Asian Journal of Chemistry; Vol. 37, No. 10 (2025), 2487-2499
ASIAN JOURNAL OF CHEMISTRY
https://doi.org/10.14233/ajchem.2025.34335

Machine Learning Models for the Identification of SARS-CoV-2 Main Protease (Mpro) Inhibitors: Development and Validation of Three-dimensional (3D) Quantitative Structure-Activity Relationship (QSAR) and Pharmacophore Models

ROXZANNE RAY[1], KALEIARESI RAJAN[1] and VASUDEVA RAO AVUPATI[1,2,*]

[1]Department of Pharmaceutical Chemistry, School of Pharmacy, IMU University (Formerly known as International Medical University), 126, Jalan Jalil Perkasa 19, Bukit Jalil, 57000 Kuala Lumpur, Malaysia
[2]Centre of Excellence for Bioactive Molecules & Drug Delivery (BMDD), Institute for Research, Development & Innovation (IRDI), IMU University (Formerly known as International Medical University), 126, Jalan Jalil Perkasa 19, Bukit Jalil, 57000 Kuala Lumpur, Malaysia

*Corresponding author: E-mail: vasudevaraoavupati@gmail.com

Received: 25 June 2025 *Accepted*: 16 September 2025 *Published online*: 30 September 2025 AJC-22133

The global pandemic of coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Recent World Health Organization (WHO) statistics show 778 million reported COVID-19 cases. To date, no specific drug has been found to treat COVID-19 effectively, largely due to the emergence of variants of concerns (VOCs). The main protease of SARS-CoV-2 is a well-established drug target to control viral replication in human host. We used a group of chemotypes with experimental Mpro inhibitory properties for the development of atom-based 3D-QSAR and ligand-based 3D-pharmacophore models using advanced machine learning strategies. The established 3D QSAR model is statistically significant ($R^2_{\text{Training set}}$ = 0.9897, Q^2 ($R^2_{\text{Test set}}$) = 0.5017), which demonstrated the model's strong predictive power. The 3D-QSAR model displays contour maps towards the positive and negative contribution of various functional groups based on the active and inactive ligands. On the other hand, we developed a ligand-based, three-point 3D pharmacophore model using 84 ligands (39 actives and 43 inactive) that has demonstrated statistically significant data related to the discrimination of active and inactive groups of compounds with a sensitivity of 97.4%, balanced accuracy of 63.8% and a perfect ROC-AUC of 1.0, internal validation revealed that AAD2 was the best-performing hypothesis. The three-point 3D-pharmacophore model shows fitness and alignment relative to the arrangement of atoms and groups within the active and inactive subsets. In summary, the 3D QSAR and pharmacophore models developed in this study could be used as a virtual screening tool to identify virtual hits as potential SARS-CoV-2 Mpro inhibitors.

Keywords: Machine learning model, SARS-CoV-2, Main protease, COVID-19, Virtual hits, Contour maps, Fitness score.

INTRODUCTION

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), the causative agent of COVID-19, has led to a global health crisis since its emergence in late 2019. Rising morbidity and death followed the virus's late 2019 appear as it has infected millions worldwide [1]. World Health Organization (WHO) statistics indicate that as of March 2025, more than 52,200 COVID-19 cases existed worldwide [2]. Although vaccine development has been a major turning point, the appearance of new variations with greater transmissibility and potential resistance has emphasized the ongoing need for effective treatment medicines [3]. Finding novel antiviral medications is a crucial area in research, which will help to solve the challenges brought on by SARS-CoV-2 and future coronaviruses [1]. Usually a difficult and time-consuming process, drug discovery, calls for years of development before clinical trials are run [4]. Creative strategies are required to speed up therapy discovery in worldwide health issues such as COVID-19. Powerful tools have developed from the computational methods such as pharmacophore modelling [5]. These approaches reduce the time and cost of experimental screening by predicting biological activity using algorithms and molecular modelling tools [6].

An essential part of drug development is the quantitative structure-activity relationship (QSAR), which effectively and economically predicts molecular actions and characteristics [7]. One of the most effective approaches for investigating a

compound's mechanism of action and guiding drug development is 3D-QSAR, which quantitatively analyzes the relationship between the chemical structure of a small molecule and its biological activity [8]. 3D-QSAR has been crucial in recent years for early drug design, structural optimization and toxicity prediction [8,9]. It has been demonstrated that preclinical research phases for small-molecule therapeutic candidates can be accelerated by deep learning techniques, especially deep QSAR [7]. Before starting a phase I clinical trial, companies such as Exscientia plc and *in silico* Medicine Inc have stated that they must finish the exploratory research phase in a year. These achievements imply that according to the Gartner hype cycle, the deep QSAR sector is nearing its productivity peak. Particularly considering newly developing infectious disorders like COVID-19, the continuous advancement and application of deep QSAR techniques may facilitate the quicker discovery of small-molecule therapeutic candidates [10].

A pharmacophore model is the spatial arrangement of chemical characteristics required for the biological action of a molecule [11]. These characteristics include aromatic rings, hydrophobic regions, hydrogen bond donors and hydrogen bond acceptors [12]. Researchers can give priority to compounds with desired medicinal properties by building 3D pharmacophore models [13]. This strategy has significantly contributed to antiviral drug development [14] and has proven effective in identifying potential antiviral agents [15]. In line with our research focus, we aimed to develop virtual screening tools for identifying virtual ligands targeting the SARS-CoV-2 main protease (Mpro). In this study, we present the development and validation of 3D-QSAR and pharmacophore models as effective virtual screening tools.

EXPERIMENTAL

Computational software: The Maestro™ graphical user interface (GUI) with Schrödinger Drug Discovery Suite integrated software was used for 2D to 3D molecular modelling, ligand energy minimization and Phase™ machine learning software for atom-based 3D QSAR and pharmacophore modelling [16].

Dataset selection

3D QSAR modelling: In this study, the compound dataset was examined and collected from the literature reported as potential inhibitors of SARS-CoV-2 Main Protease from the Binding DB database and used for the development and validation of Phase™ atom-based 3D QSAR model. The dataset was constructed from a homologous series of 36 novel benzodiazepine derivatives reported by Wang *et al*. [17] as SARS-CoV-2 inhibitors, as shown in Table-1.

3D Pharmacophore modelling: After retrieving 272 articles from binding DB, which each reported different compounds studied as possible inhibitors of the SARS-CoV-2 major protease Mpro, the ligand selection process got underway. To evaluate the dependability and significance of the reported chemicals, a comprehensive screening was carried out. Only those that were specifically linked to SARS-CoV-2 and backed up by reliable reference materials were kept. Compounds without easily accessible and reliable sources or associated with other viral targets were not included. The sample was reduced to 87 candidates after 185 chemicals were removed. Only drugs evaluated utilizing the FRET-based enzymatic assay (Fluorescence Resonance Energy Transfer), a proven technique for measuring Mpro inhibition, were chosen in order to guarantee uniformity in the assay target [18]. About 16 peer-reviewed papers that offered consistent assay conditions were found using this criterion. Out of them, 84 compounds were selected for 3D pharmacophore modelling. IC_{50} values were used to classify the chosen ligands according to their inhibitory activity. Active compounds had IC_{50} values less than 1 μM, whilst inactive compounds had values greater than 1 μM or were said to exhibit little to no inhibition [19]. A distinct division for model development was made possible by this classification. To capture a broad chemical diversity, the dataset internal validation (84 ligands, Table-2) and external validation (20 compounds, Table-3) covering non-covalent inhibitors [20] and bioactive ligands from different scaffolds such as di- and trihaloacetamides [21], boceprevir analogs [22], aldehyde-based tripeptidyl compounds [23], trisubstituted piperazine derivatives [24], benzodiazepines [17], alkynes [25], quinazoline-4-one derivatives [26], dithiocarbamates [27], spirocyclic compounds [28], 9,10-dihydrophenanthrene derivatives [29], triazine analog [30], 2-(furan-2-ylmethylene)-hydrazine-1-carbothioamide derivatives [31], peptidomimetics [32], 2-thiobenzimidazoles [33] and bioflavonoids [34].

Dataset curation: Dataset curation is the most important and crucial stage in atom-based 3D QSAR and pharmacophore modelling. Even though the dataset chosen is a possible resource for the development of the atom-based 3D QSAR and pharmacophore models, there could be existing sample errors that usually prevents the carefully selected dataset from being utilized directly for modelling purposes. The chosen dataset underwent chemical and biological data curation by utilizing different curation filters in order to create a dependable and consistent atom-based 3D QSAR and pharmacophore models. These include elimination of mixtures, elimination of redundant chemical structures, the standardization of homologous chemotypes for QSAR and heterologous for 3D pharmacophore modelling. The dataset was then manually examined using the exclusion criteria. After the curation process, the ligands were prepared for modelling simulations [16].

Ligand preparation: Two dimensional (2D) molecular structure was sketched using the 2D Sketcher tool, which was integrated into the Schrödinger Maestro™ graphics platform interface to model the ligand. By using the LigPrep software along with the Optimized Potential for Liquid Simulations (OPLS_2015) force field, all compounds in the dataset were maximized to determine the 3D lower-energy conformations that comprise the ligands. Each analyzed input structure gave a single 3D low-energy structure having set chiralities [16].

Development and validation of 3D QSAR model: The training and test set was first separated from the prepared dataset in proportions of 70% and 30%, respectively. The training and test set will be selected from the dataset using the randomization built into the Schrödinger Phase™ program [16]. The training set, which include the model compounds with a grid spacing was used to build the model. By using the formula

TABLE-1
ATOM-BASED 3D QSAR MODEL DEVELOPMENT AND VALIDATION DATASET

Ligand code	Dataset	SMILES codes	Activity IC_{50} (nM)	Predicted activity pIC_{50} (nM)
11a	Training	COC(=O)c1ccc2c(c1)NC1=C(C(=O)CCC1)C(c1ccc(C(F)(F)F)cc1)N2C(=O)CCl	0.745	0.648
7zi	Test	O=C1CCCC2=C1C(c1ccc(C(F)(F)F)cc1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.009	0.082
7zd	Training	O=C1CCCC2=C1C(C1CCCCC1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.201	-0.192
7ze	Training	O=C1CCCC2=C1C(c1ccccc1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.223	-0.213
11b	Test	N#Cc1ccc2c(c1)NC1=C(C(=O)CCC1)C(c1ccc(C(F)(F)F)cc1)N2C(=O)CCl	-0.265	-0.263
7zc	Training	O=C1CCCC2=C1C(C1CCCC1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.283	-0.274
7zh	Training	O=C1CCCC2=C1C(c1ccc(OC(F)(F)F)cc1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.307	0.032
7zg	Training	COc1ccc(C2C3=C(CCCC3=O)Nc3cc([N+](=O)[O-])ccc3N2C(=O)CCl)cc1	-0.312	-0.316
7zf	Training	Cc1ccc(C2C3=C(CCCC3=O)Nc3cc([N+](=O)[O-])ccc3N2C(=O)CCl)cc1	-0.371	-0.53
7zb	Training	O=C1CCCC2=C1C(C1CCC1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.394	-0.369
7za	Test	O=C1CCCC2=C1C(C1CC1)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.407	-0.428
7w	Test	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccc(C(F)(F)F)cc1	-0.52	-0.54
7t	Training	Cc1ccc(C2C3=C(CC(C)CC3=O)Nc3cc([N+](=O)[O-])ccc3N2C(=O)CCl)cc1	-0.525	-0.513
7s	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccccc1	-0.526	-0.602
7r	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CCCCC1	-0.56	-0.586
7z	Test	CC1C2=C(CCCC2=O)Nc2cc([N+](=O)[O-])ccc2N1C(=O)CCl	-0.567	-0.548
7k	Training	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccc(C(F)(F)F)cc1	-0.572	-0.564
7h	Training	Cc1ccc(C2C3=C(CC(C)(C)CC3=O)Nc3cc([N+](=O)[O-])ccc3N2C(=O)CCl)cc1	-0.595	-0.606
7zj	Training	O=C1CCCC2=C1C(c1ccc(F)cc1F)N(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.612	-0.523
7v	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccc(OC(F)(F)F)cc1	-0.618	-0.578
7u	Training	COc1ccc(C2C3=C(CC(C)CC3=O)Nc3cc([N+](=O)[O-])ccc3N2C(=O)CCl)cc1	-0.648	-0.639
7y	Test	O=C1CCCC2=C1CN(C(=O)CCl)c1ccc([N+](=O)[O-])cc1N2	-0.708	-0.744
7q	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CCCC1	-0.712	-0.689
7f	Test	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CCCCC1	-0.721	-0.776
7j	Training	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccc(OC(F)(F)F)cc1	-0.754	-0.613
7x	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccc(F)cc1F	-0.793	-0.781
7p	Test	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CCC1	-0.865	-0.836
7o	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CC1	-0.957	-0.88
7e	Training	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CCCC1	-0.96	-0.894
7l	Test	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2c1ccc(F)cc1F	-0.98	-1
7d	Training	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CCC1	-0.981	-0.996
7n	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C	-1.04	-1.055
7m	Training	CC1CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2	-1.061	-1.149
7c	Training	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2C1CC1	-1.125	-1.123
7b	Test	CC1C2=C(CC(C)(C)CC2=O)Nc2cc([N+](=O)[O-])ccc2N1C(=O)CCl	-1.368	-0.566
7a	Training	CC1(C)CC(=O)C2=C(C1)Nc1cc([N+](=O)[O-])ccc1N(C(=O)CCl)C2	-1.387	-1.073

pIC_{50} = -$\log_{10}[IC_{50}]$ or pEC_{50} = -$\log_{10}[EC_{50}]$, the reported anti-SARS-CoV-2 activities of the chosen dataset was first to be transformed to their respective negative logarithmic values (pIC_{50}/pEC_{50}). The Schrödinger Discovery Suite software's Phase™ module was utilized to construct atom-based 3D QSAR employing these values as dependant parameters [16].

Contour map of combined effects along with each individual effect was produced to understand the atom-based 3D QSAR model. For the combined effects contour map, the unfavourable and favourable areas will be represented by red and blue cube areas, respectively, signifying places that enhance or diminish anti-SARS-CoV-2 activity. Important structural characteristics that affect the biological action of ligands was shown by using these contour maps. In terms of biological activity, the red cubes will draw attention to areas that are not beneficial, while the blue cubes will indicate those that are [16].

The atom-based 3D QSAR model was visualized using different ligands from the dataset to understand the relative importance of each feature for instance, X (miscellaneous other atom types), W (electron-withdrawing), N (negative electrostatic), P (positive electrostatic), H (hydrophobic or nonpolar) and D (hydrogen-bond donors). From the created 3D QSAR model, graphical depictions of the contour maps for the most and least active ligands for both the training set and test set will be produced [16].

The atom-based 3D QSAR model was assessed and validated using a variety of statistical metrics published by Phase™. Using a maximum of N/5 partial least square (PLS) factors where N is the number of molecules in the training set, a predictive model was built using the PLS regression approach. The predictive 3D QSAR model that fell within the reasonable range of statistical parameters was deemed to be significant [16]. The biological activities of the test set were predicted to verify the best atom-based 3D QSAR model. Several statistical metrics, such as the regression coefficient (R^2), cross-validated coefficient (Q^2), regression's standard deviation (SD), variance ratio (F), root-mean-square error

TABLE-2
3D PHARMACOPHORE HYPOTHESIS DEVELOPMENT AND INTERNAL VALIDATION DATASET

Ligand code	SMILES code	IC_{50} (nM)	Dataset Type
L01	COc(ccc1)c2c1[nH]c(C(N[C@H](C(N[C@H](C(CO)=O)C[C@@H]3CCNC3=O)=O)CC(C)C)=O)c2	5	Active
L02	COc1c2cc(C(N[C@@H](CC(C)C)C(N[C@@H](C[C@@H]3CCNC3=O)C(CO)=O)=O)=O)[nH]c2ccc1	33	Active
L03	CC(C[C@@H](C(N[C@H](C(S([O-])(=O)=O)O)CC1CCNC1=O)=O)NC(OCc2ccccc2)=O)C	74	Active
L04	C[C@@H](c1ccccc1)NC([C@H](c2cnccc2)N(c3ccc(c4ccccc4)cc3)C(C(Br)(Br)Cl)=O)=O	9040	Inactive
L05	CC(C)(NC(N[C@@H](C(C)(C)C)C(N1C[C@H]2[C@H](C2(C)C)[C@H]1C(N[C@H](C=O)C[C@@H]3CCNC3=O)=O)=O)=O)C	9.3	Active
L06	CC(C)(C[C@@H](C(N[C@H](C=O)C[C@@H]1CCNC1=O)=O)NC([C@H](C(C)(C)C)NC(NC(C)(C)C)=O)=O)C	17	Active
L07	CC(CC(N[C@@H](C(C)(C)C)C(N1C[C@H]2[C@H](C2(C)C)[C@H]1C(N[C@H](C=O)C[C@@H]3CCNC3=O)=O)=O)=O)C	22	Active
L08	CCCCCCNC(C([C@@H](NC([C@@H]1[C@@H]2[C@@H](C2(C)C)CN1C([C@H](C(C)(C)C)NC(NC(C)(C)C)=O)=O)=O)C[C@@H]3CCNC3=O)=O)=O	720	Inactive
L09	CC([C@@H](C(N[C@H](C(N[C@H](C=O)C[C@@H]1CCNC1=O)=O)CC2CC2)=O)NC(OCc3ccccc3)=O)C	4.8	Active
L10	CC(C)[C@H](NC(=O)OCc1ccccc1)C(=O)N[C@@H](Cc1ccco1)C(=O)N[C@@H](C[C@@H]1CCNC1=O)C=O	5.2	Active
L11	O=C[C@@H](NC([C@@H](NC([C@H](C1CC1)NC(OCc2ccccc2)=O)=O)CC3CC3)=O)C[C@@H]4CCNC4=O	8.4	Active
L12	O=C[C@@H](NC([C@@H](NC(C1(NC(OCc2ccccc2)=O)CC1)=O)CC3CCCCC3)=O)C[C@@H]4CCNC4=O	350	Inactive
L13	COc1c(C(N2CCN(c3cc(Cl)c(Cl)cc3)[C@H](C(NCc4sccc4)=O)C2)=O)cncc1	170	Active
L14	COc1ccc(NC([C@@H]2CN(C(c3c(OC)ccnc3)=O)CCN2c4cc(Cl)c(Cl)cc4)=O)cc1	190	Active
L15	COc1ccc(NC([C@@H]2CN(C(c3cnccc3)=O)CCN2c4cc(Cl)c(Cl)cc4)=O)cc1	200	Active
L16	CN(Cc1cccs1)C(=O)[C@@H]1CN(CCN1c1ccc(Cl)c(Cl)c1)C(=O)c1cccnc1	1300	Inactive
L17	Clc1ccc(cc1Cl)N1CCN(C[C@H]1C(=O)NCc1ccc(Br)s1)C(=O)c1cccnc1	1800	Inactive
L18	Clc1c(Cl)cc(N2CCN(C(c3cnccc3)=O)C[C@H]2C(NS(=O)(c4sccc4)=O)=O)cc1	9500	Inactive
L19	COC(c1cc2c(N(C(CCl)=O)C(c3ccc(C(F)(F)F)cc3)C4=C(N2)CCCC4=O)cc1)=O	180	Active
L20	[O-][N+](c1cc2c(N(C(CCl)=O)C(c3ccc(C(F)(F)F)cc3)C4=C(N2)CCCC4=O)cc1)=O	1020	Active
L21	[O-][N+](c1cc2c(N(C(CCl)=O)C(C3=C(N2)CCCC3=O)C4CCCCC4)cc1)=O	1590	Active
L22	CC1(C)CC(=O)C2=C(C1)Nc1cc(ccc1N(C2C1CC1)C(=O)CCl)[N+]([O-])=O	11	Inactive
L23	CC1N(C(=O)CCl)c2ccc(cc2NC2=C1C(=O)CC(C)(C)C2)[N+]([O-])=O	23300	Inactive
L24	CC1(CC(C2=C(Nc3c(N(C(CCl)=O)C2)ccc([N+]([O-])=O)c3)C1)=O)C	24400	Inactive
L25	CC(C)([C@@H](C(N1C[C@H]2[C@H](C2(C)C)[C@H]1C(N[C@H](C#C)C[C@@H]3CCNC3=O)=O)=O)NC(C(F)(F)F)=O)C	63	Active
L26	CC(C)([C@@H](C(N1C[C@H]2[C@H](C2(C)C)[C@H]1C(N[C@H](C#CC(F)(F)F)C[C@@H]3CCNC3=O)=O)=O)NC(C(F)(F)F)=O)C	170	Active
L27	CC(C)([C@@H](C(N1C[C@H]2[C@H](C2(C)C)[C@H]1C(N[C@H](C=O)C[C@@H]3CCNC3=O)=O)=O)NC(C(F)(F)F)=O)C	290	Active
L28	CC(C)C[C@H](NC(=O)OCc1ccccc1)C(=O)N[C@@H](CCC(N)=O)C#C	3500	Inactive
L29	CNS(=O)(=O)C=C[C@H](CCC(N)=O)NC(=O)[C@H](CC(C)C)NC(=O)OCc1ccccc1	3900	Inactive
L30	CC(C[C@@H](C(N[C@H](C(C(Nc1ccccc1)=O)=O)CCC(N)=O)=O)NC(OCc2ccccc2)=O)C	80000	Inactive
L31	Oc(c1)c(O)c(O)c2c1nc(c3ccccc3)n(c4ccccc4)c2=O	83	Active
L32	CC(Cc(nc1c2c(O)c(O)c(O)c1)n(c3cc(C)c(F)cc3)c2=O)C	100	Active
L33	CC(Cc(nc1c2c(O)c(O)c(O)c1)n(c3ccccc3)c2=O)C	107	Active
L34	Oc(cc1nc(c2cccc(Br)c2)[nH]3)c(O)c(O)c1c3=O	8020	Inactive
L35	CC(C)(c1cc(c(nc2c3c(O)c(O)c(O)c2)[nH]c3=O)ccc1)C	8420	Inactive
L36	Oc(c1)c(O)c(O)c2c1nc(c3ccc(Br)cc3)[nH]c2=O	9800	Inactive
L37	Cc(ccc1)n2c1nc(CSC(NCc3cnccc3)=S)c2	8	Active
L38	CC(C[C@@H](C(N[C@H](C(S(O)(=O)=O)O)CC1CCNC1=O)=O)NC(OCc2ccccc2)=O)C	12	Active
L39	CC(C[C@@H](C(N[C@H](C(S(O)(=O)=O)O)CC1CCNC1=O)=O)NC(OCc2ccccc2)=O)C	20	Active
L40	S=C(OCc1ccccc1)NCc2cnccc2	100000	Inactive
L41	Cc(ccc1)n2c1nc(CNC(NCc3cnccc3)=S)c2	100000	Inactive
L42	Cc(ccc1)n2c1nc(CNC(NCc3cnccc3)=S)c2	3500	Inactive
L43	CC(C)C[C@H](NC(=O)OC1CC2(C1)CCN(CC2)S(C)(=O)=O)C(=O)N[C@@H](C[C@H]1CCNC1=O)C(=O)C(O)S([O-])(=O)=O	240	Active
L44	CC(C)C[C@H](NC(=O)OC1CC2(C1)CCN(CC2)C#N)C(=O)N[C@@H](C[C@H]1CCNC1=O)C=O	320	Active
L45	CC(C)C[C@H](NC(=O)OCC1CN(C1)C(=O)OC(C)(C)C)C(=O)N[C@@H](C[C@H]1CCNC1=O)C=O	330	Active
L46	CC(C)C[C@H](NC(=O)OCc1ccccc1)C(=O)N[C@@H](C[C@H]1CCNC1=O)C(=O)C(O)S([O-])(=O)=O	410	Active
L47	CC(C)C[C@H](NC(=O)OCC1CN(C1)S(C)(=O)=O)C(=O)N[C@@H](C[C@H]1CCNC1=O)C=O	4950	Inactive

L48	C[C@H](NC(=O)[C@H](N(C(=O)c1ccco1)c1ccc(cc1)-c1ccccc1)c1cccnc1)c1ccccc1	200	Active
L49	C[C@H](NC(=O)C(N(C(=O)c1ccco1)c1ccc(cc1)-c1ccccc1)c1cnccn1)c1ccccc1	280	Active
L50	C[C@H](NC(=O)C(N(C(=O)c1ccco1)c1ccc(cc1)C1CCCCC1)c1cccnc1)c1ccccc1	510	Active
L51	CC(C)C[C@H](NC(=O)OCc1ccccc1)C(=O)N[C@@H](C[C@@H]1CCNC1=O)C(O)S([O-])(=O)=O	50000	Inactive
L52	C[C@H](NC(=O)C(N(C(=O)c1ccco1)c1ccc(Cc2ccccc2)cc1)c1cccnc1)c1ccccc1	35100	Inactive
L53	CCN(CC)C(=S)SSC(=S)N(CC)CC	1040	Active
L54	OCC1Cc2c(cccc2-c2c(O)ccc(-c3ccc(Br)cc3)c12)-c1cc(ccn1)-c1ccccc1	1550	Active
L55	OCC1Cc2c(cccc2-c2c(O)ccc(-c3ccc(Br)cc3)c12)-c1cc(C=O)ccn1	1810	Active
L56	CC(=O)OCC1Cc2c(ccc(Cl)c2-c2c(O)ccc(C)c12)-c1ccccn1	8560	Inactive
L57	OCC1Cc2c(cccc2-c2c(O)ccc(-c3ccc(Br)cc3)c12)-c1cc(ccn1)-c1ccccc1	6090	Inactive
L58	OCC1Cc2c(cccc2-c2c(O)ccc(-c3ccc(Br)cc3)c12)-c1cc(C=O)ccn1	7640	Inactive
L59	NC(=S)NN=Cc1ccc(Cc2ccccc2)o1	1550	Active
L60	NC(=S)NN=Cc1ccc(o1)-c1ccc(C(O)=O)c(O)c1	1570	Active
L61	NC(=S)NN=Cc1ccc(o1)-c1ccc(cc1)C(N)=O	4000	Active
L62	NC(=S)NN=Cc1ccc(o1)-c1ccc(Cl)c(c1)C(O)=O	25400	Inactive
L63	NC(=S)NN=Cc1ccc(o1)-c1ccc(C(O)=O)c(Cl)c1	32500	Inactive
L64	NC(=S)NN=Cc1ccc(o1)-c1cccc(C(O)=O)c1Cl	40600	Inactive
L65	CC(C)C[C@H](NC(=O)[C@@H](NC(=O)c1ccc(Cl)cc1)C(C)C)C(=O)N[C@@H](C[C@@H]1CCCNC1=O)C=O	148	Active
L66	CC(C)C[C@H](NC(=O)[C@@H](NC(=O)c1ccncc1)C(C)C)C(=O)N[C@@H](C[C@@H]1CCCNC1=O)C=O	153	Active
L67	CC(C)C[C@H](NC(=O)[C@@H](NC(=O)c1ccc[nH]1)C(C)C)C(=O)N[C@@H](C[C@@H]1CCCNC1=O)C=O	172	Active
L68	O=C[C@H](C[C@@H]1CCNC1=O)NC(=O)[C@H](CC#C)NC(=O)C=Cc1ccccc1	3890	Inactive
L69	CC(C)[C@H](NC(=O)C=Cc1ccccc1)C(=O)N[C@@H](C[C@@H]1CCCNC1=O)C=O	6930	Inactive
L70	C[C@H](NC(=O)C=Cc1ccccc1)C(=O)N[C@@H](C[C@@H]1CCCNC1=O)C=O	12300	Inactive
L71	Cc1cncc(c1)-n1c(=O)nc(Nc2cc3cn(C)nc3cc2Cl)n(Cc2cc(C#N)c(F)cc2C)c1=O	2	Active
L72	Cn1cc2cc(Nc3nc(=O)n(Cc4cc5ccsc5[nH]c4=O)c(=O)n3Cc3cc(F)c(F)c(F)c3)c(Cl)cc2n1	4	Active
L73	Cn1cnc(Cn2c(=O)nc(Nc3cc4cn(C)nc4cc3Cl)n(Cc3cc(F)c(F)c(F)c3)c2=O)c1	6	Active
L74	Cc1cnc(Cn2c(=O)nc(Nc3cc4cn(C)nc4cc3Cl)n(Cc3cc(F)c(F)cc3F)c2=O)nc1	260	Inactive
L75	Cn1cc2cc(Nc3nc(=O)n(-c4cncc(c4)C(F)(F)F)c(=O)n3Cc3cc(F)c(F)c(F)c3)c(Cl)cc2n1	360	Inactive
L76	Cn1cnc(Cn2c(=O)nc(Nc3cc4cn(C)nc4cc3Cl)n(Cc3cccc4ccccc34)c2=O)n1	780	Inactive
L77	CC(C)(C)NC(=O)N[C@H](C(=O)N1C[C@H]2[C@@H]([C@H]1C(=O)NC(CC1CCC1)C(=O)C(N)=O)C2(C)C)C(C)(C)C	12100	Active
L78	Clc1ccc2n(Cc3ccccc3)c(SC(Cc3ccco3)C(=O)N3CCOC3=O)nc2c1	14900	Active
L79	Oc1ccc2nc(SC(Cc3ccco3)C(=O)N3CCOC3=O)n(Cc3ccccc3)c2c1	55800	Inactive
L80	Clc1ccc2n(Cc3ccccc3)c(SCC(=O)N3CCOC3=O)nc2c1	62200	Inactive
L81	Oc1ccc(cc1)-c1cc(=O)c2c(O)cc(O)c(-c3cc(ccc3O)-c3cc(O)c4c(cc(O)cc4=O)o3)c2o1	8300	Active
L82	Oc1cc(O)c2c(c1)oc(cc2=O)-c1ccc(O)c(O)c1	20000	Active
L83	Oc1ccc(cc1)-c1cc(=O)c2c(O)cc(O)cc2o1	281000	Inactive
L84	C[C@@]1(CC[C@H]2C(C1)=CC[C@@H]1[C@]2(C)CCC[C@@]1(C)C(O)=O)C=C	284000	Inactive

TABLE-3
3D PHARMACOPHORE HYPOTHESIS EXTERNAL VALIDATION DATASET

Compd. code	SMILES code	IC_{50} (nM)	Dataset type
C1	C[C@@H](OC(C)(C)C)[C@H](NC(=O)OCc1ccccc1)C(=O)N[C@@H](CC1CCCCC1)C(=O)N[C@@H](C[C@@H]1CCNC1=O)C=O	105	Active
C2	C[C@H](NC(=O)C(N(C(=O)c1ccco1)c1ccc(cc1)-n1cccc1)c1cccnc1)c1ccccc1	940	Active
C3	C[C@H](NC(=O)[C@H](N(C(=O)C(Br)Br)c1ccc(cc1)-c1ccccc1)c1cccnc1)c1ccc(Br)cc1	80	Active
C4	C[C@H](NC(=O)C(N(C(=O)c1ccco1)c1ccc(cc1)-c1cccs1)c1cccnc1)c1ccccc1	560	Active
C5	CC(C)C[C@H](NC(=O)OC1CC2(C1)CCN(CC2)C(=O)Cc1ccccc1)C(=O)N[C@@H](C[C@H]1CCNC1=O)C=O	350	Active
C6	CC(C)(C)NC(=O)N[C@H](C(=O)N[C@@H](CC1CC1)C(=O)N[C@@H](C[C@@H]1CCNC1=O)C=O)C(C)(C)C	26	Active
C7	CC(C)C[C@H](NC(=O)[C@@H](NC(=O)C=Cc1ccc(cc1)N(C)C)C(C)C)C(=O)N[C@@H](C[C@@H]1CCCNC1=O)C=O	592	Active
C8	CC(c1ccccc1)c1nc2cc(O)c(O)c(O)c2c(=O)n1-c1ccccc1	477	Active
C9	CC(C)(C)[C@H](NC(=O)C(F)(F)F)C(=O)N1C[C@H]2[C@@H]([C@H]1C(=O)N[C@@H](C[C@@H]1CCNC1=O)C#N)C2(C)C	760	Active
C10	CC(C)(C)[C@H](NC(=O)C(F)(F)F)C(=O)N1C[C@H]2[C@@H]([C@H]1C(=O)N[C@@H](C[C@@H]1CCNC1=O)C#C)C2(C)C	330	Active
C11	Oc1cc2n(Cc3ccccc3)c(SC(Cc3ccco3)C(=O)N3CCOC3=O)nc2cc1Cl	46100	Inactive
C12	COCCn1c(nc2ccccc2c1=O)C(=O)NCC(O)c1ccccc1	54000	Inactive

C13	C[C@H](NC(=O)[C@@H](N(C(=O)C(Br)(Br)Br)c1ccc(cc1)-c1ccccc1)c1cccnc1)c1ccccc1	7160	Inactive
C14	FC(F)(F)c1ccc(cc1)C1N(C(=O)CCl)c2ccc(cc2NC2=C1C(=O)CCC2)C#N	1840	Inactive
C15	COc1ccc(cc1)C1N(C(=O)CCl)c2ccc(cc2NC2=C1C(=O)CCC2)[N+]([O-])=O	2050	Inactive
C16	COc1cccc2[nH]c(cc12)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](CCC(N)=O)C#C	22000	Inactive
C17	S=C(NCc1cccnc1)SCc1ccccc1	9360	Inactive
C18	CC(C)C[C@H](NC(=O)OC1(C)CN(C1)C(=O)OC(C)(C)C)C(=O)N[C@@H](C[C@H]1CCNC1=O)C=O	3650	Inactive
C19	CC(=O)OCC1Cc2c(cccc2-c2c(O)ccc(-c3ccc(Br)cc3)c12)-c1cc(ccn1)-c1ccccc1	2460	Inactive
C20	CC(=O)OCC1Cc2c(cccc2-c2c(O)ccc(-c3ccc(Br)cc3)c12)-c1cc(C=O)ccn1	3200	Inactive
C21	NC(=S)NN=Cc1ccc(o1)-c1ccc(cc1)S(N)(=O)=O	5390	Inactive

(RMSE), statistical significance (P) and Pearson correlation coefficient (Pearson-R), was used to evaluate the predictive performance of the QSAR models. The best-fit model from the dataset was chosen using the combined data, which is shown in Table-4 [16].

The developed model was validated using an external test set that includes the ligands that are unlikely to be used in the construction of the atom-based 3D QSAR model. Considering this, data obtained from the test and training sets was examined. In addition, the robustness of the atom-based 3D QSAR model was confirmed using a Y-randomization test (scrambled actions). Utilizing the same independent variable structure, a new 3D QSAR model was created and the dependent variable (biological activity data) was mixed. Statistical significance was assessed for the model following several runs of this test [16].

Development and validation of 3D pharmacophore model: Schrödinger Maestro™ was used to handle the dataset for pharmacophore modelling after the final selection of 84 ligands [35]. First, conformer generation was carried out utilizing the conformer generation module in order to thoroughly investigate the accessible 3D conformational space of every molecule. This stage assisted in producing stable three-dimensional forms of the molecules that most likely depict their appearance upon binding to the target protein. Following that, LigPrep was used to standardize and optimize the structure of conformers. To guarantee uniformity and compatibility with later modelling processes, suitable ionization states, tautomeric forms and precise 3D geometries were allocated during this phase [36].

Phase™ produced 15 pharmacophore hypotheses based on the classification of 39 active and 43 inactive substances (based on their IC_{50} values). Through internal validation, each hypothesis was assessed based on how successfully it minimized the alignment of inactive ligands while identifying active ones (fitness score > 1.5). During this assessment, other performance indicators like site score, ROC curve and survival score were also considered [37].

An independent set of 21 ligands, not included in the original training dataset, was used for external validation to assess the predictive reliability and generalizability of the pharmacophore hypothesis. The top-performing hypothesis (AAD2) was used to filter these ligands, which included both active and inactive molecules. To evaluate the model's ability to distinguish between active and inactive compounds in a practical context, key performance metrics including sensitivity, specificity, accuracy, precision and ROC AUC were calculated [38].

The external validation's findings confirmed that AAD2 could accurately and reliably predict active chemicals. The model's sensitivity stayed high, suggesting that the hypothesis is still able to accurately identify actives [38]. However, the specificity revealed some room for improvement in the Mpro model, as certain inactive compounds were incorrectly classified as active. Nevertheless, the hypothesis maintained high accuracy, indicating that the model was generally effective in distinguishing between active and inactive compounds. The model's outstanding capacity to distinguish between active and inactive ligands was confirmed by the consistently high ROC AUC value [39]. These findings demonstrate that AAD2 is a robust and reliable virtual screening approach that may be used to forecast novel compounds that may have antiviral properties against SARS-CoV-2. To sum up, the external validation showed that the pharmacophore hypothesis had predictive value for further research in addition to being applicable to the training set. This makes AAD2 a useful tool for virtual screening in drug discovery and the detection of the possible inhibitors of SARS-CoV-2 [40].

RESULTS AND DISCUSSION

Atom-based 3D QSAR modelling

Alignment of 3D ligand conformations: The molecular 3D conformational alignment is crucial for developing the atom-based 3D QSAR model using Phase™ software. We used a template-ligand alignment technique, with each of the 36 dataset compounds serving as a template-ligand, matched to molecules in both training and test sets. Fig. 1 shows the conformational alignment results in a 3D workspace.

TABLE-4
3D QSAR MODEL STATISTICS

PLS	R^2	Q^2	SD	F-value	RMSE	P-value	Pearson-R
1	0.7450	0.3322	0.2034	70.1	0.30	1.39e-08	0.6563
2	0.9121	0.4315	0.1220	119.3	0.27	7.17e-13	0.7361
3	0.9522	0.5052	0.0920	146.0	0.26	1.14e-14	0.7584
4	0.9750	0.4650	0.0681	204.6	0.27	1.71e-16	0.7416
5	0.9897	0.5017	0.0449	383.0	0.26	3.89e-19	0.7440

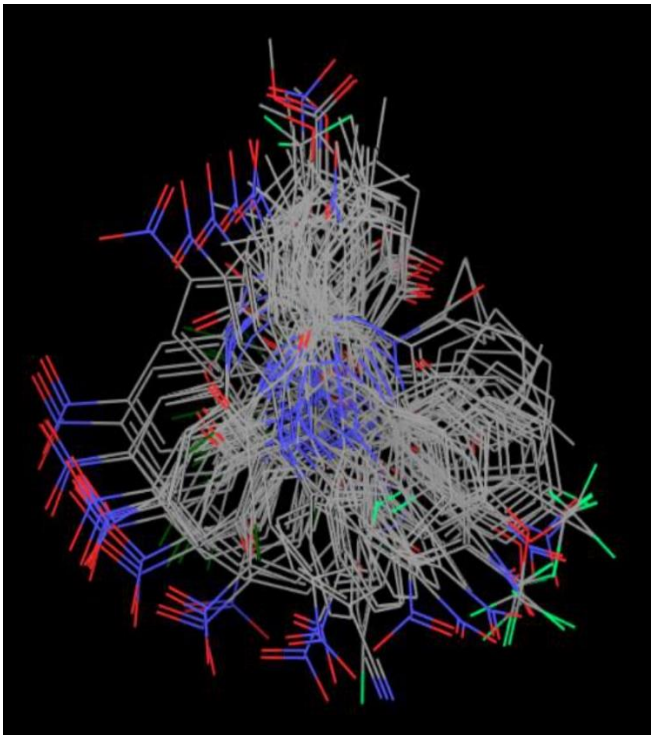

Fig. 1. Alignment of 3D ligand conformations

Atom-based 3D QSAR contour map analysis: The combined effect, electron-withdrawing and hydrophobic feature contour maps for a set of most active (**11a**) and least active (**7a**) compounds forming part of the training set and the combined effect for a set of most active (**7zd**) and least active (**7c**) compounds forming part of the test set, respectively. The favourable and unfavourable regions are displayed in contour maps that were studied, analyzed and derived logical discussion points useful for designing new molecules. The contour map results obtained are displayed in Fig. 2.

The structural characteristics and contour maps Fig. 2a-b for the training set are compounds **11a** (most active) and **7a** (least active). In the analysis of compound **11a**, the aromatic core and the substituent positions at C4 and C17 are encompassed by large blue regions on the contour map, representing favourable zones that suggest adding hydrophobic or electron rich substituents at these sites would significantly enhance the biological activity. The high pIC_{50} value observed for this compound aligns with these favourable zones, indicating an optimal 3D conformation for effective interactions with the target. Conversely, a red area around C9 indicates that introducing more electronegative substituents at this position would be unfavourable, since C9 is already an electron-deficient acyl chloride group (**14**). Adding electronegative groups here could lead to unfavourable interactions and reduced activity.

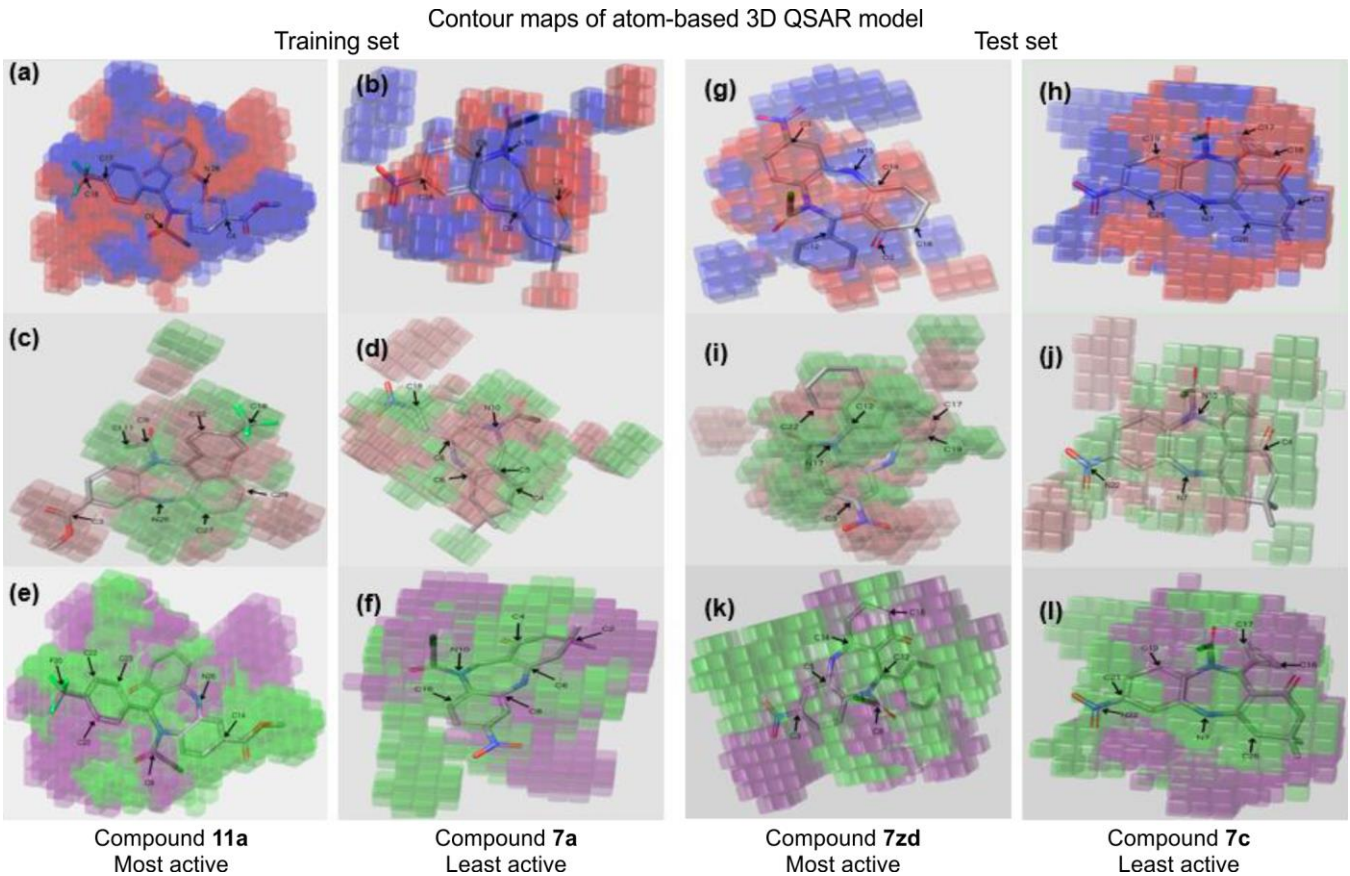

Fig. 2. The combined effect (a-b), electron-withdrawing (c-d) and hydrophobic (e-f) feature contour maps (a-e) for a set of most active (**11a**) and least active (**7a**) compounds forming part of the training set and the combined effect (g-h), electron-withdrawing (i-j) and hydrophobic (k-l) feature contour maps (g-l) for a set of most active (**7zd**) and least active (**7c**) compounds forming part of the test set. The favourable and unfavourable regions are displayed in contour maps, indicated by various colours: positive threshold colours (blue, pink and green) and negative threshold colours (red, pale green and purple) in training and test sets

Regarding compound **7a**, the contour map shows red zones around C4, which is part of the carbonyl group (C=O), suggesting that substituents at this position might hinder activity due to unfavourable interactions caused by the electronegative carbonyl. Such modifications could misalign the molecule within the active site, decreasing its efficacy. Interestingly, C9, attached to an aromatic benzene ring bonded to nitrogen, is situated within a blue, favourable region despite being electron rich. The nitrogen atom reduces the electron density at C9, making it less nucleophilic; however, this marginal electron deficiency allows hydrophobic and electron-donating substituents at C9 to enhance interactions like π-π stacking, promoting activity. Nonetheless, highly electronegative groups at C9 would further deplete electron density, potentially diminishing the biological activity.

The structural characteristics and contour maps for the test set compounds **7zd** (most active) and **7c** (least active) elucidate the differences in their biological activities, as depicted in Fig. 2g-h. In compound **7zd**, the C18 and C12 atoms, part of the benzene ring, are located within blue areas on the contour map. These regions indicate favourable interactions, such as π-π stacking and electropositive substituents, which contribute to increased activity of the compound. Conversely, C3 in **7zd** is positioned in the red zone due to its attachment to an electron-withdrawing NO_2 group, which is generally unfavourable for electronegative substituents. Nonetheless, the high activity of **7zd** persists since the overall presence of hydrophobic and π-interacting groups compensates for the unfavourable influence of this region. In contrast, compound **7c** exhibits the lowest activity within the test set, largely due to the presence of a cyclopropyl substituent at C17 and C18. This bulky, hydrophobic group causes steric hindrance, which is mapped to the red region, thus reducing activity. Although C3 is situated within a blue, favourable zone owing to its connection to the benzene ring, enabling potential π-π interactions, this positive effect is insufficient to counteract the detrimental steric and electrostatic environment around C18. Overall, these findings highlight the critical role of substituent location and nature in modulating activity; in **7c**, unfavourable steric hindrance and electrostatic interactions at C18 markedly decrease inhibitory potency, whereas in **7zd**, favourable π-π interactions within the blue regions enhance its biological activity.

The electron-withdrawing contour maps presented in Fig. 2c-d clearly illustrate regions where electron-withdrawing substituents are likely to enhance or diminish biological activity. The pink areas denote favourable regions, where such substituents are predicted to increase activity, while pale-green regions indicate unfavourable locations for electron-withdrawing groups. In compound **11a**, as shown in Fig. 2c, electron-withdrawing groups like the carbonyl at C3 and the chlorine-containing group at C9 are situated within the pink (favourable) regions of the contour map. These areas suggest that these substituents can engage in interactions with the target protein's active site that benefit from electron-withdrawing properties, thereby likely contributing to increased activity. Conversely, the pale green zones contain the N26 atom, part of an amide linkage (N–H) and C18, attached to a $–CF_3$ group. Since these regions are unfavourable for electron withdrawal, incorporating electron withdrawing substituents, would not enhance activity and might even be detrimental. For compound **7a**, depicted in Fig. 2d, the contrast in the map highlights a different pattern. In particular, the pale green zone, unfavourable for electron-withdrawing groups, includes the N19 atom bearing an NO_2 group. Its presence in this region indicates that the strong electron-withdrawing effect of NO_2 at this position does not support beneficial interactions with the active site, likely reducing its biological activity. Conversely, the pink area contains N10, part of an acyl chloride group, which exhibits electron-withdrawing character aligned with the protein's binding site preferences, positively influencing activity. Fig. 2i–j compares the most and least active compounds, with the electron density maps highlighting the crucial role of electron-withdrawing group placement. For compound **7zd**, the pink region notably includes the NO_2 group attached to N2. This strong electron-withdrawing group enhances interactions with the active site, contributing to high activity. Conversely, the pale green zone contains N17, part of an acyl chloride, indicating that despite its electron-withdrawing nature, its location does not favour interactions that bolster activity, aligning with its moderate activity profile. For compound **7c**, the map shows the pink area contains N10, part of an acyl chloride, which can favourably influence binding through stabilizing interactions. Meanwhile, N22, attached to an NO_2 group, resides in a pale green zone, indicating an unfavourable location for such a strong electron withdrawing substituent. Although the NO_2 group is inherently strongly withdrawing electrons, its placement in this region does not match the binding site preferences, likely contributing to its low biological activity. Overall, these contour maps demonstrate that the strategic positioning of electron withdrawing groups significantly impacts the activity profiles of compounds, where optimal placement aligns with regions favouring electron withdrawal and misplacement can offset potential gains.

The hydrophobic contour maps in Fig. 2e–f provide the insights into the influence of substituent positioning on biological activity through favourable or unfavourable hydrophobic interactions with the target protein. In Fig. 2e, compound **11a** shows that the C14 carbon on the benzene ring is in a green region, indicating that hydrophobic substitution here likely enhances binding affinity by favourably complementing the hydrophobic environment of binding site. Conversely, the C9 carbon attached to the acyl chloride group and the N26 atom (part of an N–H group) fall into purple zones associated with unfavourable hydrophobic interactions, which aligns with the polar nature of N–H groups that prefer hydrogen bonding rather than hydrophobic contacts. The overall weak activity of this compound reflects these unfavourable regions impacting binding efficiency. Interestingly, compound **7a**, which exhibits the lowest biological activity, presents a different scenario. Despite N7 and N10 atoms located in regions marked as favourable for hydrophobic interactions, though these are polar groups unlikely to enhance hydrophobic contact, their placement suggests that polar substituents may tolerate or are partially preferred in these areas. Nonetheless, the poor overall activity indicates that other mismatched interactions, such as the placement of carbons C8 and C2 attached to benzene rings in unfavourable (purple) zones, diminish the compound's

efficacy, highlighting the importance of correct spatial alignment within the hydrophobic contours. Moving to Fig. 2k-l, for compound **7zd**, the C12 atom attached to a benzene ring falls within a favourable green region, supporting the idea that hydrophobic interactions at this position significantly contribute to high biological activity by stabilizing binding through alignment with hydrophobic regions of the protein. However, the C3 atom bearing an electron-withdrawing NO_2 group is in an unfavourable purple zone, implying that the polarity introduced by this substituent diminishes the hydrophobic interaction potential. Despite this, the compound maintains substantial activity, emphasizing the dominant role of hydrophobic contact at C12. In contrast, compound **7c**, which displays the lowest activity among tested molecules, has the C21 atom (attached to a strongly electron-withdrawing NO_2 group) in a papered green region, a paradox given that NO_2 groups are polar and typically unfavourable for hydrophobic interactions. The presence of C17 and C18, part of cyclopropyl groups positioned in purple (unfavourable), further supports the notion that mismatched hydrophobic and electronic properties impair binding. The low activity of this compound likely results from these unfavourable interactions, underscoring that even hydrophobic groups like cyclopropyl can be ineffective if misaligned with the target's hydrophobic pockets. Overall, the contour maps elaborate that strategic placement of hydrophobic substituents in regions marked as favourable can markedly improve biological activity, while mismatched or polar groups in these pockets tend to diminish the binding efficacy. These insights underline the importance of spatial orientation and electronic properties in optimizing compound interactions within the binding site.

Statistical validation: The contour maps from 36 benzodiazepine derivatives were used for this 3D-QSAR analysis and the statistical parameters from PLS regression with 1 to 5 factors are listed in Table-4. Using the n/5 rule, where n is the number of training compounds, a maximum of 5 components was set. This ensures the number of variable factors does not exceed one-fifth of the training set size, preventing overfitting. Key statistical parameter was used for assessing the model`s accuracy and predictability such as the regression coefficient (R^2), cross-validated coefficient (Q^2), regression's standard deviation (SD), variance ratio (F), root-mean-square error (RMSE), statistical significance (P) and Pearson correlation coefficient (Pearson-R).

The dataset was initially divided into training and test sets randomly. Multiple compound combinations were analyzed using partial least squares (PLS) to identify the most reliable model. According to Ren *et al.* [8] and Praveen *et al.* [41] an optimal model should have a regression coefficient (R^2) close to 1.000, indicating a good fit and a cross-validation coefficient (Q^2) above 0.5, reflecting strong predictive ability. Among the five PLS models generated, PLS 5 was the best, with $R^2 = 0.9897$, demonstrating excellent goodness-of-fit and $Q^2 = 0.5017$, indicating good predictive power. It also had the lowest RMSE (0.26) and standard deviation (SD = 0.0449), confirming high accuracy in activity prediction. The model's statistical significance was supported by a high F-value (383.0) and a very low P-value (3.89×10^{-19}). Although PLS 5 had a slightly lower Pearson-R value (0.7440) compared to PLS 3 (0.7584), its overall balanced performance underscores its robustness and suitability for further validation.

3D Pharmacophore modelling

Generation of pharmacophore hypotheses: The pharmacophore hypotheses generated in this study revealed key chemical features associated with ligand bioactivity against SARS-CoV-2 Mpro. Several hypotheses such as AAA3, AAH2 and AAD4, likely emphasized the importance of hydrogen bond acceptors (A), aromatic rings (R) and hydrophobic groups (H), highlighting the critical roles these features play in effective ligand binding. Hypotheses like AAH2 and AAA2, which incorporate multiple hydrogen bond acceptors and aromatic rings, suggest that the presence of these features enhances interaction potential within the active site, consistent with the importance of hydrogen bonding and π-π stacking interactions in ligand recognition.

Conversely, hypotheses such as AAD3 and AAD1, which include hydrogen bond donors (D), indicate that hydrogen bond donation from ligands also contributes significantly to activity, possibly by establishing stabilizing interactions with acceptor residues in the protease pocket. The diverse set of 15 hypotheses highlights the complexity of features affecting bioactivity and emphasizes the necessity to account for multiple interaction modes.

Comparison of these hypotheses against a dataset of active and inactive ligands using fitness scores and mapping statistics showed that certain hypotheses, such as AAA3 and AAH2, more effectively distinguished actives from inactives, underscoring their potential for virtual screening and lead optimization. This analysis emphasizes that features like hydrogen bond acceptors, aromaticity and hydrophobic groups are essential pharmacophoric elements for designing potent inhibitors of SARS-CoV-2 Mpro. The performance metrics of various 3D pharmacophore hypothesis developed in this study are shown in Table-5.

The AUC values for the four pharmacophore hypotheses (AAD2, AAA1, AAD3 and AAD1) are compared, AAD2 achieving the highest AUC of 1.0. This indicates that AAD2 demonstrated flawless classification performance in distinguishing active from inactive compounds, effectively differentiating between them. Although AAD2 exhibited an excellent ROC-AUC, its overall accuracy was 60.7%, with a balanced

TABLE-5
PERFORMANCE METRICS OF 3D PHARMACOPHORE HYPOTHESES

Hypothesis	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1-score (%)	AUC
AAD2	97.4	30.2	62.2	55.9	71.0	1.00
AAA1	87.5	20.0	50.0	46.7	60.9	0.75
AAD3	62.5	9.1	31.6	33.3	43.5	0.47
AAD1	75.0	22.2	47.1	46.2	57.1	0.57

accuracy of 63.8%, reflecting a good compromise between sensitivity and specificity. The reduced specificity is explained by the structural similarity between some active and inactive compounds and partial pharmacophoric matches in inactive substances that may contribute to misclassification. Chemically, the pharmacophore features identified such as hydrogen bond donors and aromatic regions align well with known structures of Mpro inhibitors, supporting the chemical plausibility of AAD2. These combined statistical and chemical validations establish AAD2 as the most reliable model among those tested, providing a solid basis for further external validation.

The TPR/FPR data were analyzed to further validate the predictive performance of the AAD2 pharmacophore hypothesis by assessing the impact of different fitness score thresholds on classification performance. The plot illustrates fluctuations in the true positive rate (TPR) and false positive rate (FPR) across a range of thresholds, with an ideal cutoff such as 1.5 identified as the point where a suitable balance between sensitivity (high TPR) and specificity (low FPR) is achieved. At this threshold, the model effectively distinguishes active compounds from inactives, minimizing misclassification of inactive molecules while accurately capturing active ones. This analysis highlights the robustness and resilience of the AAD2 model in providing reliable predictions across various cutoff points.

Selection of 3D pharmacophore hypothesis: AAD2 is the chosen pharmacophore hypothesis, which has three important parts: two hydrogen bond acceptors (A3 and A4) and one hydrogen bond donor (D8). These features show the spatial and electrical conditions that must be met for ligands to bind effectively to the active site of the SARS-CoV-2 major protease (Mpro). The distances and angles between these features determine how they are arranged in space. Fig. 3a shows that under the AAD2 theory, the distance from A3 to A4 is 4.14 Å, from A4 to D8 is 5.27 Å and from D8 to A3 is 3.15 Å. Fig. 3b shows that the angles between these characteristics were 91.5º for A3, 36.6º for A4 and 51.8º for D8.

The selection of the AAD2 hypothesis as the most suitable for virtual screening was based on its high sensitivity and reasonable balance of accurate metrics. Specifically, AAD2 accurately mapped 38 of 39 active ligands with a fitness score above 1.5, corresponding to a sensitivity of 97.4% and only misclassified one active as inactive (false negative). This indicates a strong ability to identify true actives, which is crucial in early-stage screening to avoid missing promising compounds. In terms of specificity, AAD2 correctly classified 13 out of 43 inactives as true negatives, with 30 misclassified as false positives, resulting in a specificity of 30.2%. While this low specificity indicates a higher false-positive rate, such a trade-off is often acceptable in pharmacophore-based virtual scree-

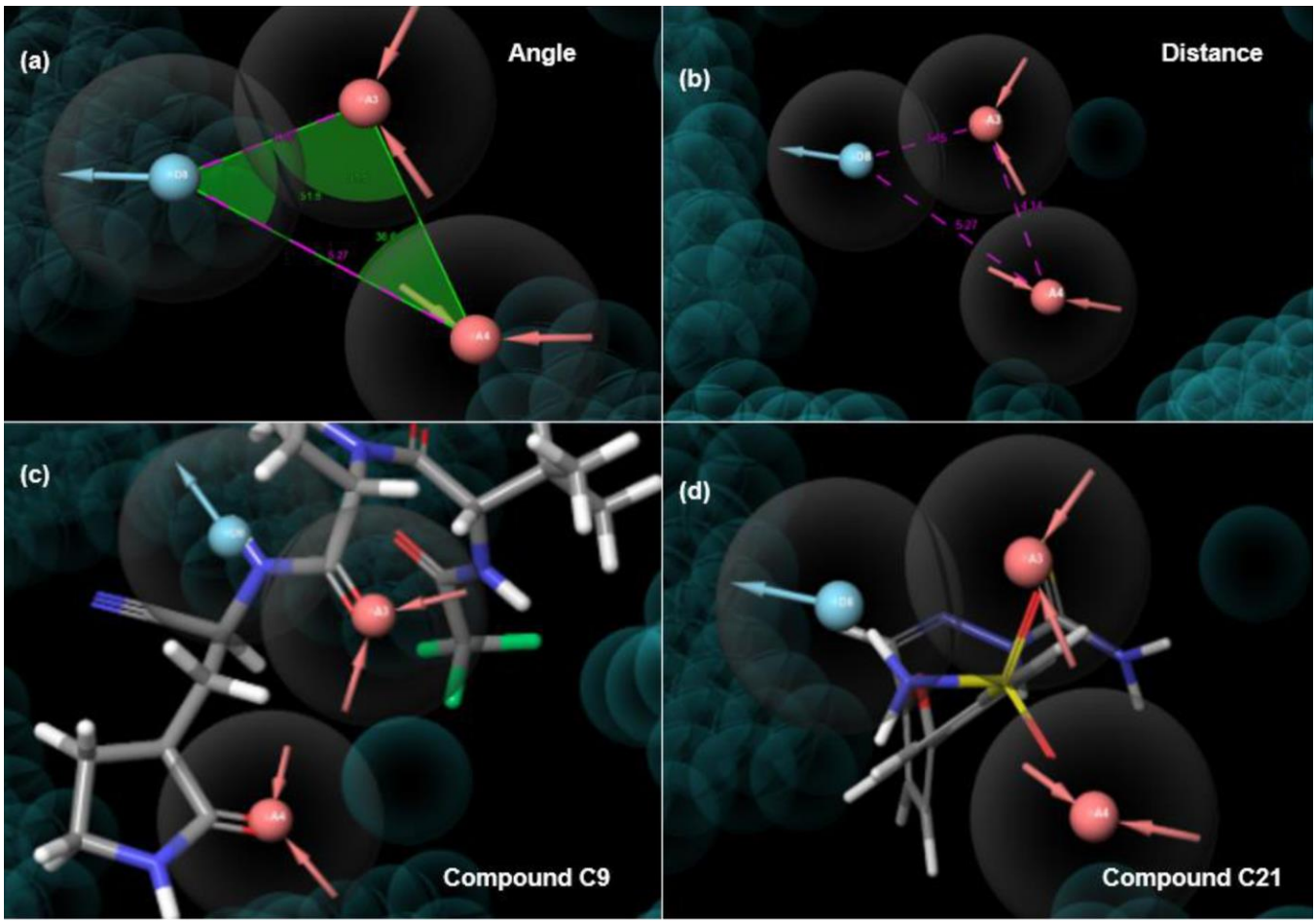

Fig. 3. Distance (a) and angle (b) measurements of the selected 3D pharmacophore hypothesis AAD2 (acceptor in red colour and donor in blue colour pharmacophore sites); best-fit (c) and least-fit (b) ligands from the external validation dataset

ning, where prioritizing sensitivity helps ensure that potential actives are not overlooked. Overall, the performance metrics (Table-6) of AAD2 demonstrate its effectiveness in early virtual screening efforts, emphasizing the importance of sensitivity over specificity in this context. The use of such a hypothesis facilitates identifying a broad set of candidate ligands for subsequent validation, aligning with the goal of minimizing false negatives during the initial drug discovery stages.

External validation of selected pharmacophore hypothesis: The external validation of the AAD2 pharmacophore hypothesis involved testing its ability to predict activity of ligands outside the training set. Twenty ligands were randomly selected and evaluated using the Phase™ module with AAD2, applying a threshold of 1.5 to fitness scores ligands scoring ≥ 1.5 were provisionally predicted as active, while those scoring < 1.5 were predicted to be inactive. The activation status was verified against experimental IC_{50} values, with ≤ 1000 nM defining active compounds. Among the predicted actives, 10 ligands (C1–C10) were correctly identified (true positives), whereas 4 (C12, C13, C16, C18) were false positives. Conversely, 5 ligands (C14, C15, C17, C21, C8) predicted as inactive were correctly labeled (true negatives), except for C8, which was incorrectly classified as inactive despite being active; this discrepancy arose from multiple conformers generated during screening, highlighting the importance of considering the highest scoring conformer for validation. The validation metrics, including sensitivity, specificity and ROC-AUC, were calculated using the specified formula, with the hypothesis demonstrating high sensitivity, moderate specificity and an excellent ROC-AUC of 1.0, indicating the model's strong potential for identifying active SARS-CoV-2 Mpro inhibitors.

During the external validation process, ligand C9 achieved the highest fitness score of 2.19 among the 20 evaluated ligands, indicating an excellent match with the pharmacophore model AAD2. The analysis demonstrated that C9's features aligned well with the three key components of the hypothesis: two hydrogen bond acceptors (A3 and A4) and one hydrogen bond donor (D8). Fig. 3c illustrates that C9 is positioned very close to the optimal spatial arrangement and possesses the necessary chemical groups required for effective interaction with the SARS-CoV-2 main protease, underscoring its potential as a promising inhibitor.

Ligand C21 exhibited the lowest fitness score of 1.14 during the external validation, indicating a poor fit with the pharmacophore model. Fig. 3d clearly demonstrates that the functional groups of C21 do not align appropriately with the hypothesized pharmacophoric features either they fail to reach the correct spatial locations or are oriented incorrectly. This poor alignment may result from the ligand lacking the appropriate chemical properties or from its structural conformation not accommodating the required 3D shape for optimal interaction. As observed, ligands with higher fitness scores generally show better spatial and chemical alignment, whereas those with lower scores correspond to weaker matches, supporting the validity of fitness scores as a reliable indicator of a compound's compatibility with the pharmacophoric criteria.

AAD2 was selected as the optimal pharmacophore hypothesis since it demonstrated superior performance across multiple evaluation metrics. Notably, AAD2 achieved the highest sensitivity at 97.4%, indicating it effectively identified nearly all active compounds. Despite its relatively low specificity at 30.2%, it still outperformed the other hypotheses in classifying actives accurately. The model's AUC was perfect at 1.0, reflecting its ability to perfectly discriminate between active and inactive compounds, a significant advantage over other models with AUCs ranging from 0.47 to 0.75. Moreover, AAD2 exhibited a precision of 55.9% and an F1-score of 71.0%, balancing the true positive rate while minimizing false positives. Its overall performance metrics suggest it is a reliable and robust model for virtual screening, especially for identifying potential SARS-CoV-2 inhibitors.

TABLE-6
EXTERNAL VALIDATION METRICS FOR AAD2 HYPOTHESIS

Validation metric	Formula	Value
Sensitivity [42]	$\frac{TP}{TP+FN}$	90.9%
Specificity [42]	$\frac{TN}{TN+FP}$	50.0%
Accuracy [43]	$\frac{TP+TN}{TP+TN+FP+FN}$	75.0%
Balanced accuracy [44]	$\frac{Sensitivity+Specificity}{2}$	70.4%
Precision [43]	$\frac{TP}{TP+FP}$	66.7%
F1 score [45]	$\frac{2\times Precision\times Sensitivity}{Precision+Sensitivity}$	76.9%
Matthews correlation coefficient (MCC) [46]	$\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	0.459%
Enrichment factor (EF) [46]	$\frac{\text{Actives found in Top X\%}}{\text{Expected actives by random selectin in Top X\%}}$	1.3
ROC-AUC	–	1.0

The spatial arrangement of features within AAD2 confirms that its pharmacophoric elements are well-distributed in three-dimensional space, allowing compatible ligands with suitable functional groups to bind effectively. Specifically, the distances between features A3 and A4 (4.14 Å), A4 and D8 (5.27 Å) and D8 and A3 (3.15 Å) are optimal for accommodating ligands with complementary groups. The angular geometries 91.5º for A3, 36.6º for A4 and 51.8º for D8 further define the shape and directionality necessary for binding. Correct alignment of these distances and angles ensures that only ligands with appropriate 3D orientation and chemical properties will match the pharmacophore, making AAD2 an effective template for virtual screening against SARS-CoV-2 targets.

Validation of the model using a compound set yielded 10 true positives, 1 false negative, 5 false positives and 5 true negatives. The sensitivity reached 90.9%, indicating strong capability in detecting active compounds, while the specificity was moderate at 50%. Overall accuracy was 75%, with a balanced accuracy of 70.4% and an MCC of 0.459 reflecting an average prediction quality. The F1-score of 0.769 and a precision of 66.7% suggest a favourable balance in predicting true actives without excessive false positives. Furthermore, the enrichment factor of 1.3 shows active compounds are somewhat prioritized compared to random screening. Importantly, the external ROC-AUC of 1.0 underscores its excellent discriminatory power, confirming the reliability of AAD2 for virtual screening efforts targeting COVID-19 therapeutics. The ligands tested against the AAD2 pharmacophore exhibited a strong correlation between their fit scores and inhibitory potency (IC_{50} values). Ligands with fitness scores above 2.0 such as cyano-hexane (C9, score 2.15) with an IC_{50} of 760 nM and ureido-butyramide (C6, score 2.04) with an IC_{50} of 26 nM possess functional groups capable of forming robust hydrogen bonds, aligning well with the pharmacophore features. In contrast, ligands with lower fitness scores (< 2.0), such as hydroxyquinazoline (C12, score 1.82), which has an IC_{50} of 54,000 nM, tend to contain bulkier or rigid groups like aromatic rings or sulfonamides. These groups can hinder the optimal binding or fail to form effective hydrogen bonds, leading to poor pharmacophoric alignment and reduced activity. This relationship underscores that high fit scores correspond to better chemical complementarity and stronger inhibitory activity, validating the pharmacophore model as a predictive tool for identifying potent SARS-CoV-2 inhibitors.

Conclusion

In summary, the 3D-QSAR model for benzodiazepine derivatives, using PLS 5, shows strong predictive power with key structural features identified through contour maps. Future improvements could include expanding the dataset, employing advanced machine learning and conducting *in vitro* validation. The 3D pharmacophore model (AAD2) exhibits excellent discrimination (ROC-AUC 1.0, sensitivity 97.4%), with two hydrogen bond acceptors and one donor as crucial features for identifying potent SARS-CoV-2 main protease inhibitors. Small, polar groups enhance binding, while bulky groups reduce fitness. Its strong correlation with IC_{50} highlights its efficiency as a screening tool. Future directions include adding targets like PLpro and nsp16, integrating deep learning and experimental validation. These strategies support pandemic preparedness and antiviral drug discovery.

ACKNOWLEDGEMENTS

The authors are thankful to the IMU University's Vice Chancellor, Dean, School of Pharmacy; Dean, School of Postgraduate Studies; Director and Deputy Directors of Institute for Research, Development & Innovation (IRDI) for providing the facilities to complete the proposed research project. This research project was funded by IMU University Joint-Committee on Research & Ethics, Project ID No. BSc (Hons.) in Pharmaceutical Chemistry: BPC I-2024(06).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this article.

REFERENCES

1. Y.M. Bar-On, A. Flamholz, R. Phillips and R. Milo, *eLife*, **9**, e57309 (2020); https://doi.org/10.7554/eLife.57309
2. B. Hu, H. Guo, P. Zhou and Z.L. Shi, *Nat. Rev. Microbiol.*, **19**, 141 (2021); https://doi.org/10.1038/s41579-020-00459-7.
3. D.D. Singh, A. Parveen and D.K. Yadav, *Front. Cell Infect. Microbiol.*, **11**, 777212 (2021); https://doi.org/10.3389/fcimb.2021.777212
4. A. Schuhmacher, O. Gassmann and M. Hinder, *J. Transl. Med.*, **14**, 105 (2016); https://doi.org/10.1186/s12967-016-0838-4
5. S.A. Khedkar, A.K. Malde, E.C. Coutinho and S. Srivastava, *Med. Chem.*, **3**, 187 (2007); https://doi.org/10.2174/157340607780059521
6. J. Zhang, T. Xiao, Y. Cai and B. Chen, *Curr. Opin. Virol.*, **50**, 173 (2021); https://doi.org/10.1016/j.coviro.2021.08.010
7. S. Kwon, H. Bae, J. Jo and S. Yoon, *BMC Bioinform.*, **20**, 521 (2019); https://doi.org/10.1186/s12859-019-3135-4
8. S. Ren, X. Liu, Y. Huang, C. Zhu, W. Zhu and H. Dong, *J. Mol. Struct.*, **1321**, 140286 (2025); https://doi.org/10.1016/j.molstruc.2024.140286
9. M.S. Bahia, O. Kaspi, M. Touitou, I. Binayev, S. Dhail, J. Spiegel, N. Khazanov, A. Yosipof and H. Senderowitz, *Mol. Inform.*, **42**, 2200186 (2023); https://doi.org/10.1002/minf.202200186
10. A. Tropsha, O. Isayev, A. Varnek, G. Schneider and A. Cherkasov, *Nat. Rev. Drug Discov.*, **23**, 141 (2024); https://doi.org/10.1038/s41573-023-00832-0
11. D. Giordano, C. Biancaniello, M. A. Argenio and A. Facchiano, *Pharmaceuticals*, **15**, 578 (2022); https://doi.org/10.3390/ph15050578
12. D.B. Kitchen, H. Decornez, J.R. Furr and J. Bajorath, *Nat. Rev. Drug Discov.*, **3**, 935 (2004); https://doi.org/10.1038/nrd1549
13. V. M. Balaramnavar, K. Ahmad, M. Saeed, I. Ahmad, M. Kamal and T. Jawed, *RSC Adv.*, 10, 40264 (2020); https://doi.org/10.1039/D0RA06038K
14. W.-J. Shin and B.L. Seong, *Expert Opin. Drug Discov.*, **8**, 411 (2013); https://doi.org/10.1517/17460441.2013.767795
15. A. Kutlushina, A. Khakimova, T. Madzhidov and P. Polishchuk, *Molecules*, **23**, 3094 (2018); https://doi.org/10.3390/molecules23123094
16. K.A. Qusay, A. Vasudeva Rao, H. Husniza and Z.T. Najeeb, *Res. J. Biotechnol.*, **16**, 50 (2021).

17. F. Wang, R. Zeng, J. Qiao, A. Xia, Y. Li, F. Li, Y. Wu, Y. Liu, X. Zhao, J. Lei and S. Yang, *Bioorg. Med. Chem. Lett*., **92**, 129407 (2023); https://doi.org/10.1016/j.bmcl.2023.129407
18. C. Fischer, T.J. Van Oers, M.J. van Belkum, T. Lamer, A. Romney, P. Chen, M.J. Lemieux and J.C. Vederas, *RSC Adv*., **14**, 35438 (2024); https://doi.org/10.1039/D4RA06573E
19. C.A. Lipinski, F. Lombardo, B.W. Dominy and P.J. Feeney, *Adv. Drug Deliv. Rev*., **23**, 3 (1997); https://doi.org/10.1016/S0169-409X(96)00423-1
20. N. Kitamura, M.D. Sacco, C. Ma, Y. Hu, J.A. Townsend, X. Meng, F. Zhang, X. Zhang, M. Ba, T. Szeto, A. Kukuljac, M.T. Marty, D. Schultz, S. Cherry, Y. Xiang, Y. Chen and J. Wang, *J. Med. Chem.*, **65**, 2848 (2022); https://doi.org/10.1021/acs.jmedchem.1c00509
21. C. Ma, Z. Xia, M.D. Sacco, Y. Hu, J.A. Townsend, X. Meng, J. Choza, H. Tan, J. Jang, M.V. Gongora, X. Zhang, F. Zhang, Y. Xiang, M.T. Marty, Y. Chen and J. Wang, *J. Am. Chem. Soc.*, **143**, 20697 (2021); https://doi.org/10.1021/jacs.1c08060
22. Y.R. Alugubelli, Z.Z. Geng, K.S. Yang, N. Shaabani, K. Khatua, X.R. Ma, E.C. Vatansever, C.-C. Cho, Y. Ma, J. Xiao, L.R. Blankenship, G. Yu, B. Sankaran, P. Li, R. Allen, H. Ji, S. Xu and W.R. Liu, *Eur. J. Med. Chem.*, **240**, 114596 (2022); https://doi.org/10.1016/j.ejmech.2022.114596
23. Y. Ma, K.S. Yang, Z.Z. Geng, Y.R. Alugubelli, N. Shaabani, E.C. Vatansever, X.R. Ma, C.-C. Cho, K. Khatua, L.R. Blankenship, J. Xiao, G. Yu, B. Sankaran, P. Li, R. Allen, H. Ji, S. Xu and W. R. Liu, *Eur. J. Med. Chem*., **240**, 114570 (2022); https://doi.org/10.1016/j.ejmech.2022.114570
24. S. Gao, L. Song, K. Sylvester, B. Mercorelli, A. Loregian, K. Toth, R.H. Weiße, A. Useini, N. Sträter, M. Yang, B. Ye, A.E. Tollefson, C.E. Müller, X. Liu and P. Zhan, *J. Med. Chem.*, **66**, 16426 (2023); https://doi.org/10.1021/acs.jmedchem.3c01876
25. C. Ngo, W. Fried, S. Aliyari, J. Feng, C. Qin, S. Zhang, H. Yang, J. Shanaa, P. Feng, G. Cheng, X.S. Chen and C. Zhang, *J. Med. Chem.*, **66**, 12237 (2023); https://doi.org/10.1021/acs.jmedchem.3c00810
26. K. Zhang, T. Wang, M. Li, M. Liu, H. Tang, L. Wang, K. Ye, J. Yang, S. Jiang, Y. Xiao, Y. Xie, M. Lu and X. Zhang, *Eur. J. Med. Chem.*, **257**, 115487 (2023); https://doi.org/10.1016/j.ejmech.2023.115487
27. L. Brier, H. Hassan, X. Hanoulle, V. Landry, D. Moschidi, J. Dumont, L. Desmarets, Y. Rouillé, A. Herledan, S. Warenghem, C. Piveteau, P. Carré, S. Ikherbane, F.-X. Cantrelle, E. Dupré, J. Dubuisson, S. Belouzard, F. Leroux, B. Deprez and J. Charton, *Eur. J. Med. Chem*., **250**, 115186 (2023); https://doi.org/10.1016/j.ejmech.2023.115186
28. C.S. Dampalla, A.D. Rathnayake, A.C. Galasiti Kankanamalage, Y. Kim, K.D. Perera, H.N. Nguyen, M.J. Miller, T.K. Madden, H.R. Picard, H.A. Thurman, M.M. Kashipathy, L. Liu, K.P. Battaile, S. Lovell, K.-O. Chang and W.C. Groutas, *J. Med. Chem.*, **65**, 7818 (2022); https://doi.org/10.1021/acs.jmedchem.2c00224
29. J.-W. Zhang, Y. Xiong, F. Wang, F.-M. Zhang, X. Yang, G.-Q. Lin, P. Tian, G. Ge and D. Gao, *Eur. J. Med. Chem.*, **228**, 114030 (2022); https://doi.org/10.1016/j.ejmech.2021.114030
30. C.S.B. Chia and M. Inoue, *ACS Med. Chem. Lett.*, **13**, 1394 (2022); https://doi.org/10.1021/acsmedchemlett.2c00349
31. X. Dou, Q. Sun, G. Xu, Y. Liu, C. Zhang, B. Wang, Y. Lu, Z. Guo, L. Su, T. Huo, X. Zhao and C. Wang, *Eur. J. Med. Chem.*, **238**, 114508 (2022); https://doi.org/10.1016/j.ejmech.2022.114508
32. H. Wang, R. Pei, X. Li, W. Deng, S. Xing, Y. Zhang, C. Zhang, S. He, H. Sun, S. Xiao, J. Xiong, Y. Zhang, X. Chen, Y. Wang, Y. Guo, B. Zhang and L. Shang, *Eur. J. Med. Chem.*, **238**, 114458 (2022); https://doi.org/10.1016/j.ejmech.2022.114458
33. D. Deodato, N. Asad and T.M. Dore, *Bioorg. Med. Chem. Lett.*, **72**, 128867 (2022); https://doi.org/10.1016/j.bmcl.2022.128867
34. Y.B. Ryu, H.J. Jeong, J.H. Kim, Y.M. Kim, J.-Y. Park, D. Kim, T.T. H. Nguyen, S.-J. Park, J.S. Chang, K.H. Park, M.-C. Rho and W.S. Lee, *Bioorg. Med. Chem*., 18, 7940 (2010); https://doi.org/10.1016/j.bmc.2010.09.035
35. G. Arthur, W. Oliver, B. Klaus, S. Thomas, I. Gökhan, B. Sharon, T. Isabelle, D. Pierre and L. Thierry, *Front. Mol. Biosci*., **7**, 599059 (2020); https://doi.org/10.3389/fmolb.2020.599059
36. J.C. Shelley, A. Cholleti, L.L. Frye, J.R. Greenwood, M.R. Timlin and M. Uchimaya, *J. Comput. Aided Mol. Des.*, **21**, 681 (2007); https://doi.org/10.1007/s10822-007-9133-z
37. S.Y. Yang, *Drug Discov. Today*, **15**, 444 (2010); https://doi.org/10.1016/j.drudis.2010.03.013
38. A. Vuorinen and D. Schuster, *Methods*, **71**, 113 (2015); https://doi.org/10.1016/j.ymeth.2014.10.013
39. N. Moussa, A. Hassan and S. Gharaghani, *Heliyon*, **7**, e06605 (2021); https://doi.org/10.1016/j.heliyon.2021.e06605
40. D. Giordano, C. Biancaniello, M.A. Argenio and A. Facchiano, *Pharmaceuticals*, **15**, 646 (2022); https://doi.org/10.3390/ph15050646
41. S. Parveen, L. Shahbaz, N. Shafiq, M. Rashid, M. Mohany and M. Zhu, *RSC Adv.*, **15**, 2045 (2025); https://doi.org/10.1039/D4RA06536K
42. K. Chu, *Emerg. Med*., **11**, 175 (1999); https://doi.org/10.1046/j.1442-2026.1999.00041.x
43. G. Varoquaux and O. Colliot, in eds.: O. Colliot, Evaluating Machine Learning Models and their Diagnostic Value, In: Machine Learning for Brain Disorders, New York, NY, USA: Humana, Chap. 20 (2023).
44. Machine Learning Model Evaluation Metrics: Accuracy, Precision, Recall, F1; https://www.sanfoundry.com/machine-learning-model-evaluation-metrics-accuracy-precision-recall-f1/?utm_source=chatgpt.com (Accessed on 12 March 2025).
45. D. Chicco and G. Jurman, *BMC Genomics*, **21**, 6 (2020); https://doi.org/10.1186/s12864-019-6413-7
46. D. Arnal, C. Moya, L. Filippelli, J. Segura-Garcia and S. Maicas, *BioData Min*., **16**, 1 (2023); https://doi.org/10.1186/s13040-022-00318-6