36	[22]
58		H	0.10	6.99	[22]



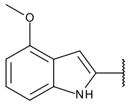
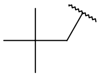
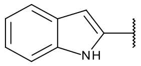
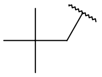
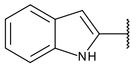
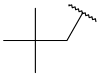
No.	R_1	R_2	R_3	IC_{50} (μM)	pIC_{50}	Ref.
59			CH_3	0.04	7.46	[22]
60			H	0.02	7.7	[22]
61			CH_3	0.11	6.98	[22]

Table 1 (continued)

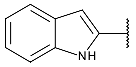
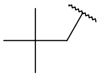
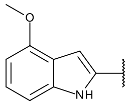
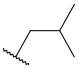
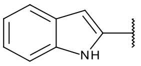
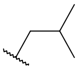
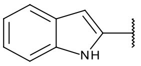
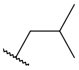
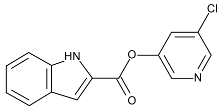
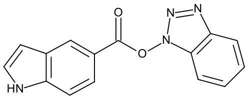
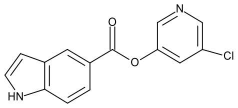
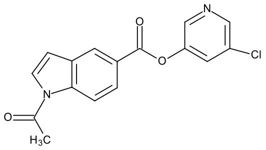
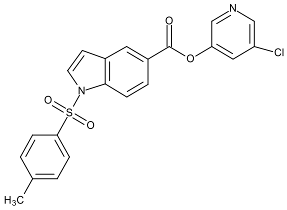
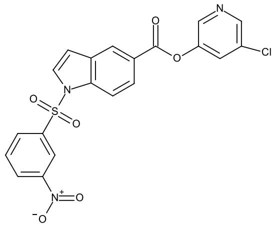
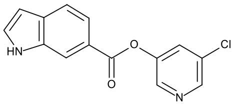
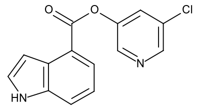
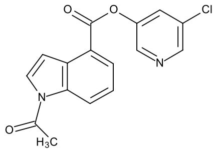
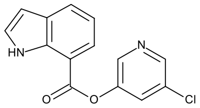
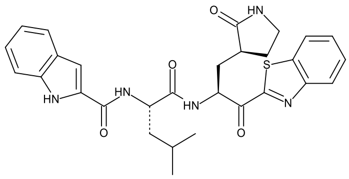
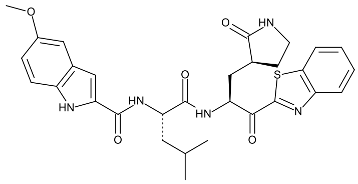
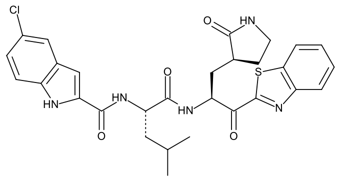
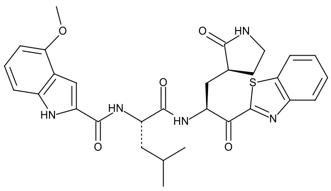
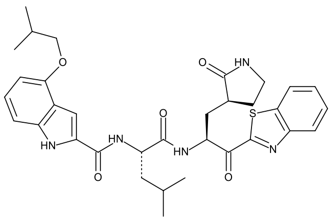
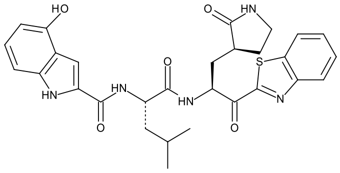
62			C_2H_5	0.11	6.95	[22]
63			CH_3	0.05	7.28	[22]
64			H	0.04	7.42	[22]
65			CH_3	0.13	6.88	[22]
66				0.07	7.19	[23]
67				0.2	6.7	[24]
68				0.31	6.51	[24]
69				0.4	6.4	[24]
70				0.37	6.43	[24]
71				0.09	7.05	[24]
72				0.23	6.64	[24]

Table 1 (continued)

73		0.03	7.52	[24]
74		1.08	5.97	[24]
75		0.08	7.1	[24]
76		1.5	5.82	[25]
77		4.6	5.34	[25]
78		4.8	5.32	[25]
79		0.74	6.13	[25]
80		5.2	5.28	[25]
81		1.5	5.82	[25]

optimization, four separate random partitions were performed.

Descriptors

There are three categories of optimal descriptors in CORAL software, including SMILES-based, graph-based and a combination of SMILES with molecular graph descriptors as hybrid descriptors. The optimal descriptors used in this research to construct the QSAR model are a combination of hydrogen suppression graph (HSG) and SMILES descriptors. The below equation indicates the optimal type of molecular descriptors for QSAR modeling for pIC_{50} of isatin and indole-based compounds as SARS $3CL^{pro}$ inhibitors:

$$\begin{aligned} \mathrm{DCW(T,\ N)} = & \sum \mathrm{CW(S_k)} + \sum \mathrm{CW(SS_k)} \\ & + \sum \mathrm{CW(SSS_k)} + \mathrm{CW(BOND)} \\ & + \mathrm{CW(NOSP)} + \mathrm{CW(HALO)} \\ & + \mathrm{CW(HARD)} + \mathrm{CW(PAIR)} \\ & + \mathrm{CW(Cmax)} + \mathrm{CW(Nmax)} \\ & + \mathrm{CW(Omax)} + \mathrm{CW(Smax)} \\ & + \mathrm{CW(C5)} + \mathrm{CW(C6)} \end{aligned} \tag{1}$$

where, Sk, SSk and SSk are one, two and three-character SMILES features, respectively. BOND represents a global SMILES descriptor that demonstrate the presence/absence of various bonds including double (=), triple (#), and stereochemical (@) bonds. The NOSE indicates the presence/absence of nitrogen, oxygen, sulfur, and phosphorus atoms in the SMILES symbol of molecules. HALO is the presence/absence of halogen in the structure of molecules. HARD is the combination of BOND, NOSP, and HALO in the structure of compounds. Cmax, Nmax, and O max show the maximum number of rings (the range 0–9), the maximum number of nitrogen atoms, and the maximum number of oxygen atoms in the molecular structure, respectively. In addition, C5 and C6 indicate the presence of five- and six-membered rings in the molecular structures, respectively. The CW(x) represents the correlation weight of a SMILES feature or an HSG invariant.

The following equation indicates the correlation between the sum of correlation weights (DCW) of the optimal descriptors and pIC_{50} of the compounds:

$$\mathrm{pIC_{50}} = \mathrm{a} + \mathrm{b} \times \mathrm{DCW(T*,\ N*)} \tag{2}$$

a is the intercept point and b is the slope of the line obtained by the least-squares method. DCW (Descriptors of Correlation Weights) is the sum of correlation weights for the optimal descriptor derived from HSG and SMILES and calculated by Monte Carlo optimization. The T* and N* indicate the optimal threshold value and the number of Monte Carlo optimization cycles, respectively.

A flowchart of a Monte Carlo optimization cycle is presented by Sokolovic et al. [26]. At first cycle, the CW(x) of features is randomly generated and then optimized based on the proposed objective function. There are different objective functions to obtain a reliable QSAR model in CORAL software. TF0, TF1 are two objective functions that we used here to obtain correlation weights for attributes and compare the extracted models based on each of them [27, 28].

$$\mathrm{TF_0} = \mathrm{R_{TRN}} + \mathrm{R_{iTRN}} - |\mathrm{R_{TRN}} - \mathrm{R_{iTRN}}| \times \mathrm{c} \tag{3}$$

$$\mathrm{TF_1} = \mathrm{TF_0} + \mathrm{IIC} \times c' \tag{4}$$

The R_{ATRN} and R_{PTRN} denote the correlation coefficients between the experimental and predicted pIC_{50} for the active training and passive training sets, respectively and, c and c' represent empirical values which are generally constant.

The IIC_{CAL} for calibration (CAL) set is obtained according to the following equation:

$$\mathrm{IIC} = \mathrm{R_{CAL}} \times \frac{\min\left({}^{-}\mathrm{MAE_{CAL,}}\,{}^{+}\mathrm{MAE_{CAL,}}\right)}{\min\left({}^{-}\mathrm{MAE_{CAL,}}\,{}^{+}\mathrm{MAE_{CAL,}}\right)} \tag{5}$$

The R_{CAL} indicates the correlation coefficient for the calibration set. MAE_{CAL} (Mean Absolute Error for calibration set) is calculated based on Eqs 6 to 8:

$$\begin{gathered} {}^{-}\mathrm{MAE_{CAL}} = -\frac{1}{\mathrm{N}} \sum_{\mathrm{K}=1}^{\mathrm{N}} |\Delta_\mathrm{K}| \Delta_\mathrm{K} < 0, \mathrm{N}^{-} \\ \text{is the number of } \Delta\mathrm{k} < 0 \end{gathered} \tag{6}$$

$$\begin{gathered} {}^{+}\mathrm{MAE_{CAL}} = +\frac{1}{\mathrm{N}} \sum_{\mathrm{K}=1}^{\mathrm{N}} |\Delta_\mathrm{K}| \Delta_\mathrm{K} \geq 0, \mathrm{N}^{+} \\ \text{is the number of } \Delta\mathrm{k} \geq 0 \end{gathered} \tag{7}$$

$$\Delta_\mathrm{k} = \mathrm{Exerimental_k} - \mathrm{predicted_k} \tag{8}$$

The 'k' is the index (1, 2... N) and the experimental k and predicted k are related to the pIC_{50}. The CWs for each attribute of Split 1 is provided as an example in Additional file 1: Table S1, total number of attributes is 383.

QSAR model Validation

There are various criteria for evaluating the predictive ability of QSAR models, such as internal validation, external validation, and Y-scrambling. In this study, some standard statistical criteria were used to check the validity of the QSAR models, such as coefficient of determination (R^2), concordance correlation coefficient (CCC), Q^2, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, standard error of estimation (s), mean absolute error (MAE), r^2_m and new Y-scrambling criteria ($C_{R^2_p}$) [29–32]. In addition, the IIC of models was used to improve the predictability of the models [33, 34].

Applicability domain

The range of compounds for which a QSAR model can make reliable predictions is defined based on the applicability domain (AD) of model as the Organization of Economic Co-operation and Development (OECD) principle 3. Here, the AD is calculated based on the distribution of SMILES features in the training and calibration sets and is defined as "Defect_{A_K}"[17].

$$\text{Defect}_{F_K} = \frac{|P_{TRN}(A_K) - P_{CAL}(F_K)|}{N_{TRN}(A_K) + N_{CAL}(F_K)} \tag{9}$$

where $P_{TRN(Fk)}$ and $_{PCAL(Fk)}$ represent the probabilities of kth feature (F_k) in the training and calibration set, respectively; $N_{TRN(Fk)}$ and $N_{CAL(Fk)}$ denote the frequency of kth feature (F_k) in the training and calibration set, respectively.

$$\text{Defect}_{\text{Molecule}} = \sum_{i=1}^{F_K} \text{Defect}_{F_K} \tag{10}$$

According to the SMILES of molecules, the molecule is included in AD if:

$$\text{Defect}_{\text{Molecule}} < 2 \times \overline{\text{Defect}_{\text{TRN}}} \tag{11}$$

where $\overline{\text{Defect}}_{\text{TRN}}$ is the average $\text{Defect}_{\text{molecule}}$ in the training set.

The interpretation of QSAR models

CORAL software provides a simple approach to interpret QSAR models. Three categories of features can be extracted with numerical data of correlation weights in several Monte Carlo optimization cycles: (I) features with a positive correlation weight in all runs that increase the endpoint; (II) features with a negative correlation weight in all runs that decrease the endpoint; and also (III) features with both negative and positive correlation weight in different optimization runs, these features have an undefined role and not be classified as an increasing/decreasing promoters of the endpoint [35].

Molecular docking study

Molecular docking method as a common virtual screening technique can help to find the most favorable ligand binding mode in protein for computer-aided drug discovery [36–38]. The X-ray crystallographic structures of SARS-COV-2 3CL^{pro} were obtained from the Protein Data Bank (PDB: 6XHO) based on a good experimental resolution (1.45 Å), R-value free (0.239), and R-value work (0.211). The native ligand in active site of this protein was ethyl (4R)-4-({N-[(4-methoxy-1H-indol-2-yl)carbonyl]-L-leucyl}amino)-5-[(3S)-2-oxopyrrolidin-3-yl]pentanoate (Query on V34), thus we use this pdb code for molecular docking of indole derivatives. The selected receptor for molecular docking simulation was the x-ray structure of SARS-COV-1 (PDB ID: 1UK4) based on a good experimental resolution (2.5 Å), R-value free (0.231), and R-value work (0.213). The native ligand in active site of this protein was 5-mer peptide. 6XHO and1 UK4 structures consist of a dimer composed of two identical sequences. The side chain A was chosen for molecular docking and the side chain B was removed. The protein structure was prepared using adding hydrogens removing water molecules and native ligands. Then, the Kollmann charges were assigned to the receptor. All compounds were sketched using the by ChemOffice15 (PerkinElmer Inc.), and assigned gasteiger charges and energy optimization of ligands using the steepest descent algorithm carried out by Open Babel [39]. The docking studies were done with the Smina program. Smina is a version of AutoDock Vina with a modified scoring function that is particularly optimized to offer high-throughput scoring (http://smina.sf.net) [40].

The grid parameter file is according to the grid box that comprised 20 × 20 × 20 points with 1 Å space and was centered on the active site of SARS-COV-2 3CL^{pro} ($x=9.412$, $y=1.383$, and $z=8.836$). The grid parameter file is according to the grid box that comprised 14 × 14 × 14 points with 1 Å space and was centered on the active site of SARS-COV-1 ($x=66.036$, $y=3.288$, and $z=5.254$).

The X-ray crystallographic structures of SARS-COV-1, SARS-COV-2 3CL^{pro} were obtained from the Protein Data Bank (PDB: 1UK and 6XHO). The structures of compounds were drawn by BIOVIA Discovery Studio Visualizer 2021. The calculation of energy optimization was done using the steepest descent method. Smina was performed with default settings for three proteins and 9 best conformations of ligand

Table 2 Statistical parameters of QSAR models for prediction of pIC_{50}

Split	Target function	Set	n	R^2	CCC	IIC	Q^2	$Q^2_{F_1}$	$Q^2_{F_2}$	$Q^2_{F_3}$	s	MAE	r^2_m	$C_{R^2_P}$
1	TF0	ATRN	25	0.9992	0.9996	0.9225	0.9990				0.030	0.017		0.9663
		PTRN	20	0.9991	0.9855	0.5687	0.9989				0.195	0.159		0.9761
		CAL	16	0.7308	0.8293	0.6755	0.6265	0.6274	0.6027	0.5712	0.762	0.571	0.6807	0.6995
		VAL	20	0.6200	0.7652	0.5845	0.5353				0.7476	0.5390	0.6173	
	TF1	ATRN	25	0.9419	0.9701	0.6470	0.9330				0.253	0.211		0.9317
		PTRN	20	0.9470	0.9322	0.5838	0.9343				0.414	0.342		0.9174
		CAL	16	0.9229	0.9173	0.9606	0.9015	0.8788	0.8708	0.8605	0.435	0.364	0.5968	0.9043
		VAL	20	0.8804	0.9235	0.8546	0.8603				0.3770	0.3123	0.8545	
2	TF0	ATRN	24	0.9995	0.9997	0.8459	0.9994				0.026	0.018		0.9767
		PTRN	19	0.9995	0.9678	0.9998	0.9994				0.299	0.262		0.9646
		CAL	16	0.6102	0.6710	0.2277	0.5350	0.3128	0.2694	0.4007	0.908	0.656	0.3894	0.5836
		VAL	22	0.7387	0.8535	0.7343	0.6710				0.6126	0.4913	0.7166	
	TF1	ATRN	24	0.9407	0.9694	0.6928	0.9300				0.280	0.225		0.9214
		PTRN	19	0.9405	0.9245	0.2175	0.9192				0.423	0.334		0.9240
		CAL	16	0.9044	0.9487	0.9509	0.8773	0.9090	0.9033	0.9206	0.330	0.260	0.7991	0.8963
		VAL	22	0.8258	0.9055	0.6769	0.7964				0.4563	0.3460	0.7153	
3	TF0	ATRN	23	0.9995	0.9998	0.9167	0.9994				0.024	0.016		0.9861
		PTRN	20	0.9979	0.9847	0.9990	0.9976				0.209	0.193		0.9506
		CAL	16	0.7578	0.8256	0.7414	0.6978	0.5360	0.5276	0.6404	0.705	0.546	0.6655	0.7177
		VAL	22	0.7342	0.7536	0.4182	0.6885				1.0149	0.7729	0.5618	
	TF1	ATRN	23	0.9581	0.9786	0.7529	0.9514				0.221	0.170		0.9430
		PTRN	20	0.9283	0.9419	0.9607	0.9131				0.383	0.316		0.8977
		CAL	16	0.8668	0.9126	0.9310	0.8226	0.7897	0.7858	0.8370	0.475	0.361	0.7721	0.8214
		VAL	22	0.9170	0.9172	0.4932	0.8975				0.5134	0.3963	0.7490	
4	TF0	ATRN	24	0.9992	0.9996	0.5998	0.9990				0.030	0.019		0.9765
		PTRN	21	0.9990	0.9679	0.3260	0.9988				0.265	0.193		0.9816
		CAL	16	0.6548	0.7855	0.5333	0.5865	0.6178	0.5361	0.5700	0.747	0.605	0.5882	0.6127
		VAL	20	0.6223	0.7862	0.4861	0.5550				0.7126	0.4863	0.5345	
	TF1	ATRN	24	0.9580	0.9786	0.8282	0.9494				0.217	0.168		0.9414
		PTRN	21	0.9569	0.9651	0.4967	0.9447				0.292	0.227		0.9414
		CAL	16	0.8786	0.9322	0.9373	0.8404	0.8795	0.8538	0.8644	0.420	0.343	0.8631	0.8476
		VAL	20	0.8090	0.8887	0.3711	0.7787				0.5356	0.3886	0.7850	

were introduced (Additional file 1: Table S4). The computational docking approach was evaluated based on the root-mean-square deviation (RMSD) value from redocking the co-crystalized native ligand back into the active pocket site of the receptor [41].

Results and discussion

QSAR models

To build the reliable QSAR models, two objective functions were used: objective function without IIC (TF0) and with IIC (TF1). The range of finding the optimal threshold value (T) and the number of epochs (N) were 1–3 and 1–15, respectively. The QSAR models to predict the inhibitory activity against SARS $3CL^{pro}$ for four splits were built based on TF1 are given below:

Split 1:

$$pIC50 = 2.4816(\pm 0.0328) + 0.0572(\pm 0.0005) \times DCW(1{,}14) \tag{12}$$

$$R^2_{ATRN} = 0.94,\ n_{TRN} = 25;$$
$$R^2_{PTRN} = 0.95,\ n_{PTRN} = 20;$$
$$R^2_{CAL} = 0.92,\ n_{CAL} = 16;$$
$$R^2_{VAL} = 0.88,\ n_{VAL} = 20$$

Split 2:

$$pIC50 = -0.0804(\pm 0.0679) + 0.0972(\pm 0.0010) \times DCW(1{,}12) \tag{13}$$

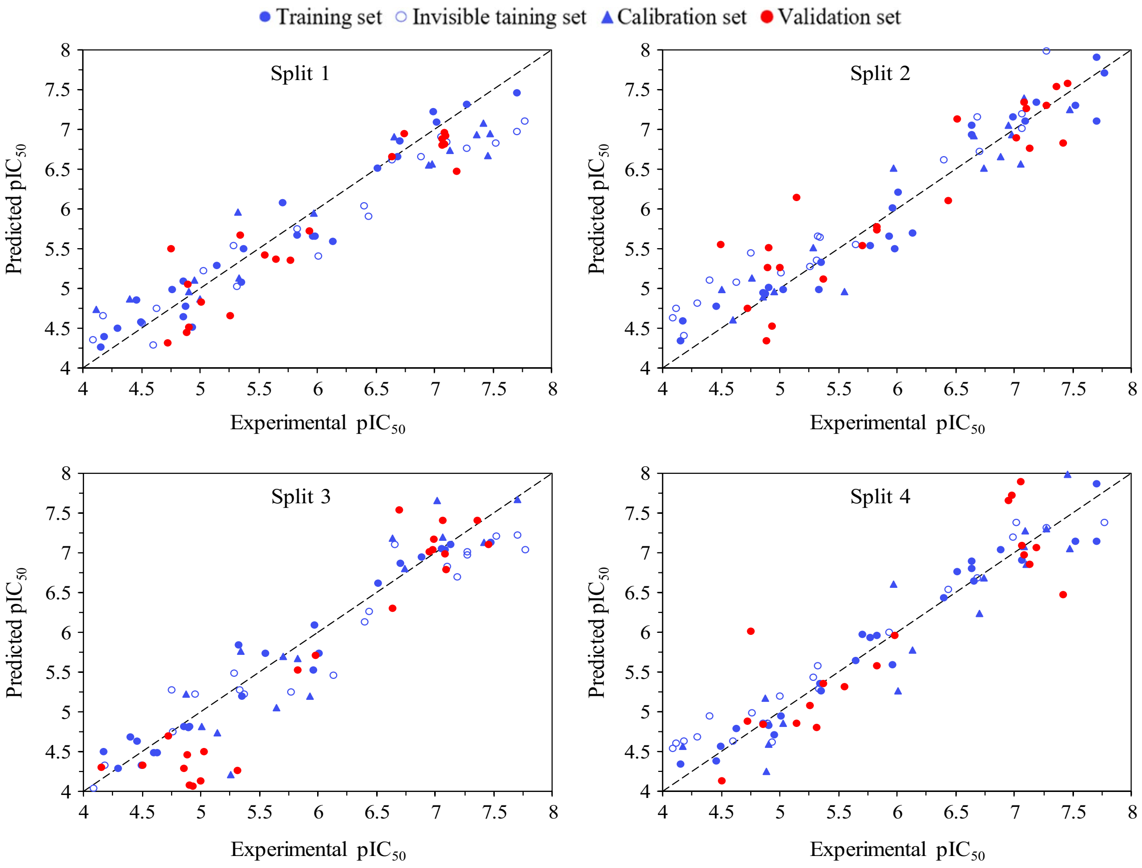


Fig. 1 The graphical representation of the observed versus prediccted pIC_{50} for split 1 to 4

$$R^2_{ATRN} = 0.94,\ n_{ATRN} = 24;$$
$$R^2_{PTRN} = 0.94,\ n_{PTRN} = 19;$$
$$R^2_{CAL} = 0.90,\ n_{CAL} = 16;$$
$$R^2_{VAL} = 0.83,\ n_{VAL} = 22$$

Split 3:

$$pIC50 = -0.1674(\pm 0.0477) + 0.1226(\pm 0.0010) \times DCW(1,6) \qquad (14)$$

$$R^2_{ATRN} = 0.96,\ n_{ATRN} = 23;$$
$$R^2_{PTRN} = 0.93,\ n_{PTRN} = 20;$$
$$R^2_{CAL} = 0.87,\ n_{CAL} = 16;$$
$$R^2_{VAL} = 0.92,\ n_{VAL} = 22$$

Split 4:

$$pIC50 = 0.3203(\pm 0.0545) + 0.1004(\pm 0.0011) \times DCW(1,10) \qquad (15)$$

$$R^2_{ATRN} = 0.96,\ n_{ATRN} = 24;$$
$$R^2_{PTRN} = 0.96,\ n_{PTRN} = 21;$$
$$R^2_{CAL} = 0.88,\ n_{CAL} = 16;$$
$$R^2_{VAL} = 0.81,\ n_{VAL} = 20$$

where R^2_{ATRN}, R^2_{PTRN} R^2_{CAL}, and R^2_{VAL} are coefficient of determination for active training, passive training, calibration, and validation set, respectively. n_{ATRN}, n_{PTRN}, n_{CAL}, and n_{VAL} indicate the number of molecules in the training, calibration, and validation set, respectively.

Table 2 indicates the statistical criteria of QSAR models for predicting of pIC_{50} isatin and indole derivatives based on TF0 and TF1 for each split. Regarding the QSAR models, the models developed based on IIC (TF1) are more predictive than the models developed using TF1. Therefore, it can be stated that the QSAR models built with the modified objective function TF1 using IIC are more reliable and robust than the models built by the objective function TF0. Thus, the QSAR model built for split 3 with TF1 was selected as the best

Table 3 The list of structural attributes increases or decrease the pIC_{50} of isatin and indole derivatives based on the Split 3 model for three independent probes

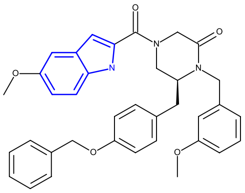
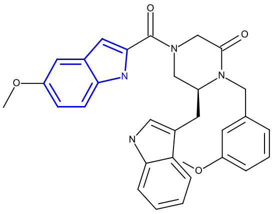
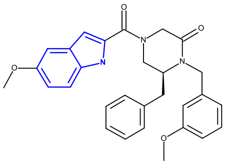
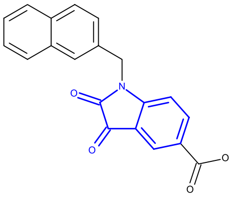
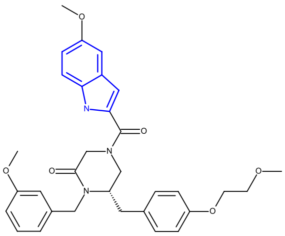
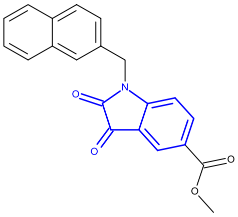
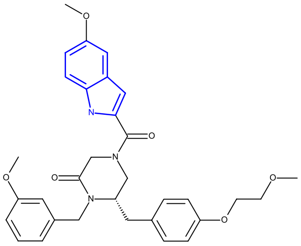
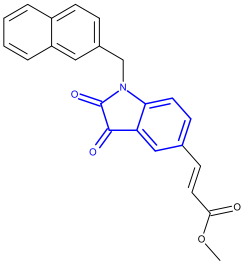
SA_K	Cws Probe 1	CWs Probe 2	CWs Probe 3	NSs	NSc	NSv	Defect $[SA_k]$	Comments
++++N–-B2==	0.81977	1.44792	3.3629	23	20	16	0	Presence of nitrogen with double bond
++++N–-O===	0.55248	1.60119	0.9952	23	20	16	0	Presence of nitrogen with oxygen
++++O–-B2==	3.11773	2.3733	4.25607	23	20	16	0	Presence of oxygen with double bond
1……….	2.69283	0.10613	3.13178	23	20	16	0	Presence of at least one ring
O…(……	0.4556	0.00742	0.17344	23	20	16	0	Combination of aliphatic oxygen with branching
O…=……	0.19142	0.11911	0.49273	23	20	16	0	Combination of aliphatic oxygen with double bond
3……….	0.32927	0.44463	2.31771	22	19	16	0.0011	Presence of at least three rings
=…(……	0.1846	0.14941	0.0791	22	20	16	0.0011	Combination of double bound with branching
O…=…(…	0.15551	0.31959	0.3984	22	20	16	0.0011	Presence of oxygen with double bond and branching
c…2……	1.17117	1.05998	0.02425	22	19	16	0.0011	Presence of aromatic carbon in second ring
c…c…2…	0.18057	0.67353	0.16569	22	19	16	0.0011	Presence of two consecutive aromatic carbon in second ring
N…(……	0.3027	0.16119	0.46805	19	15	14	0.0015	Combination of nitrogen with branching
BOND10000000	0.33172	0.06765	0.32395	18	12	7	0.0138	Presence of double bounds
C…(…=…	1.0509	0.49087	0.32091	16	20	15	0.0078	Presence of aliphatic carbon with branching and double bond
c…1……	0.04992	2.36042	0.34589	14	14	12	0.0054	Presence of aromatic carbon in first ring
++++N–-S===	– 0.60102	– 0.03774	– 0.7567	13	10	8	0.0031	Presence of nitrogen with sulfur
N…1……	– 1.44414	– 2.57584	– 0.89628	7	2	4	0.0082	Presence of aliphatic nitrogen in firth ring
C…N…(…	– 0.13702	– 1.66665	– 1.60758	8	8	6	0.0054	Presence of consecutive aliphatic carbon with aliphatic nitrogen with branching
4…c…(…	– 0.85897	– 0.24461	– 0.27774	6	6	5	0.0047	Presence of aromatic carbon with branching in fourth ring
N…3…C…	– 0.85012	– 0.27675	– 0.27575	4	0	1	0.0223	Presence of aliphatic nitrogen and carbon in third ring
C…(…4…	– 0.7243	– 0.07997	– 0.11114	3	5	0	1	Presence of aliphatic carbon with branching in fourth ring
C5…AH.2…	– 0.36619	– 0.87409	– 0.56113	3	4	4	0.0171	Presence of two five-member rings with aromaticity and heteroatoms
s…4……	– 0.58556	– 0.70714	– 0.26015	3	3	3	0.0095	Presence of aromatic sulphur in the fourth ring
++++I–-N===	– 0.95359	– 1.09861	– 0.30426	2	2	1	0.0082	Presence of iodine with nitrogen
C…c…2…	– 0.84725	– 0.3365	– 0.29322	2	0	1	0.0082	Prsence of consecutive aliphatic carbon with aromatic carbon in second ring
[…C…@…	– 0.26604	– 1.1151	– 0.15705	2	3	3	0.0201	Presence of aliphatic carbon with stereo-chemical (3D) bond

model because the coefficient of determination (R^2) was the highest for the validation set of this model.

Y-randomization test (Y-test) was done by CORAL software to confirm the non-chance correlation of developed QSAR models. After ten repetitions of new random models were developed and the values of average value of R^2 were found below 0.1 (see Additional file 1: Table S2). These values confirm that the correlation between pIC_{50} and molecular attributes is not based on chance correlation. Moreover, for the Y-randomization test, the value of CR^2p for all models was more than 0.8 (Table 2).

Additional file 1: Table S3 shows the SMILES symbol of isatin and indole derivatives, the set of each compound, the observed and calculated pIC_{50} of four models, and AD in four splits using TF1. The average $\overline{\text{Defect}}_{\text{TRN}}$ for Split 1 to 4 of constructed models based of TF0 are 5.91, 3.19, 5.18, and 5.05, respectively. So, compounds fall into AD if DefectSMILES < 11.82, 6.38, 10.36, and 10.10, for split 1 to 4 respectively. The percentages of data set in the AD of models were 82, 82,

Table 4 The average predicted pIC_{50}, IC_{50}, affinity, based on four models for eight extracted compounds from CHEMBL data search

Structure	pIC_{50}	IC_{50} (μM)	Affinity (Kcal mol^{-1})	PDB	Structure	pIC_{50}	IC_{50} (μM)	Affinity (Kcal mol^{-1})	PDB
 CHEMBL4524939 Indole scaffold	7.99	0.010	-9.7	6XHO	 CHEMBL4443007 Indole scaffold	8.30	0.005	-10.1	6XHO
 CHEMBL4458417 Indole scaffold	7.36	0.043	-9.6	6XHO	 CHEMBL383761 Isatin scaffold	7.69	0.020	-9.1	1UK4
 CHEMBL4452760 Indole scaffold	8.11	0.008	-10.1	6XHO	 CHEMBL210543 Isatin scaffold	7.99	0.010	-9.4	1UK4
 CHEMBL4565907 Indole scaffold	7.6	0.025	-9.7	6XHO	 CHEMBL3103276 Isatin scaffold	7.89	0.013	-9.4	1UK4

83, and 88 for splits 1–4, respectively. This revealed that the four prediction models were capable of predicting more than 80% of the new data (Additional file 1: Table S3).

Figure 1 displays the plots of the calculated versus observed pIC_{50} of SARS $3CL^{pro}$ inhibitors for four models developed based on TF1. It also shows that there is good agreement between the observed and experimental pIC_{50}.

Mechanistic interpretation

Mechanistic interpretation as the fifth OECD principle of QSAR modeling states that the molecular features responsible for increased or decreased activity should be investigated whenever possible. The interpretation of the model can help to design and identify new isatin- and indole-based derivatives. The list of structural features extracted from the best QSAR model (split 3) for three independent probes is shown in Table 3. A short description of these descriptors is presented in the comments

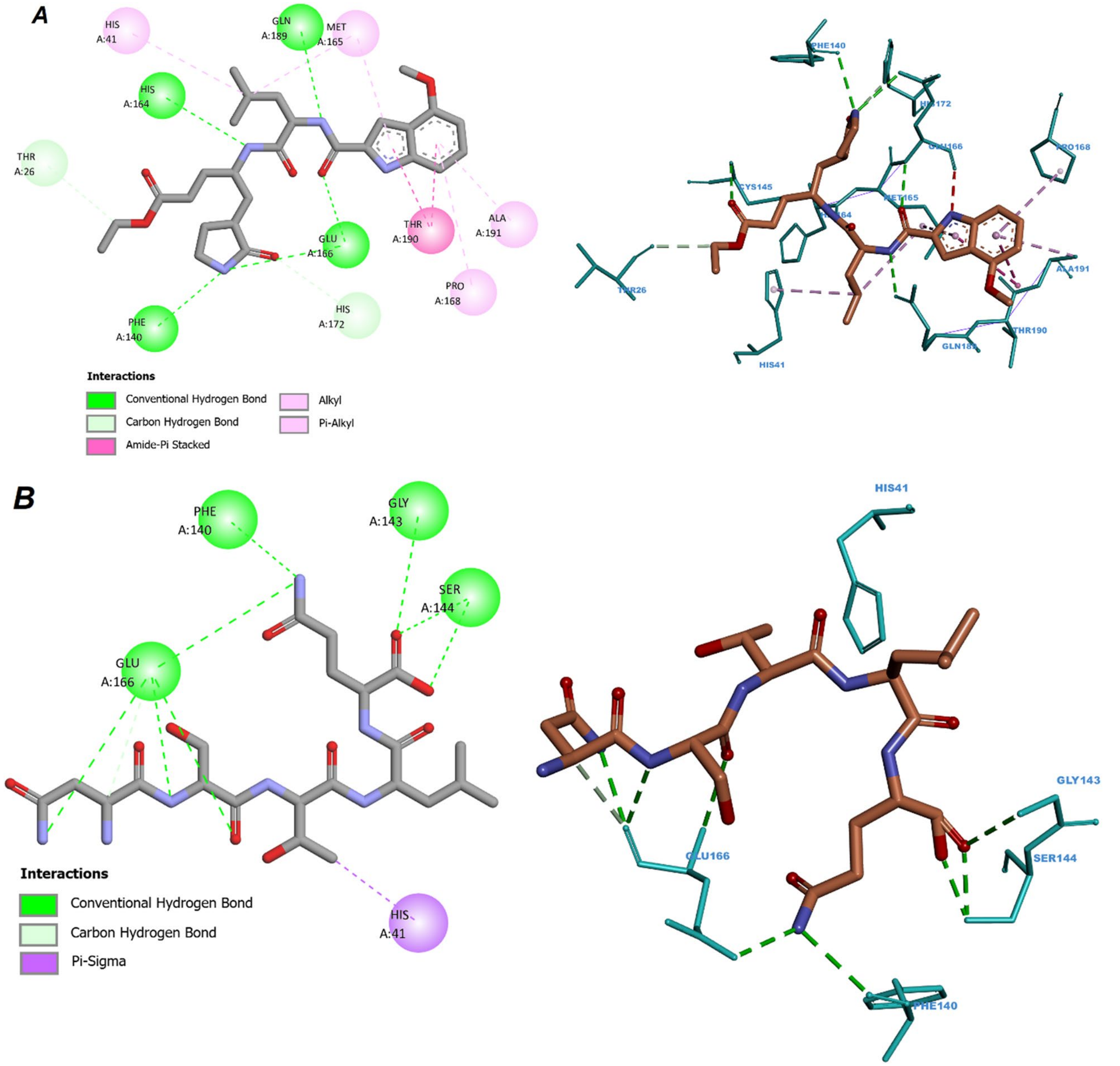


Fig. 2 V34 interaction patterns with active residues in the SARS-COV-2 3CLpro pocket (A), 5-mer peptide interaction patterns with active residues in the SARS-COV-1 pocket (B)

column of Table 3 which shows the structural features of increasing or decreasing pIC_{50} of isatin and indole derivatives. The identified promotors in the increase of pIC_{50} include the presence of nitrogen with double bond, presence of nitrogen with oxygen, presence of oxygen with double bond, presence of at least one ring, combination of aliphatic oxygen with double bond, presence of oxygen with double bond and branching and presence of aromatic carbon in first ring. The promoters of decrease of SARS $3CL^{pro}$ inhibitory activity of isatin and indole derivatives are the presence of nitrogen with sulfur, presence of consecutive aliphatic carbon with aliphatic nitrogen with branching, presence of aromatic carbon with branching in fourth ring and presence of aliphatic carbon with branching in fourth ring.

Based on the favorable structural features and using the most active molecules among the 81 inhibitors which were gathered from literature, some compounds synthesized in various studies were extracted from ChEMBL database. In the ChEMBL database, newly synthesized

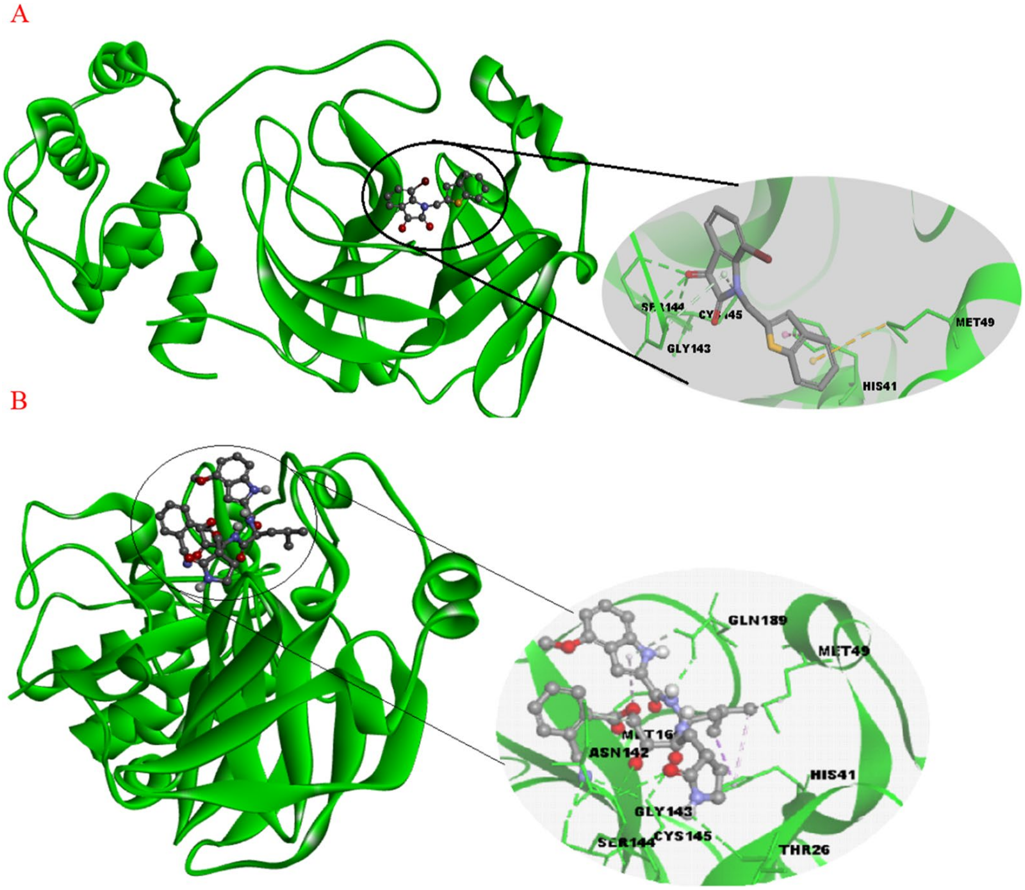


Fig. 3 Three-dimensional diagram of compound 12(**A**) and 53(**B**) into the binding pocket of SARS-COV-1 3CL^{p}

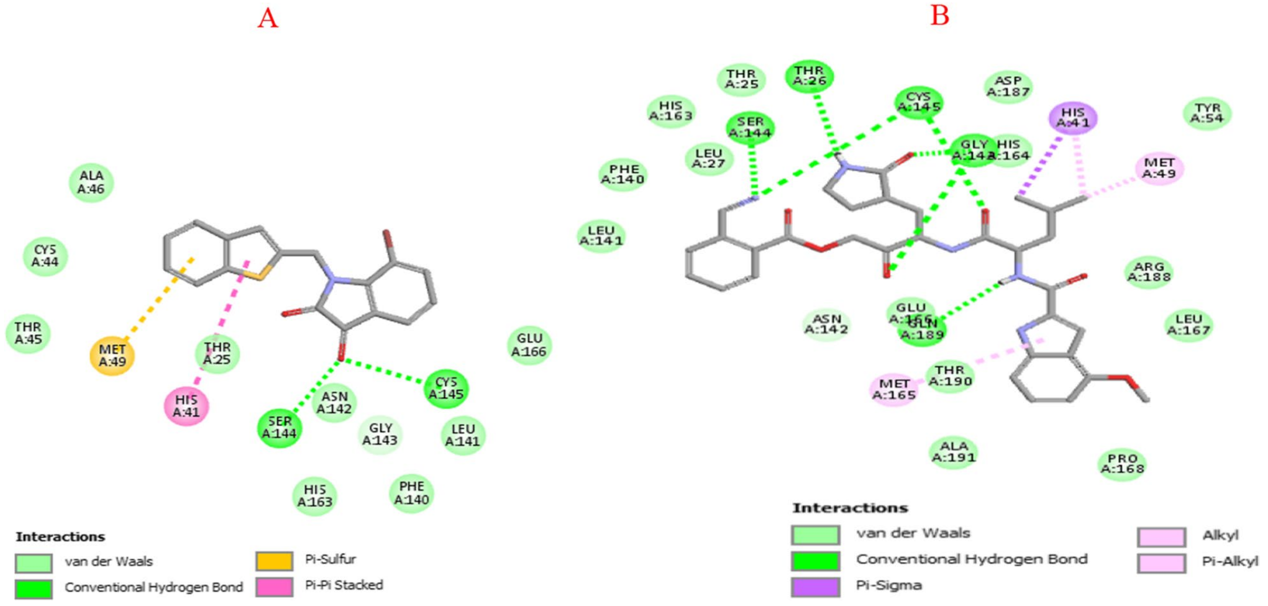


Fig. 4 Two-dimensional diagram of compound 12 (**A**) and 53 (**B**) interactions with binding site residues of SARS-COV-1 3CL^{pro}

compounds can be extracted with percentage similarity with desired compound, so we entered the ligand with the highest activity into ChEMBL and extracted some similar compounds from this database. The inhibitory activity (pIC_{50}) of selected structures was calculated using best QSAR model (Split 3). Finally, eight most active compounds (isatin and indole scaffolds with most pIC_{50}) were selected and introduced which are listed in Table 4. The predicted pIC_{50} range for the extracted compounds based on average prediction of four models was between 7.35 and 8.30. The AD analysis of these compounds based on the Split 3 model (the best model) shows that they fall into AD except for CHEMBL3103276.

Molecular docking analysis

First, we perform a re-docking of the V34 ligand with the SARS-COV-2 $3CL^{pro}$ and 5-mer peptide with SARS-COV-1 receptors; this is done to validate the molecular docking protocol and also to get insight into the reference active amino acid residues involved in interactions inside the SARS-COV-2 $3CL^{pro}$ and SARS-COV-1 protein pocket (PDB code: 6XHO and 1UK4). Figure 2 displays 3D and 2D visualizations of the re-docking pathways of V34 inside the COVID-2 $3CL^{pro}$ and 5-mer peptide inside the SARS-COV-1 protein pockets with − 8.07 and − 9.4 kcal/mol, respectively. Figures indicate that the re-dock V34 located in the active site of SARS-COV-2 $3CL^{pro}$ interacts with the THR26, HIS41, PHE140, CYS145, HIS164, MET165, GLU166, PRO168, HIS172, GLN189, THR190, and ALA191. Also, the re-dock 5-mer peptide located in the active site of SARS-COV-1 interacts with the HIS41, PHE140, GLY143, SER144, and GLU166. These interactions were hydrophobic and hydrogen bonds. The root-mean-square deviation (RMSD) values were 0.14 and 1.1 Å for native and re-docked ligands of V34 and 5-mer peptide, respectively; which are lower than the tolerable marginal value of 2 Å (Additional file 1: Fig. S1).

Figure 3a and b shows that the compound 12 and 53 were placed into the binding pocket of SARS-COV-1 $3CL^{pro}$ by representing three-dimensional diagram. Two-dimensional diagram of compound 12 and 53 interactions was presented in Fig. 4a and b the compounds formed some important interactions with binding site residues of SARS-COV-1 3CLpro. As the molecular docking results are shown in Fig. 3a, the compound 12 formed two hydrogen bond interactions with SER144 and CYS145 the binding site of SARS-COV-1 $3CL^{pro}$. Also, it has two hydrophobic interactions with HIS41 and MET49. Moreover, ALA46, CYS44, THR45, THR25, ASN142, GLY143, HIS163, PHE140, LEU141 and GLU166 have van der Walls interaction with the protein. Figure 3b shows various interactions of compound 57 with HIS41, MET49 and MET165, along with some hydrophobic interactions. In addition, the complex formed hydrogen bond interactions with residues SER144, THR26, CYS145, GLY143 and GLN189. LEU141, PHE140, HIS163, LEU27, THR25, ASN142, GLU166, THR190, ALA191, TYR54, ARG188, LEU167 and PRO168 had van der Walls interaction with the protein.

Comparing the molecular docking results of re-docked native ligands and compounds 12 and 53 as the most activist compounds; we can notice that all compounds 12 and 53 interacted with the majority of active residues in the COV-2 3CLpro and SARS-COV-1 pockets with which native ligands interacted.

Molecular docking results agree with some promoters regarding the increase in pIC_{50} in QSAR models; for instance, compounds 12 and 53 contain oxygen with double bonds, at least one ring, and branching, all of which interact with amino acids residues in protein active sites via hydrogen bonds and hydrophobic interactions.

Hexachlorophene was used as a SARS $3CL^{pro}$ standard inhibitor ($IC_{50} = 5$ μM) according to Liu et al. [42]. We docked Hexachlorophene into the active site of 6XHO. The best binding mode of the Hexachlorophene in the

Table 5 ADMET prediction for eight extracted compounds from CHEMBL

Compound	Human intestinal absorption	ClogP	Ames test	Acute oral toxicity	Drug likeness	Drug score
CHEMBL4524939	+ (0.9816)	5.36	No	III	− 2.47	0.16
CHEMBL4458417	+ (0.9816)	4.02	No	III	6.12	0.56
CHEMBL4452760	+ (0.9792)	3.85	No	III	1.76	0.46
CHEMBL4565907	+ (0.9774)	4.00	No	III	2.59	0.51
CHEMBL4443007	+ (0.9816)	4.05	Yes	III	6.75	0.48
CHEMBL383761	+ (0.9670)	2.67	No	III	− 3.73	0.34
CHEMBL210543	+ (0.9914)	3.10	No	III	− 4.66	0.32
CHEMBL3103276	+ (0.9761)	3.43	No	III	− 5.60	0.17

binding site of SARS-COV-1 $3CL^{pro}$ (pdb: 6XHO) was − 8.05 kcal/mol.

Eight extracted compounds from CHEMBL based on scaffold of isatin or indole were docked into 1UK4 and 6XHO as well. Two and three-dimensional diagrams of the interaction of the eight ligands from CHEMBLE with their receptors are presented in Additional file 1: Fig. S2. Molecular docking analysis shows that these ligands with the majority of active residues in the COV-2 3CLpro and SARS-COV-1 pockets with which native ligands interacted. As before we mentioned it for the activist compounds 12 and 53. It confirmed that indole and isatin are important cores in interaction with targets. As can be seen in Table 4, all eight compounds had higher binding energy compared to the most active compounds in data set and hexachlorophene. The results present a very good correlation between results obtained from Monte Carlo optimization modeling and molecular docking studies.

ADMET results

In silico ADMET (absorption, distribution, metabolism, excretion, and toxicity) screening of compounds can reduce the cost and time associated with the in vitro assay and/or in vivo experiments [43]. AdmetSAR online database was used to predict ADMET properties of extracted isatin- and indole-based compounds [44]. As ADMET properties are shown in Table 5, all eight compounds showed positive results for human intestinal absorption. Furthermore, it is necessary to check whether the proposed molecules are non-toxic because it plays an important role in the selection of drugs. Ames test was negative for all compounds except CHEMBL4443007 and based on acute oral toxicity all compounds were classified as non-toxic.

The Osiris Property Explorer (OPE) tool was used to assess the fragment-based drug-likeness of the extracted compounds [45, 46]. A positive value (0.1–10) indicates that the compound mainly contains fragments that are often found in commercial drugs. Also, using this program, the overall drug scores were evaluated that combines drug-likeness, ClogP, ClogS, molecular weight, and toxicity risk factors in one single value where the frequency of occurrence of each fragment is determined within the collection of approved drugs and within Fluka non-medicinal chemicals.

Finally, based on the results of the OPE study, compounds CHEMBL4458417 and CHEMBL4565907 both containing an indole scaffold with the positive values of drug-likeness and the highest drug-score can be introduced as selected leads.

Conclusion

Four simple, predictive, and reliable QSAR models were developed for the pIC_{50} values of 81 isatin and indole derivatives that inhibit SARS $3CL^{pro}$ using Monte Carlo with the index of ideality of correlation (IIC) as the objective function. The statistical parameters of the models were suitable with high predictive power ($R^2_{Val} = 0.81$–0.92, and MAE = 0.31–0.40). The four proposed models were satisfactory for predicting new isatin and indole derivatives as candidates for SARS $3CL^{pro}$ inhibitors and can be used for pre-synthesis evaluation of new isatin and indole derivatives. A mechanistic interpretation of the models was done by examining the correlation weights of the different extracted molecular features extracted in several Monte Carlo optimization runs. These features were used to extract eight new and more active isatin and indole derivatives from the ChEMBL database. The activity of new compounds was further verified by molecular docking studies. The activity of the new compounds was further confirmed by molecular docking studies. The binding energy of these molecules with residues of active site were in correlation with calculated pIC_{50}. Finally, the compounds CHEMBL4458417 and CHEMBL4565907 both containing an indole scaffold with the positive values of drug-likeness and the highest drug-score were introduced as selected leads.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13065-023-00947-w.

Additional file1: Table S1. CWs for each attribute of Split 1. **Table S2.** The results of Y-randomization test for all splits constructed based on TF1. **Table S3.** SMILES notations of isatin and indole derivatives, the compound set, their experimental, predicted pIC_{50}, and applicability domain in four splits using TF1. **Table S4.** The affinity of nine conformations docked into SARS-COV-1 3CLpro (PDB: 1UK4 and 6XHO) for compounds 12 and 53. **Figure S1**. 3D superposition of original (black) and re-docked (yellow) (A) V34 ligand in the 6XHO (RMSD=0.14 Å), (A) 5-mer peptide ligand in the 1UK4 (RMSD=1.1 Å). **Figure S2.** Two and three-dimensional diagram of (A) CHEMBL4524939 (B) CHEMBL4458417 (C) CHEMBL4452760 (D) CHEMBL4565907 (E) CHEMBL4443007 interactions with binding site residues of SARS-COV-1 3CLpro (6XHO) and (F) CHEMBL383761 (G) CHEMBL210543 (H) CHEMBL3103276interactions with binding site residues of SARS-COV-1 (1UK4).

Acknowledgements

The authors would like to express their deepest gratitude to Dr. Alla P. Toropova and Dr. Andrey A. Toropov for providing the CORAL software.

Author contributions

S.A. conceived of the presented idea. N.S. performed drawing of structures and developed the models and wrote original draft. S.A. performed model validation and interpretation. F.S. contributed to molecular docking. A.A. performed in silico ADMET screening of compounds. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 November 2022 Accepted: 31 March 2023
Published online: 07 April 2023

References

1. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung SH. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: peptidomimetics and small molecule chemotherapy. J Med Chem. 2016;59(14):6595–628.
2. Zhou L, Liu Y, Zhang W, Wei P, Huang C, Pei J, Yuan Y, Lai L. Isatin compounds as noncovalent SARS coronavirus 3C-like protease inhibitors. J Med Chem. 2006;49(12):3440–3.
3. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science. 2003;300(5626):1763–7.
4. Liu Y, Liang C, Xin L, Ren X, Tian L, Ju X, Li H, Wang Y, Zhao Q, Liu H, et al. The development of Coronavirus 3C-Like protease (3CL(pro)) inhibitors from 2010 to 2020. Eur J Med Chem. 2020;206: 112711.
5. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.
6. Chu CK, Gadthula S, Chen X, Choo H, Olgen S, Barnard DL, Sidwell RW. Antiviral activity of nucleoside analogues against SARS-coronavirus (SARS-CoV). Antiviral Chem Chemother. 2006;17(5):285–9.
7. Morgenstern B, Michaelis M, Baer PC, Doerr HW, Cinatl J Jr. Ribavirin and interferon-beta synergistically inhibit SARS-associated coronavirus replication in animal and human cell lines. Biochem Biophys Res Commun. 2005;326(4):905–8.
8. Koren G, King S, Knowles S, Phillips E. Ribavirin in the treatment of SARS: a new trick for an old drug? CMAJ. 2003;168(10):1289–92.
9. Tai DY. Pharmacologic treatment of SARS: current knowledge and recommendations. Ann Acad Med Singap. 2007;36(6):438.
10. Corona A, Meleddu R, Esposito F, Distinto S, Bianco G, Masaoka T, Maccioni E, Menendez-Arias L, Alcaro S, Le Grice SF, et al. Ribonuclease H/DNA polymerase HIV-1 reverse transcriptase dual inhibitor: mechanistic studies on the allosteric mode of action of isatin-based compound RMNC6. PLoS ONE. 2016;11(1): e0147225.
11. Meleddu R, Distinto S, Corona A, Tramontano E, Bianco G, Melis C, Cottiglia F, Maccioni E. Isatin thiazoline hybrids as dual inhibitors of HIV-1 reverse transcriptase. J Enzyme Inhib Med Chem. 2017;32(1):130–6.
12. Webber SE, Tikhe J, Worland ST, Fuhrman SA, Hendrickson TF, Matthews DA, Love RA, Patick AK, Meador JW, Ferre RA. Design, synthesis, and evaluation of nonpeptidic inhibitors of human rhinovirus 3C protease. J Med Chem. 1996;39(26):5072–82.
13. Gao F, Yang H, Lu T, Chen Z, Ma L, Xu Z, Schaffer P, Lu G. Design, synthesis and anti-mycobacterial activity evaluation of benzofuran-isatin hybrids. Eur J Med Chem. 2018;159:277–81.
14. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol. 2003;331(5):991–1004.
15. Chen L-R, Wang Y-C, Lin YW, Chou S-Y, Chen S-F, Liu LT, Wu Y-T, Kuo C-J, Chen TS-S, Juang S-H. Synthesis and evaluation of isatin derivatives as effective SARS coronavirus 3CL protease inhibitors. Bioorg Med Chem Lett. 2005;15(12):3058–62.
16. Liu W, Zhu H-M, Niu G-J, Shi E-Z, Chen J, Sun B, Chen W-Q, Zhou H-G, Yang C. Synthesis, modification and docking studies of 5-sulfonyl isatin derivatives as SARS-CoV 3C-like protease inhibitors. Bioorg Med Chem. 2014;22(1):292–302.
17. 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.
18. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. Molecules. 2015;20(7):13384–421.
19. Lin X, Li X, Lin X. A review on applications of computational methods in drug screening and design. Molecules. 2020. https://doi.org/10.3390/molecules25061375.
20. Hamzehali H, Lotfi S, Ahmadi S, Kumar P. Quantitative structure–activity relationship modeling for predication of inhibition potencies of imatinib derivatives using SMILES attributes. Sci Rep. 2022;12(1):21708.
21. 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.
22. Hoffman RL, Kania RS, Brothers MA, Davies JF, Ferre RA, Gajiwala KS, He M, Hogan RJ, Kozminski K, Li LY. Discovery of ketone-based covalent inhibitors of coronavirus 3CL proteases for the potential therapeutic treatment of COVID-19. J Med Chem. 2020;63(21):12725–47.
23. Zhang J, Pettersson HI, Huitema C, Niu C, Yin J, James MN, Eltis LD, Vederas JC. Design, synthesis, and evaluation of inhibitors for severe acute respiratory syndrome 3C-like protease based on phthalhydrazide ketones or heteroaromatic esters. J Med Chem. 2007;50(8):1850–64.
24. Ghosh AK, Gong G, Grum-Tokars V, Mulhearn DC, Baker SC, Coughlin M, Prabhakar BS, Sleeman K, Johnson ME, Mesecar AD. Design, synthesis and antiviral efficacy of a series of potent chloropyridyl ester-derived SARS-CoV 3CLpro inhibitors. Bioorg Med Chem Lett. 2008;18(20):5684–8.
25. Thanigaimalai P, Konno S, Yamamoto T, Koiwai Y, Taguchi A, Takayama K, Yakushiji F, Akaji K, Chen S-E, Naser-Tavakolian A. Development of potent dipeptide-type SARS-CoV 3CL protease inhibitors with novel P3 scaffolds: design, synthesis, biological evaluation, and docking studies. Eur J Med Chem. 2013;68:372–84.
26. Soković D, Stanković V, Toskić D, Lilić L, Ranković G, Ranković J, Nedin-Ranković G, Veselinović AM. Monte Carlo-based QSAR modeling of dimeric pyridinium compounds and drug design of new potent acetylcholine esterase inhibitors for potential therapy of myasthenia gravis. Struct Chem. 2016;27:1511–9.
27. Ahmadi S. Mathematical modeling of cytotoxicity of metal oxide nanoparticles using the index of ideality correlation criteria. Chemosphere. 2020;242: 125192.
28. Toropova AP, Toropov AA. Use of the index of ideality of correlation to improve models of eco-toxicity. Environ Sci Pollut Res. 2018;25(31):31771–5.
29. Aher R, Roy K. Exploring the structural requirements in multiple chemical scaffolds for the selective inhibition of Plasmodium falciparum calcium-dependent protein kinase-1 (Pf CDPK-1) by 3D-pharmacophore modelling, and docking studies. SAR QSAR Environ Res. 2017;28(5):390–414.
30. Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair RM, Branham WS, Dial SL, Moland CL, Sheehan DM. QSAR models using a large diverse set of estrogens. J Chem Inf Comput Sci. 2001;41(1):186–95.
31. Schuurmann G, Ebert R-U, Chen J, Wang B, Kuhne R. External validation and prediction employing the predictive squared correlation coefficient Test set activity mean vs training set activity mean. J Chem Inf Model. 2008;48(11):2140–5.
32. Daoui O, Elkhattabi S, Chtita S, Elkhalabi R, Zgou H, Benjelloun AT. QSAR, molecular docking and ADMET properties in silico studies of novel 4, 5, 6, 7-tetrahydrobenzo [D]-thiazol-2-Yl derivatives derived from dimedone as potent anti-tumor agents through inhibition of C-Met receptor tyrosine kinase. Heliyon. 2021;7(7): e07463.
33. Kumar P, Kumar A. CORAL: QSAR models of CB1 cannabinoid receptor inhibitors based on local and global SMILES attributes with the index of

ideality of correlation and the correlation contradiction index. Chemom Intell Lab Syst. 2020;200: 103982.

34. Toropov AA, Toropova AP, Marzo M, Benfenati E. Use of the index of ideality of correlation to improve aquatic solubility model. J Mol Graph Model. 2020;96: 107525.
35. Ahmadi S, Mardinia F, Azimi N, Qomi M, Balali E. Prediction of chalcone derivative cytotoxicity activity against MCF-7 human breast cancer cell by Monte Carlo method. J Mol Struct. 2019;1181:305–11.
36. Mahapatra MK, Karuppasamy M. Fundamental considerations in drug design. In: Computer Aided Drug Design (CADD): from ligand-based methods to structure-based approaches. Elsevier; 2022: 17–55.
37. Daoui O, Elkhattabi S, Chtita S. Rational design of novel pyridine-based drugs candidates for lymphoma therapy. J Mol Struct. 2022;1270: 133964.
38. Daoui O, Nour H, Abchir O, Elkhattabi S, Bakhouch M, Chtita S. A computer-aided drug design approach to explore novel type II inhibitors of c-Met receptor tyrosine kinase for cancer therapy: QSAR, molecular docking, ADMET and molecular dynamics simulations. J Biomol Struct Dyn. 2022. https://doi.org/10.1080/07391102.2022.2124456.
39. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminform. 2011;3(1):1–14.
40. 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.
41. Daoui O, Elkhattabi S, Chtita S. Rational identification of small molecules derived from 9, 10-dihydrophenanthrene as potential inhibitors of 3CLpro enzyme for COVID-19 therapy: a computer-aided drug design approach. Struct Chem. 2022;33(5):1667–90.
42. Liu Y-C, Huang V, Chao T-C, Hsiao C-D, Lin A, Chang M-F, Chow L-P. Screening of drugs by FRET analysis identifies inhibitors of SARS-CoV 3CL protease. Biochem Biophys Res Commun. 2005;333(1):194–9.
43. Alam S, Khan F. Virtual screening, docking, ADMET and system pharmacology studies on Garcinia caged Xanthone derivatives for anticancer activity. Sci Rep. 2018;8(1):1–16.
44. Matin MM, Hasan MS, Uzzaman M, Bhuiyan MMH, Kibria SM, Hossain ME, Roshid MH. Synthesis, spectroscopic characterization, molecular docking, and ADMET studies of mannopyranoside esters as antimicrobial agents. J Mol Struct. 2020;1222: 128821.
45. Almasirad A, Mousavi Z, Tajik M, Assarzadeh MJ, Shafiee A. Synthesis, analgesic and anti-inflammatory activities of new methyl-imidazolyl-1, 3, 4-oxadiazoles and 1, 2, 4-triazoles. Daru J Pharm Sci. 2014;22(1):1–8.
46. Faizi M, Jahani R, Ebadi SA, Tabatabai SA, Rezaee E, Lotfaliei M, Amini M, Almasirad A. Novel 4-thiazolidinone derivatives as agonists of benzodiazepine receptors: design, synthesis and pharmacological evaluation. EXCLI J. 2017;16:52.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC