COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Docking-generated multiple ligand poses for bootstrapping bioactivity classifying Machine Learning: Repurposing covalent inhibitors for COVID-19-related TMPRSS2 as case study

Ma'mon M. Hatmal [a], Omar Abuyaman [a], Mutasem Taha [b,*]

[a] Department of Medical Laboratory Sciences, Faculty of Applied Medical Sciences, The Hashemite University, PO Box 330127, Zarqa 13133, Jordan
[b] Department of Pharmaceutical Sciences, Faculty of Pharmacy, University of Jordan, Amman 11942, Jordan

ARTICLE INFO

Article history:
Received 22 March 2021
Received in revised form 3 August 2021
Accepted 16 August 2021
Available online 19 August 2021

Keywords:
Docking
Scoring
Bootstrapping
Ligand-receptor contact fingerprints
Machine learning
Covalent docking

ABSTRACT

In the present work we introduce the use of multiple docked poses for bootstrapping machine learning-based QSAR modelling. Ligand-receptor contact fingerprints are implemented as descriptor variables. We implemented this method for the discovery of potential inhibitors of the serine protease enzyme TMPRSS2 involved the infectivity of coronaviruses. Several machine learners were scanned, however, Xgboost, support vector machines (SVM) and random forests (RF) were the best with testing set accuracies reaching 90%. Three potential hits were identified upon using the method to scan known untested FDA approved drugs against TMPRSS2. Subsequent molecular dynamics simulation and covalent docking supported the results of the new computational approach.

1. Introduction

Docking algorithms explore the conformational space of bound ligand(s) by sampling numerous ligand conformations/poses within the targeted binding site. The binding enthalpy of each docked pose/conformation is evaluated employing force fields that calculate interactions between ligand-protein complementary groups [1–4]. However, docking engines ignore entropic contributions in binding and therefore need to be guided by scoring functions to select realistic docked poses [4–33]. Modern docking methods can reproduce an experimental ligand crystallographic pose within high-ranking solutions. However, scoring functions are usually unable to evaluate binding free energies to accurately rank docked poses [1,3,34–37].

The concept of Ligand-Receptor Contacts Fingerprints (LRCFs) is well established [38]. It proceeds by either mapping out all binding site atoms that contact a list of docked potent ligands and evade inactive compounds [38–41], or identify binding site atoms that frequently contact a particular bound ligand during molecular dynamics or related simulations [42–44]. Significant binding site contacts can be transformed into pharmacophores [38–44]. A related concept to LRCFs is the interaction fingerprints (IFPs) [45–51], which were used as postdock equivalents to scoring functions [45,46,52] or as means for virtual screening by similarity search [46,49,53,54].

Machine learning (ML) in drug discovery involves the implementation of statistical means for learning and predicting molecular properties [55–62]. The following are popular ML algorithms used in computer-aided drug design and discovery and were evaluated in the current research: Random Forest (RF) [61,63]; Naïve Bayesian (NB) [64–68]; eXtreme Gradient Boosting (XGBoost) [69,70]; K-nearest neighbors (kNN) [71]; Support vector machine (SVM) [72,73]; probabilistic neural networks (PNN) [74–79]; and multilayer perceptron (MLP) [80,81]. However, ML needs to be validated vis-à-vis statistical accuracy. Bootstrapping (BS) is a commonly used statistical approach to assign accuracy values for ML models. BS is a resampling method that uses random sampling with replacement to assign accuracy measurements to sample estimates particularly in situations of limited training data [82–84].

We herein propose to use multiple docked poses (up to hundreds per ligand), generated by a number of docking engines/scoring functions for a list of active and inactive ligands, to bootstrap bioactivity ML classifiers using LRCFs descriptors. The rational is based on a plausible assumption: Since docking algorithms implement reasonable enthalpy estimating methods, then all docked conformers/poses are enthalpically plausible, particularly for potent ligands. Therefore, ML-based convergence, among numer-

[*] Corresponding author.
E-mail address: mutasem@ju.edu.jo (M. Taha).

https://doi.org/10.1016/j.csbj.2021.08.023

ous docked poses/conformers of potent ligands, on certain group of atom contacts within the binding site highlights the significance of those contacts and their collective ability to classify docked virtual screening hits, i.e., as active or inactive. This method is reminiscent of 4D-QSAR, which uses molecular dynamics simulations to create conformation-dependent molecular descriptors [179]. Still, our approach is less computationally extensive and more suited for virtual screening purposes.

We applied this innovative concept, i.e., bootstrapping ML with multiple docked poses, to virtually screen and repurpose known drug molecules against the COVID-19-related enzyme transmembrane serine protease 2 (TMPRSS2).

TMPRSS2 is a cell surface serine protease involved in priming entry proteins necessary for respiratory viruses, e.g., influenza and COVID19, to cross cellular membranes [85–94]. Several reversible [90] and irreversible [95,96] inhibitors have been reported for TMPRSS2 [86,88–92], e.g., nafamostat and camostat [97–102].

TMPRSS2 had no known crystallographic structure until recently (PDB code: 7MEQ, Released 2021–04-21), nevertheless, it was successfully modeled based on its close homologue hepsin (PDB code 5ce1) [92]. Incidentally, our proposed approach of bootstrapping ML with multiple docked poses is expected to be particularly useful in such a case where ligand–receptor crystallographic structure is unavailable to help identify critical ligand-receptor binding interactions.

Irreversible "covalent" inhibitors usually include reactive electrophilic "warheads" capable of reacting with nucleophilic center (s) within targeted binding sites [103–106]. However, for successful covalent inhibition, irreversible bond formation should be preceded by reversible recognition [107,108] that specifically allows the warhead close proximity to the targeted nucleophilic center (Ser478 in TMPRSS2) [107,108]. In fact, presence of reactive warhead in particular inhibitor represents an added value to enhance bioactivity rather than an essential precondition for bioactivity. This is why we were prompted to use docking tools developed for reversible binders to bootstrap and generate ML models intended for the identification of covalent binders, as shown below.

The following are reported warheads against serine hydrolases. (i) β-lactam rings [106,109]. (ii) Carbamates [106,109–111] (iii) Nitriles [112,113]. Two famous nitrile-based protease covalent inhibitors are vildagliptin and saxagliptin [114,115].(iv) Michael acceptors [107,116–119]. (v) Nitro groups [120–132]. (vi) Aromatic esters. [97,133–136]. Table 1 shows the different warheads and appropriate example on each.

The computational workflow in this project is shown in Fig. 1. Firstly, a diverse set of known inhibitors are docked into TMPRSS2 binding pocket using 3 docking engines and 9 scoring functions. The docked ligands include active and inactive compounds, i.e., against TMPRSS2, of reversible and irreversible binding capacity. Reversible and irreversible active inhibitors are rather easy to define. However, inactive reversible and irreversible inhibitors warrant more explanation: These are molecules with or without covalent warheads, respectively, which were explicitly reported as being inactive against TMPRSS2.

The collected compounds were ionized, tautomerized and docked/scored into TMPRSS2. Identical or closely similar docked poses were filtered out using proper RMSD filter. Reasonable docked poses (as judged by plausible consensus among docking engines and scoring functions) were split into training and testing sets for ML. Several ML methods were evaluated and the best were used to predict the TMPRSS2 bioactivity classification of docked poses of FDA-approved drugs. Promising hits were further evaluated by covalent docking and molecular dynamics simulations. Three hits were found to be potential potent inhibitors of TMPRSS2, namely, capreomycin, aspoxicillin and fosamprenavir.

2. Material and methods

2.1. Data collection

The literature was carefully searched for TMRPSS2 inhibitors. The search identified 1064 TMPRSS2 inhibitors that can be unequivocally classified as "active" ($IC_{50} \leq 1000$ nM) or "Inactive" ($IC_{50} \geq 10{,}000$ nM). Details about these compounds are as follows: Crystallographic ligands bound to the close homologue of TMPRSS2, human hepsin, were collected and considered. In one crystallographic complex, namely, 5ce1, the ligand is reported to have anti-TMPRSS2 IC_{50} of 100 nM [137] and was therefore considered as "active, while in another crystallographic complex, namely, 1p57, the ligand is reported to have IC_{50} value of 40,000 nM, and accordingly was considered as "Inactive". Additionally, a group of 99 reversible TMPRSS2 inhibitors were collected [90], out of which 88 compounds were reported to have Ki values $\leq$ 1000 nM, and were therefore categorized as "actives", while 2 compounds were reported with Ki values $\geq$ 10,000 nM and were categorized as "inactives". Two additional covalent-bond forming TMPRSS2 inhibitors were also included as "actives", namely, nafamostat (IC_{50} = 100 nM) and camostat (IC_{50} = 1000 nM) [97]. Both contain aromatic ester warheadsand inhibit TMPRSS2 by covalent bonding [138,139]. Additionally, 972 established "inactives" were collected [97], out of which 150 were equipped with covalent warheads (30 nitro compounds, 29 Michael acceptors, 24 nitriles, 35 carbamates, 23 β-lactams and 9 aromatic esters).

However, to reduce the number of modeled compounds, it was decided to use principal component analysis (PCA) to visually select diverse representative molecules from the collected compounds as in Fig. 2.

This shortlisted the modelled compounds into 107 divided into 15 "actives" (of which two are irreversible and 13 reversible inhibitors) and 92 "inactives" (of which 12 have nitro warheads, 9 with nitrile warheads, 6 Michael acceptors, 18 carbamates, 16 β-lactams, 2 aromatic esters, one α-keto-amide and 26 compounds lacking any reactive warheads). Table 2 shows the chemical structures of the modeled compounds and their reported bioactivities (if available) [90,97,137,140].

Prior to docking, the modelled compounds were appropriately ionized using the Prepare Ligands protocol in Discovery Studio (version 4.5) assuming pH range of 6.5–8.5. Additionally, the same protocol was also used to generate tautomeric forms of each compound. This yielded 338 tautomeric forms (30 for the "active" inhibitors and 308 for the "inactive" members) for subsequent docking studies.

2.2. Homology modelling

2.2.1. Template search

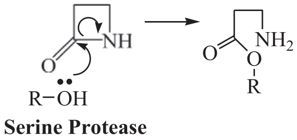
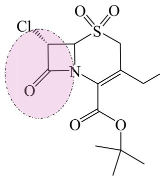
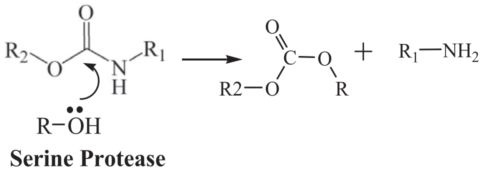
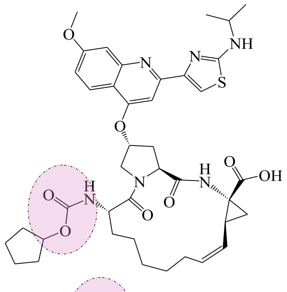
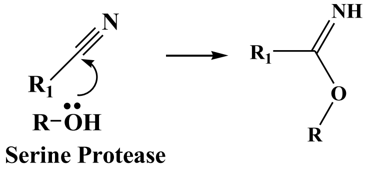
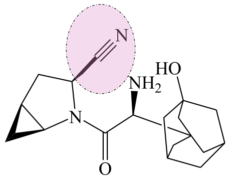
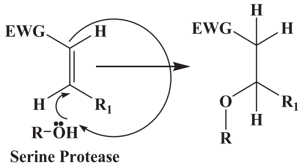
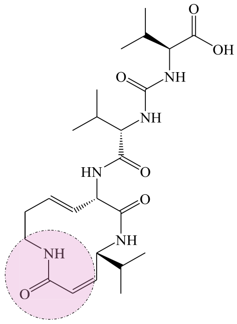
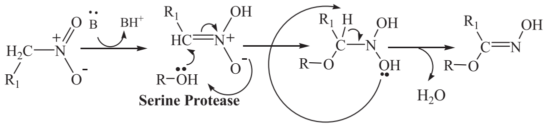
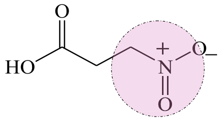
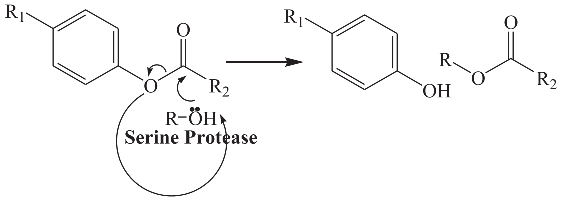
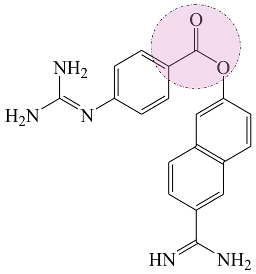
Template search with BLAST and HHBlits has been performed against the SWISS-MODEL template library (last update: 2020-04-15, last included PDB release: 2020-04-10) [141,142]. A total of 1166 templates were found.

2.2.2. Model building

Models are built based on the target-template alignment using ProMod3 [143]. Coordinates which are conserved between the target and the template are copied from the template to the model. Insertions and deletions are remodelled using a fragment library. Side chains are then rebuilt. Finally, the geometry of the resulting model is regularized by using a force field. In case loop modelling with ProMod3 fails, an alternative model is built with PROMOD-II [144].

Table 1

Major warheads used for covalent inhibition of serine proteases, their mechanisms of action, and examples on each.

Warhead	Nucleophile-Electrophile Chemistry	Example		Target
		Structure	Name	
β-Lactam			L-647957	Elastase
Carbamate			Ciluprevir	NS3/4A protease (Hepatitis C)
Nitrile			Saxagliptin	Dipeptidyl Peptidase 4
Michael Acceptors	 EWG: Electron withdrawing group		Syringolin A	β5 chymotrypsin-like proteasomal subunit
Nitro			3-Nitro-propionate	Isocitrate lyase
Aromatic ester			Nafamostat	Prostasin

2.2.3. Ligand modelling

The ligand present in the template structure (5ce1) was transferred by homology to the TMPRSS2 model because it satisfies the requested criteria by SWISS-MODEL: (a) The ligand is annotated as biologically relevant Hepsin inhibitor, (b) the ligand is in contact with the model, (c) the ligand is not clashing with the protein, (d) the residues in contact with the ligand are conserved between the target and the template.

2.2.4. Model quality estimation

The global and per-residue model quality has been assessed using tools implemented in SWISSS-MODEL including: QMEAN (Qualitative Model Energy Analysis) score and MolProbity score [145]. The former is a composite of 6 energy values within the homology model matrix related to protein nativeness with score values $\leq$ 4.0 indicating poor quality homology models. The later, on the other hand, combines several protein parameters including:

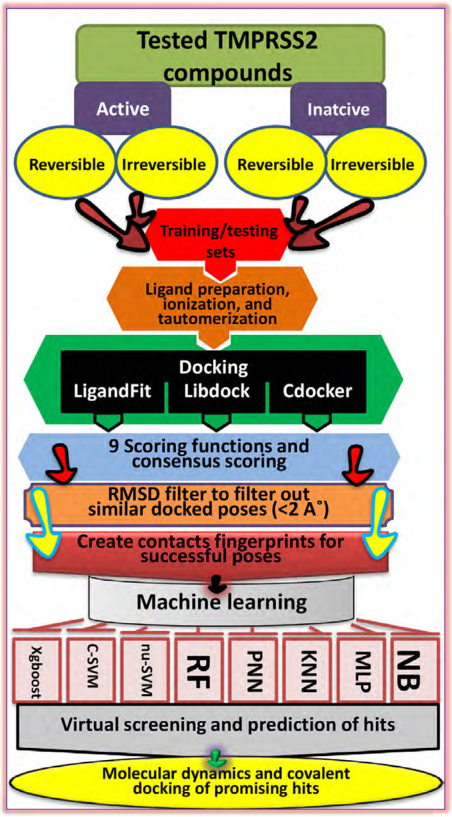

Tested TMPRSS2 compounds
Active
Inatcive
Reversible
Irreversible
Reversible
Irreversible
Training/testing sets
Ligand preparation, ionization, and tautomerization
Docking
LigandFit
Libdock
Cdocker
9 Scoring functions and consensus scoring
RMSD filter to filter out similar docked poses (<2 A˚)
Create contacts fingerprints for successful poses
Machine learning
Xgboost
C-SVM
nu-SVM
RF
PNN
KNN
MLP
NB
Virtual screening and prediction of hits
Molecular dynamics and covalent docking of promising hits

Fig. 1. Computational workflow implemented in the current project.

clash score, Ramachandran Plot criteria (Ramachandran Favored and Ramachandran Outliers), rotamer outliers, C-Beta deviations (geometry problems around alpha-carbons), bad bonds, and bad angles [146].

2.3. Docking

The collected set of inhibitors (**1**–**107**, Table 1) were docked into the homology model of TMPRSS2 using 3 docking engines: LibDock [26,147], LigandFit [148] and CDOCKER [149,150]. The binding pocket was defined as the cavity volume occupied by hepsin ligand within the homology model. Docking details are found in Supplementary Sections SM1-SM3.

2.3.1. Scoring of docked poses

Highest ranking docked conformers/poses generated by LibDock, LigandFit, and CDOCKER were scored using 9 scoring functions: Jain [6,15], LigScore1, LigScore2 [148], PLP1, PLP2 [151], PMF, PMF04 [152,153], -CDOCKER Energy and -CDOCKER Interaction Energy [149].

LigScore1 and LigScore2 scores were calculated employing CFF force field (version 1.02) and using grid-based energies with a grid extension of 7.5 Å across the binding site. PMF scores were calculated employing cutoff distances of 12.0 Å for carbon-carbon interactions and other atomic interactions, while PMF04 scores were calculated employing cutoff values of 6.0 and 9.0 Å for carbon-carbon interactions and other atomic interactions, respectively. -CDOCKER Energy and -CDOCKER Interaction Energy were calculated using Momany-Rone ligand partial charge method.

It was decided to select docked conformers/poses based on consensus among the 9 scoring functions [154,155]. The consensus function assigned a value of 1 for any molecular pose ranked within the highest 20% by the particular scoring function; otherwise, it was assigned a zero value (i.e., fit was within the lowest 80%). Subsequently, the consensus function summed up the scores for each molecular pose/conformer and ranked the molecular orientations. Docked poses of a particular ligand that achieved consensus among at least 4 scoring functions were selected for subsequent processing.

2.3.2. RMSD filtering

The RMSD filter implemented in Discovery Studio 4.5 was used. This filter calculates the Root Mean Square Deviation (RMSD) of ligand poses (in Å). Only heavy atoms were included for RMSD calculation (i.e., hydrogen atoms were excluded). RMSD values were calculated with respect to all docked poses of a particular compound. Poses with an RMSD <2.0 Å were considered duplicates of which only the one having higher consensus score was retained.

2.4. Ligand-Receptor fingerprints

The docked poses/conformers of each modelled compound were evaluated to identify their closest binding site atoms. A binding site atom that occurs within $\leq$ 2.5 Å of any atom within docked ligand pose is allocated an intermolecular contact value of "one", otherwise it is given a contact value of "zero". Distance evaluations were automatically performed employing an *in-house* made FORTRAN package. Eventually, a 2D matrix is built where each row corresponds to docked ligands poses and each column corresponds to different binding site atom. The matrix is filled with binary code, whereby "zeros" correspond to inter-atomic distances >2.5 Å and "ones" for distances binding site atoms at distances $\leq$ 2.5 Å.

2.5. Machine learning

Seven orthogonal ML were scanned, namely, RF, XGBoost, kNN, PNN, SVM, NB, and MLP.

2.5.1. Random Forest (RF)

RF is a multipurpose ML strategy for classification based on ensemble of Decision Trees (DTs) [61]. Each tree predicts a classification independently and "votes" for the related class. Most of the votes decide the overall RF predictions [70]. We implemented RF learner node within KNIME Analytics Platform (Version 4.1.3) with the following settings: Splitting criterion is the Information Gain Ratio (which normalizes the standard information gain by the split entropy to overcome any unfair preference for nominal splits with many child nodes), Number of trees = 100. No limitations were imposed on the number of levels or minimum node size. The accuracy was calculated using out-of-bag internal validation.

2.5.2. eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting (XGBoost, or XGB) relies on the ensemble of weak DT-type models to create boosted DT-type models [69,70,159]. We implemented the XGBoost Learner node within KNIME Analytics Platform (Version 4.1.3) with the following settings: Tree booster was implemented with depth wise grow policy, boosting rounds = 100, Eta = 0.3, Gamma = 0, maximum depth = 6, minimum child weight = 1, maximum delta step = 0, subsampling rate = 1, column sampling rate by tree = 1, column sampling rate by level = 1, lambda = 1, Alpha = 0, sketch epsilon = 0.03, scaled position weight = 1. Maximum number of bins = 256, Sample type (uniform), Normalize type (tree), and Dropout rate = 0.

2.5.3. k-Nearest Neighbors (kNN)

The kNN classifier depends on a distance learning methodology that calculates the activity value of an unknown member based on the bioactivities of a certain number (k) of nearest neighbors

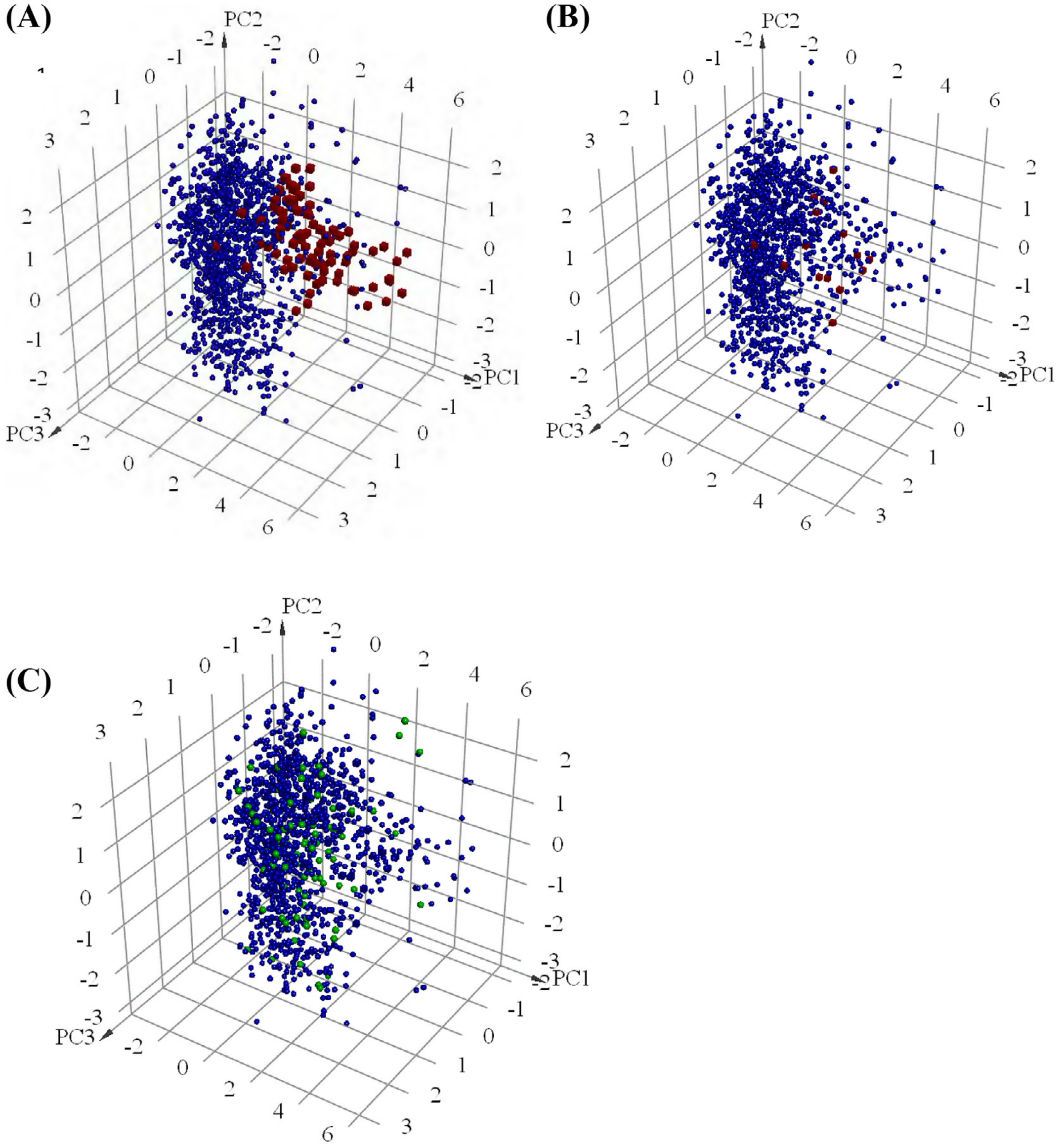


Fig. 2. Three-dimensional plots showing three main principal components based on 8 physicochemical descriptors (LogP, Molecular Weight, hydrogen bond donors, hydrogen bond acceptors, Rotatable Bonds, Number of Rings, Number of Aromatic Rings, Molecular Fractional Polar Surface Area) calculated for **(A)** All collected "actives" (red cubes, ■) compared to all established "inactives" (blue spheres, ●), **(B)** "actives" used in current modelling (red cubes, ■) compared to all collected compounds (blue spheres, ●), (C) "inactives" used in current modelling (green spheres, ●), compared to all collected compounds (blue spheres, ●). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(kNNs) in the training set. In this classifier, the similarity is measured by a distance metric [160]. We implemented kNN Learner node within KNIME Analytics Platform (Version 4.1.3) with k scanned from 3 to 6.

2.5.4. Probabilistic neural network (PNN)

Trains a probabilistic neural network (PNN) based on the Dynamic Decay Adjustment method on labeled data using Constructive Training of Probabilistic Neural Networks as the underlying algorithm [161,162]. We implemented PNN Learner node within KNIME Analytics Platform (Version 4.1.3) using PNN theta minus = 0.2 and theta plus = 0.4 and without specifying maximum number of epochs so that the PNN process is repeated until stable rule model is achieved.

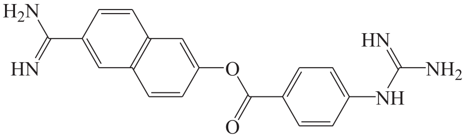
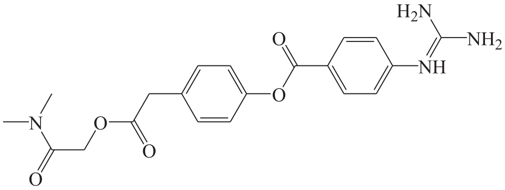
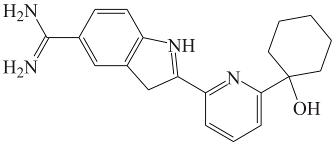
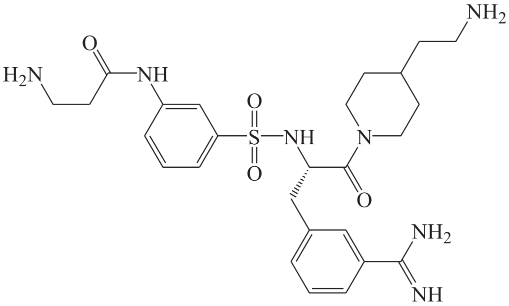
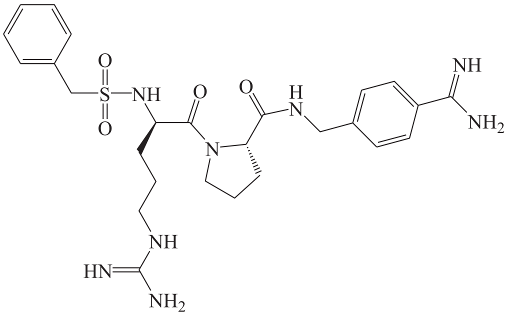
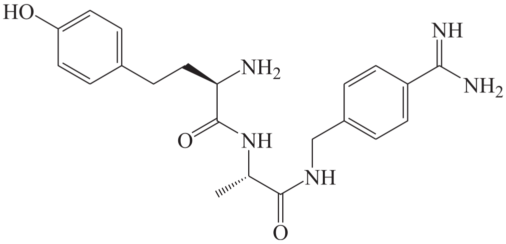
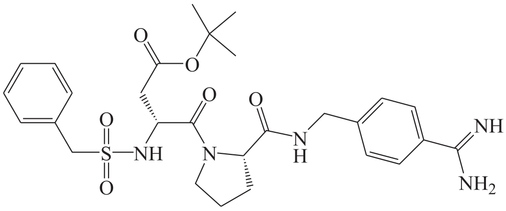
2.5.5. Naïve Bayesian (NB)

NB is a simple classifier whereby class labels are predicted and assigned to external observations based on vectors of descriptors for some finite set of training observations. NB classifier assumes each descriptor to contribute independently to the probability that certain observation (e.g., compound) belongs to a particular class (e.g., active or inactive) [163,164]. The probability of certain observation to belong to certain class is the multiplication of the individual probabilities of that class within each individual descriptor [164–166]. We implemented NB learner node within KNIME Analytics Platform (Version 4.1.3) with the following parameters: Default probability = 0.0001, minimum standard deviation = 0.0001, threshold standard deviation = 0.0 and maximum number of unique nominal values per attribute = 20.

2.5.6. Multilayer perceptron (MLP)

It is an implementation of the RProp algorithm for multilayer feed forward networks [167]. MLP has the capacity to learn nonlinear models in real time. MLP can have one or more nonlinear hidden layers between the input and output layers. For each hidden layer, different numbers of hidden neurons can be assigned. Each

Table 2
Chemical structures of the modeled compounds and their reported bioactivities.

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
1	Nafamostat		100	Aromatic ester	[97]
2*	Camostat		1000	Aromatic ester	[97]
3	5ce1-Hepsin		100	None	PDB code: 5CE1
4	[76][a]		8	None	[90]
5*	[24][a]		19	None	[90]
6	[66][a]		20	None	[90]
7	[25][a]		19	None	[90]

(continued on next page)

Table 2 (*continued*)

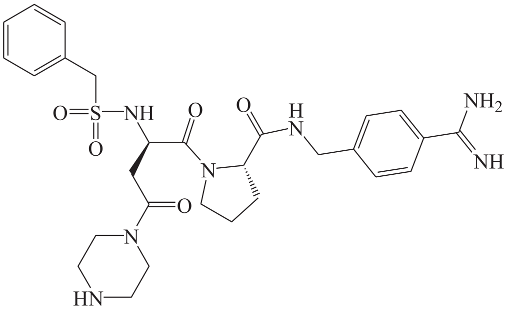
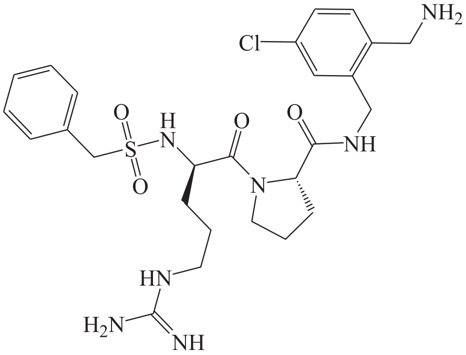
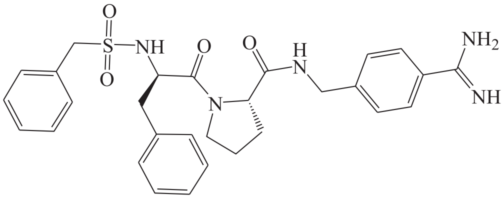
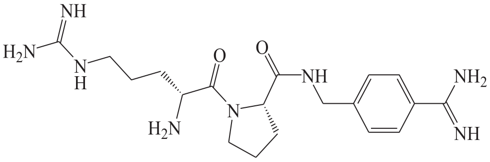
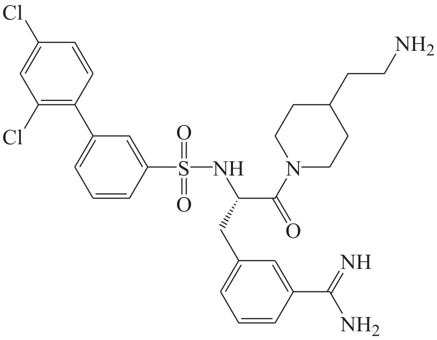
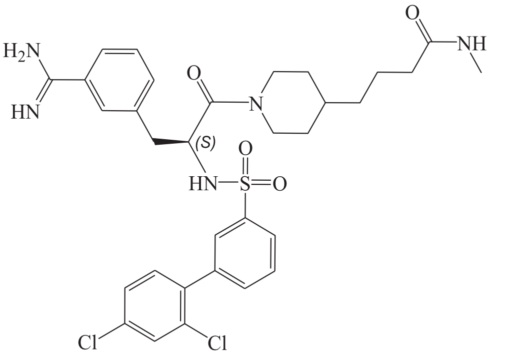
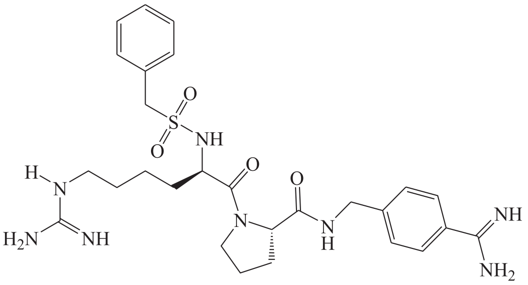
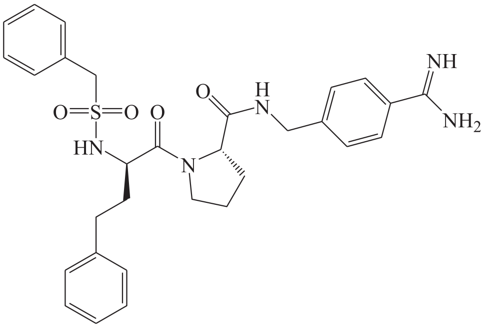
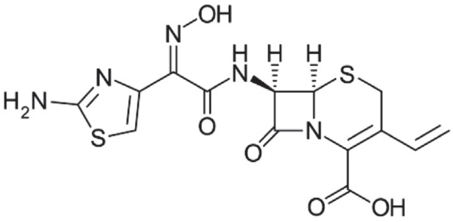
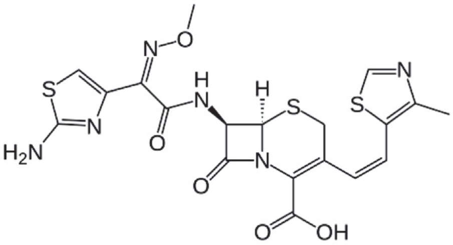
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
8	[54][a]		45	None	[90]
9*	[71][a]		60	None	[90]
10*	[31][a]		160	None	[90]
11	[69][a]		220	None	[90]
12	[92][a]		0.9	None	[90]

Table 2 (*continued*)

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
13	[93][a]		4	None	[90]
14[*]	[27][a]		21	None	[90]
15	[29][a]		50	None	[90]
16	Cefdinir		Inactive[**]	β-lactam	[97]
17	Cefditoren		Inactive[**]	β-lactam	[97]

(*continued on next page*)

Table 2 (*continued*)

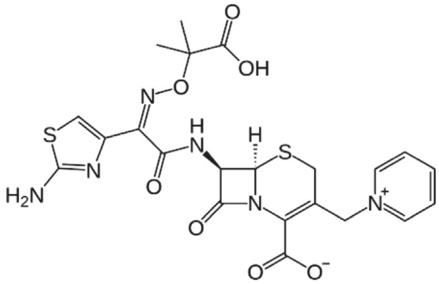
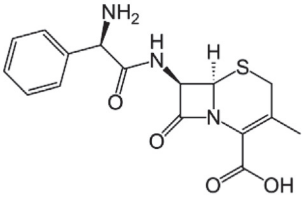
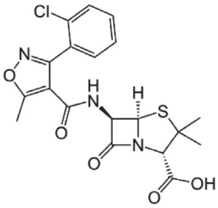
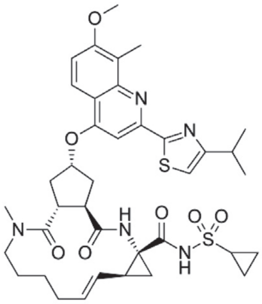
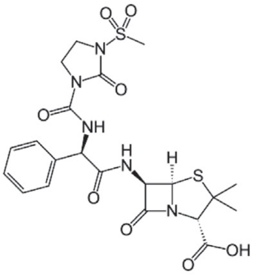
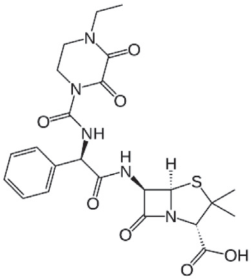
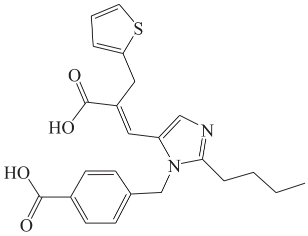
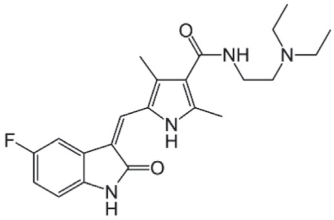
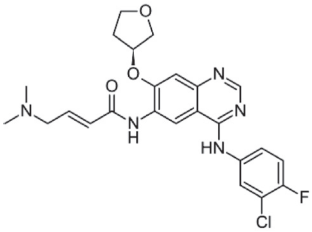
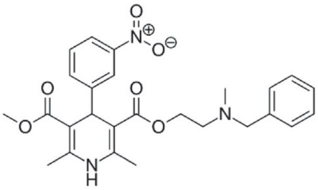
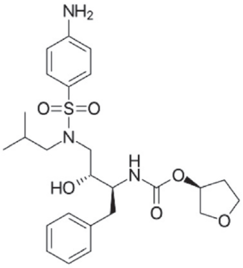
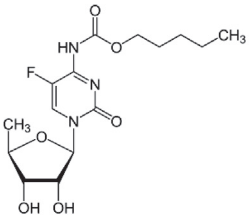
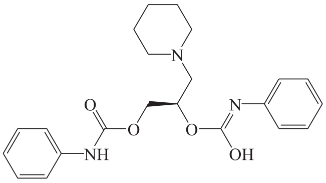
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
18	Ceftazidime		Inactive**	β-lactam	[97]
19*	Cephalexin		Inactive**	β-lactam	[97]
20	Cloxacillin		Inactive**	β-lactam	[97]
21*	Simeprevir		Inactive**	None	[97]
22	Mezlocillin		Inactive**	β-lactam	[97]
23	Piperacillin		Inactive**	β-lactam	[97]

Table 2 (*continued*)

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
24*	Eprosartan		Inactive**	Michael Acceptor	[97]
25	Sunitinib		Inactive**	Michael Acceptor	[97]
26*	Afatinib		Inactive**	Michael Acceptor	[97]
27	Nicardipine		Inactive**	Nitro	[97]
28	Amprenavir		Inactive**	Carbamate	[97]
29	Capecitabine		Inactive**	Carbamate	[97]
30	Diperodon		Inactive**	Carbamate	[97]

(*continued on next page*)

Table 2 (*continued*)

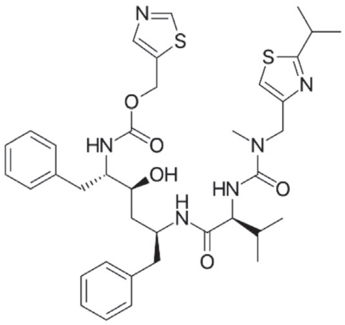
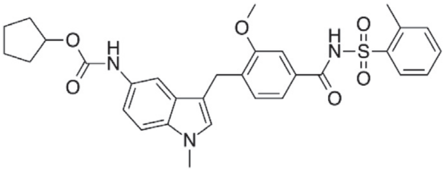
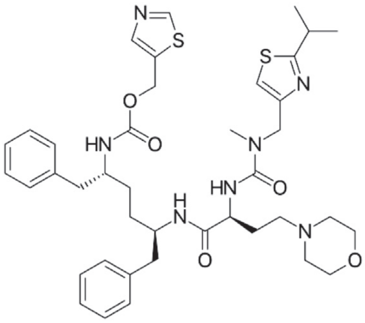
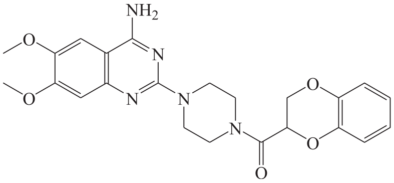
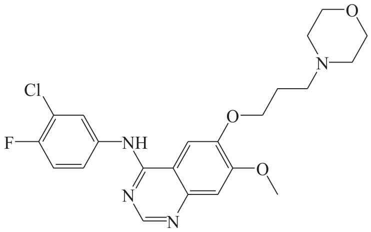
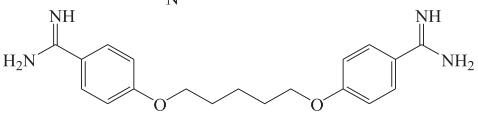
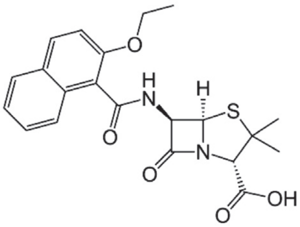
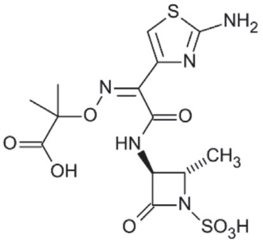
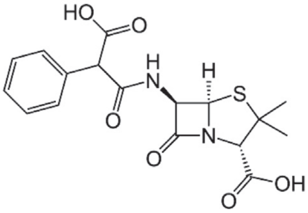
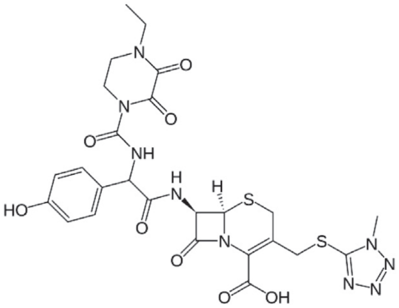
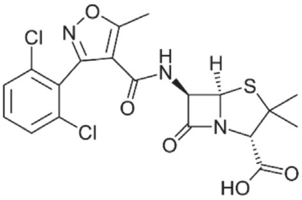
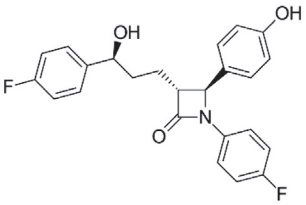
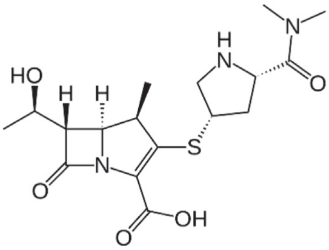
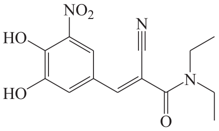
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
31*	Ritonavir		Inactive**	Carbamate	[97]
32	Zafirlukast		Inactive**	Carbamate	[97]
33*	Cobicistat		Inactive**	Carbamate	[97]
34	Doxazosin		Inactive**	None	[97]
35	Gefitinib		Inactive**	None	[97]
36	Pentamidine		Inactive**	None	[97]
37	Nafcillin		Inactive**	β-lactam	[97]

Table 2 (*continued*)

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
38	Aztreonam		Inactive[**]	β-lactam	[97]
39[*]	Carbenicillin		Inactive[**]	β-lactam	[97]
40	Cefoperazone		Inactive[**]	β-lactam	[97]
41	Dicloxacillin		Inactive[**]	β-lactam	[97]
42	Ezetimibe		Inactive[**]	β-lactam	[97]
43	Meropenem		Inactive[**]	β-lactam	[97]
44	Entacapone		Inactive[**]	Michael Acceptor	[97]

(*continued on next page*)

Table 2 (*continued*)

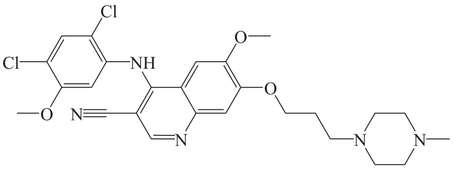
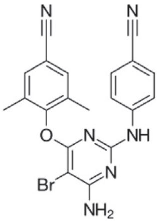
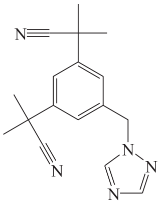
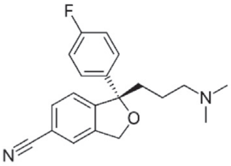
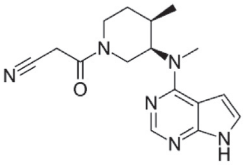
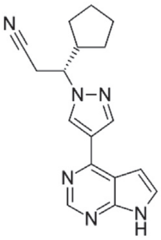
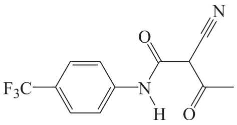
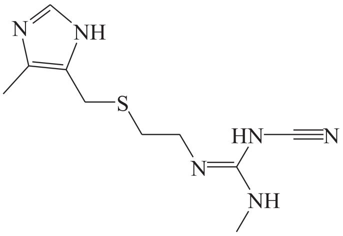
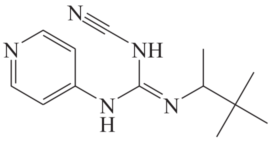
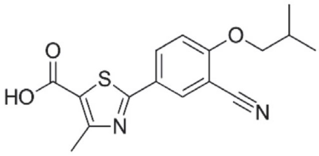
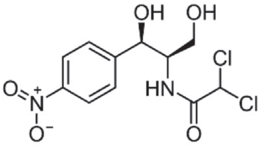
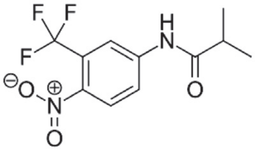
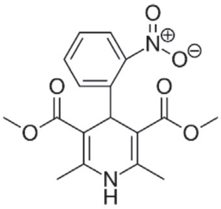
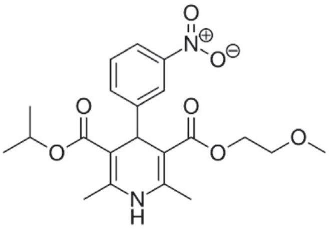
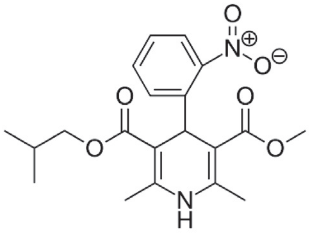
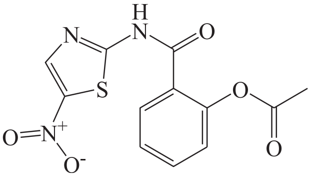
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
45	Bosutinib		Inactive**	Michael Acceptor	[97]
46	Etravirine		Inactive**	Nitrile	[97]
47	Anastrozole		Inactive**	Nitrile	[97]
48	Escitalopram		Inactive**	Nitrile	[97]
49	Tofacitinib		Inactive**	Nitrile	[97]
50	Ruxolitinib		Inactive**	Nitrile	[97]
51	Teriflunomide		Inactive**	Nitrile	[97]
52	Cimetidine		Inactive**	Nitrile	[97]

Table 2 (*continued*)

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
53*	Pinacidil		Inactive**	Nitrile	[97]
54	Febuxostat		Inactive**	Nitrile	[97]
55	Chloramphenicol		Inactive**	Nitro	[97]
56*	Flutamide		Inactive**	Nitro	[97]
57*	Nifedipine		Inactive**	Nitro	[97]
58	Nimodipine		Inactive**	Nitro	[97]
59	Nisoldipine		Inactive**	Nitro	[97]
60	Nitazoxanide		Inactive**	Nitro	[97]

(*continued on next page*)

Table 2 (*continued*)

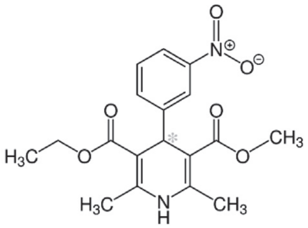
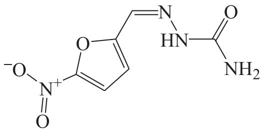
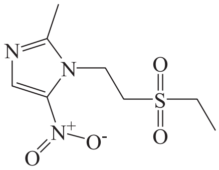
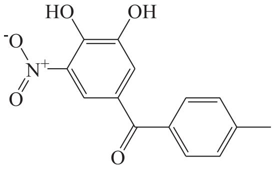
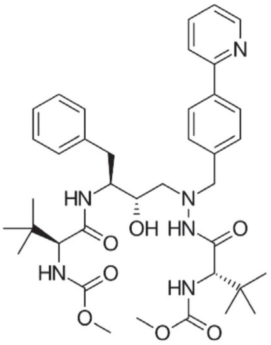
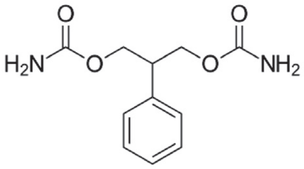
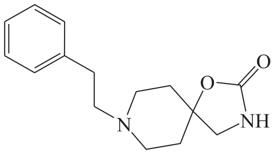
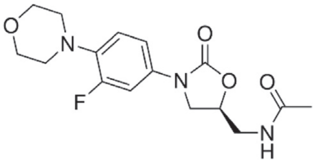
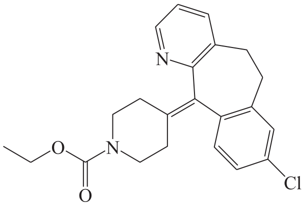
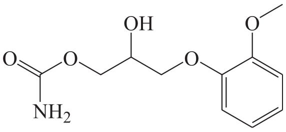
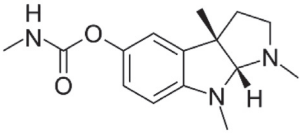
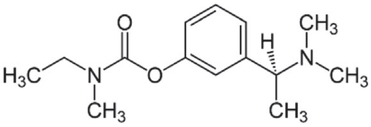
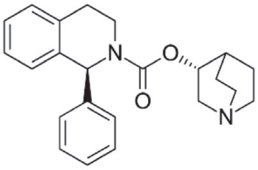
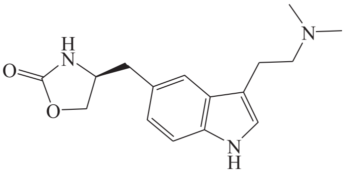
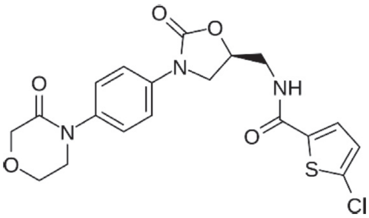
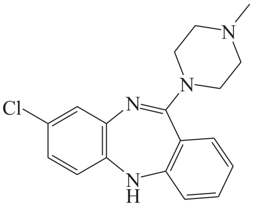
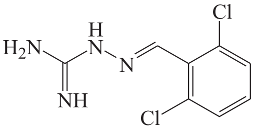
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
61	Nitrendipine		Inactive**	Nitro	[97]
62*	Nitrofurazone		Inactive**	Nitro	[97]
63	Tinidazole		Inactive**	Nitro	[97]
64	Tolcapone		Inactive**	Nitro	[97]
65*	Atazanavir		Inactive**	Carbamate	[97]
66	Felbamate		Inactive**	Carbamate	[97]
67	Fenspiride		Inactive**	Carbamate	[97]
68	Linezolid		Inactive**	Carbamate	[97]

Table 2 (*continued*)

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
69*	Loratadine		Inactive**	Carbamate	[97]
70*	Methocarbamol		Inactive**	Carbamate	[97]
71	Physostigmine		Inactive**	Carbamate	[97]
72	Rivastigmine		Inactive**	Carbamate	[97]
73	Solifenacin		Inactive**	Carbamate	[97]
74	Zolmitriptan		Inactive**	Carbamate	[97]
75	Rivaroxaban		Inactive**	Carbamate	[97]
76	Clozapine		Inactive**	None	[97]
77	Guanabenz		Inactive**	None	[97]

(*continued on next page*)

Table 2 (*continued*)

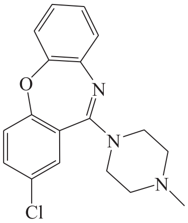
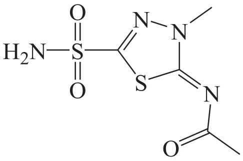
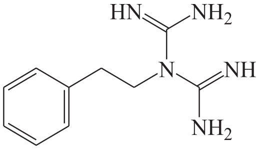
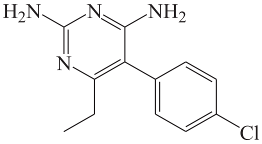
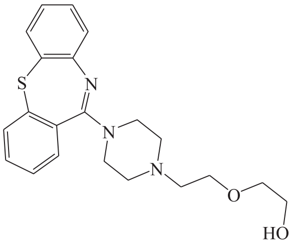
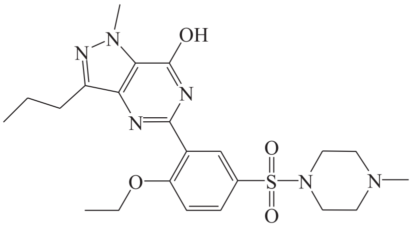
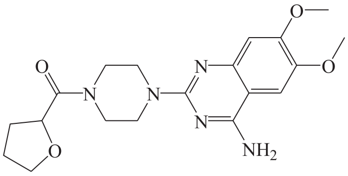
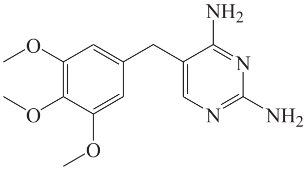
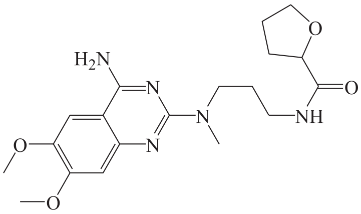
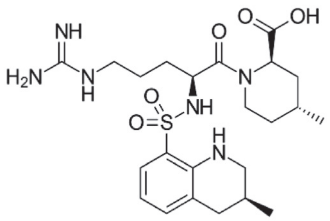
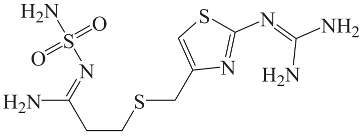
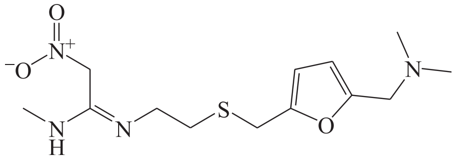
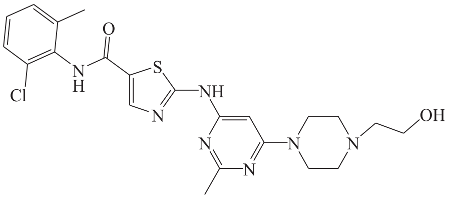
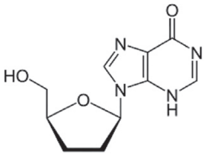
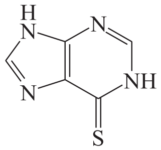
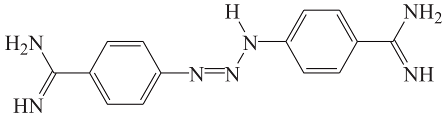
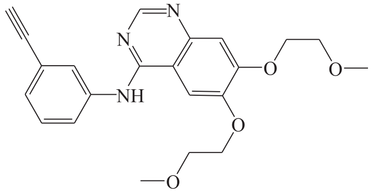
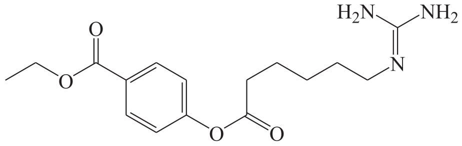
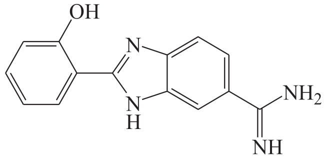
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
78	Loxapine		Inactive**	None	[97]
79	Methazolamide		Inactive**	None	[97]
80	Phenformin		Inactive**	None	[97]
81	Pyrimethamine		Inactive**	None	[97]
82	Quetiapine		Inactive**	None	[97]
83	Sildenafil		Inactive**	None	[97]
84*	Terazosin		Inactive**	None	[97]
85*	Trimethoprim		Inactive**	None	[97]
86	Alfuzosin		Inactive**	None	[97]

Table 2 (*continued*)

Compound		Structure	IC$_{50}$ or Ki (nM)	Warhead	Reference
Number	Name				
87	Argatroban		Inactive**	None	[97]
88	Famotidine		Inactive**	None	[97]
89	Ranitidine		Inactive**	None	[97]
90	Dasatinib		Inactive**	None	[97]
91	Didanosine		Inactive**	None	[97]
92	Mercaptopurine		Inactive**	None	[97]
93*	Diminazene		Inactive**	None	[97]
94	Erlotinib		Inactive**	None	[97]
95	Gabexate		Inactive**	Aromaticester	[97]
96	1p57-Hepsin		40,000	None	[138]

(*continued on next page*)

Table 2 (*continued*)

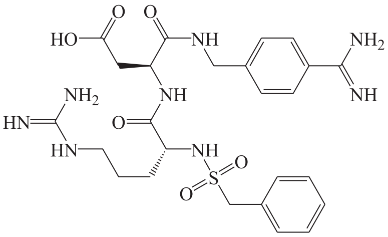
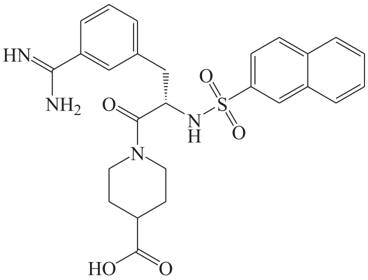
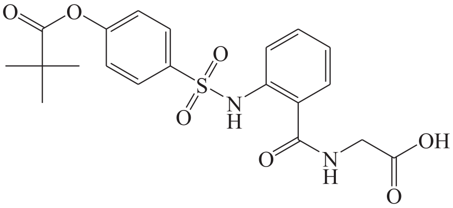
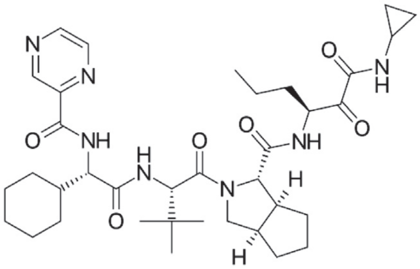
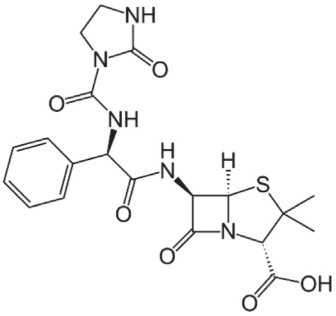
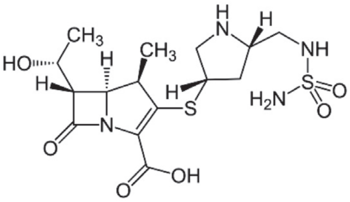
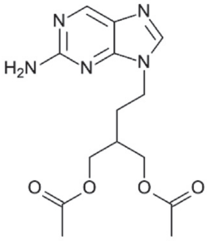
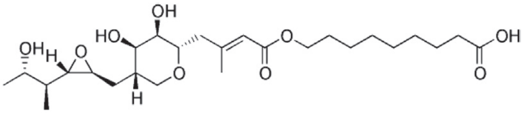
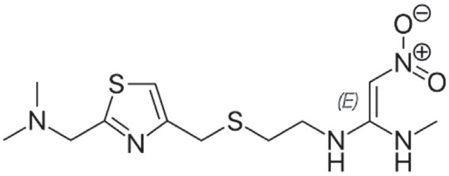
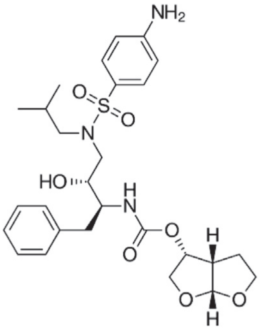
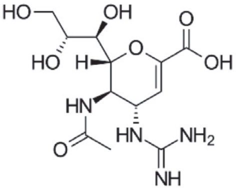
Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
97	[53][a]		10,000	None	[90]
98	[78][a]		20,000	None	[90]
99	Sivelestat		Inactive**	Aromatic ester	[97]
100	Telaprevir		Inactive**	α-keto-amide	[97]
101	Azlocillin		Inactive**	β-lactam	[97]
102	Doripenem		Inactive**	β-lactam	[97]

Table 2 (*continued*)

Compound		Structure	IC_{50} or Ki (nM)	Warhead	Reference
Number	Name				
103*	Famciclovir		Inactive**	None	[97]
104	Mupirocin		Inactive**	Michael Acceptor	[97]
105	Nizatidine		Inactive**	Nitro	[97]
106	Darunavir		Inactive**	Carbamate	[97]
107	Zanamivir		Inactive**	None	[97]

[*] These compounds were used as testing compounds in machine learning.
[**] Defined as "Inactive" by the particular reference.
[a] Numbers in brackets represent the number of each compound in the original literature citation.

hidden neuron gives a weighted linear summation for the values from the previous layer, and the nonlinear activation function is followed. The output values are reported after the output layer transforms the values from the last hidden layer. We implemented MLP learner node within KNIME Analytics Platform (Version 4.1.3) with the following optimized parameters: Maximum number of iterations = 100, Number of hidden layers = 3, and number of hidden neurons per layer = 100.

2.5.7. Support vector machine (SVM)

Support vector machine (SVM) chooses a small number of boundary instances called support vectors to create discriminatory function to separates training observations into distinct classes with widest possible boundaries. SVM allows the effective use of a multitude of kernels to allow classification. A key feature of SVMs is the attempt to minimize the error on training data and reduce the computational complexity of models to avoid over fitting by tuning the factors involved in the process [168,73]. Two SVM method were attempted, namely, C-SVM and nu-SVM. C and nu are regularization parameters that penalize misclassifications. C ranges from 0 to infinity while nu ranges between 0 and 1 and represents the lower and upper bound on the number of examples that are support vectors and that lie on the wrong side of the hyperplane. The following default settings were used in both SVM methods as implemented in the WEKA-KNIME (version 4.1.3) LibSVM node, these include: Kernel Cache (Cache Size = 40.0), kernel type is radial basis function: exp(-gamma*|u-v|^2), and loss function is 0.1, kernel coefficients epsilon = 0.001 and Gamma = 0.00. However, in nu-SVM the optimized nu value of 0.1 was used (identified using Bayesian Optimization (TPE) implemented in KNIME).

2.5.8. ML model evaluation

ML models were evaluated by calculating their accuracies (Eq. (1)) and Cohen's kappa values (Eq. (2)) [169–171] against the training and testing sets (Table 2).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{N} \tag{1}$$

where, TP is the true positive (correctly classified actives), TN true negatives (truly classified inactives), and n is the total number of evaluated compounds.

$$K = \frac{P_0 + P_e}{1 - P_e} \tag{2}$$

where P_o is the relative observed agreement among raters (i.e., accuracy), and P_e is the hypothetical probability of chance agreement. This is done by using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement, then kappa = 1. If there is no agreement among the raters other than what would be expected by chance (as given by P_e), kappa = 0. Negative Cohen's kappa value implies the agreement is worse than random [172].

Evaluation against the training set involves removing 20% (i.e., leave-20%-out or 5-fold cross-validation) of the data points (i.e., compounds), then building the particular ML model from the remaining data. The model is then used for classifying the removed compounds. The process is repeated until all training data points are removed from the training list and predicted at least once. Accuracy is calculated based on comparing classification results with actual bioactivity classes. On the other hand, evaluation against the testing set involves calculating the accuracy Cohen's kappa of the particular ML model by comparing its classification results with the actual bioactivity classes of the testing set [173,174].

2.5.9. Virtual screening

To collect FDA-approved drug molecules for virtual screening, we employed SMARTS codes corresponding to reactive warheads in Table 1 to screen an *in house* built list of FDA-approved drugs (2111 molecules) for molecules of covalent warheads. Screening was performed within DiscoveryStudio (version 4.5) environment. The resulting list was then compared with TMPRSS2-related ligands reported in the literature (active and inactive) [90,97]. Only untested compounds (155 molecules) were kept for subsequent virtual screening as potential TMPRSS2 inhibitors. Supplementary Table SM-2 shows screened compounds and their corresponding chemical structures and warheads.

The evaluated compounds were prepared, docked, scored, RMSD-filtered and have their LRCFs determined exactly in the same manner as described for the modelled training and testing compounds (sections 2.3–2.4). Subsequently, their bioactivity classifications were predicted using successful ML models.

2.6. Molecular dynamics

Docked poses of capreomycin, aspoxicillin, fosamprenavir corresponding to highest consensus scores were solvated by VMD in TIP3 water molecules. The complexes were then neutralized with NaCl and the systems were minimized by conjugate gradient minimization until 10 KJ/mol/nm. Simulations were run using CHARMM22 force field for proteins and general forcefield for drug-like molecules (CGenFF) (https://cgenff.paramchem.org) for ligands. NVE ensemble was arbitrary applied. SHAKE constraints were applied to all hydrogen atoms. Step size was set to be 2 femto-seconds (fs). During the heating phase, the temperature of the system was raised linearly from 0 to 310 K over 155,000 steps with a time step of 0.002 ps. Equilibration of the solvent molecules was achieved in 1,000,000 steps of simulation (2 ns). This was followed by 200 ns of production simulation for data collection, during which structures were stored every 1 ns. Minimization, heating, equilibration, and heating–cooling simulations were performed using OpenMM software (http://openmm.org/). Results from all simulationswere visualized using Discovery Studio (version 4.5, Biovia, USA).

2.7. Covalent docking

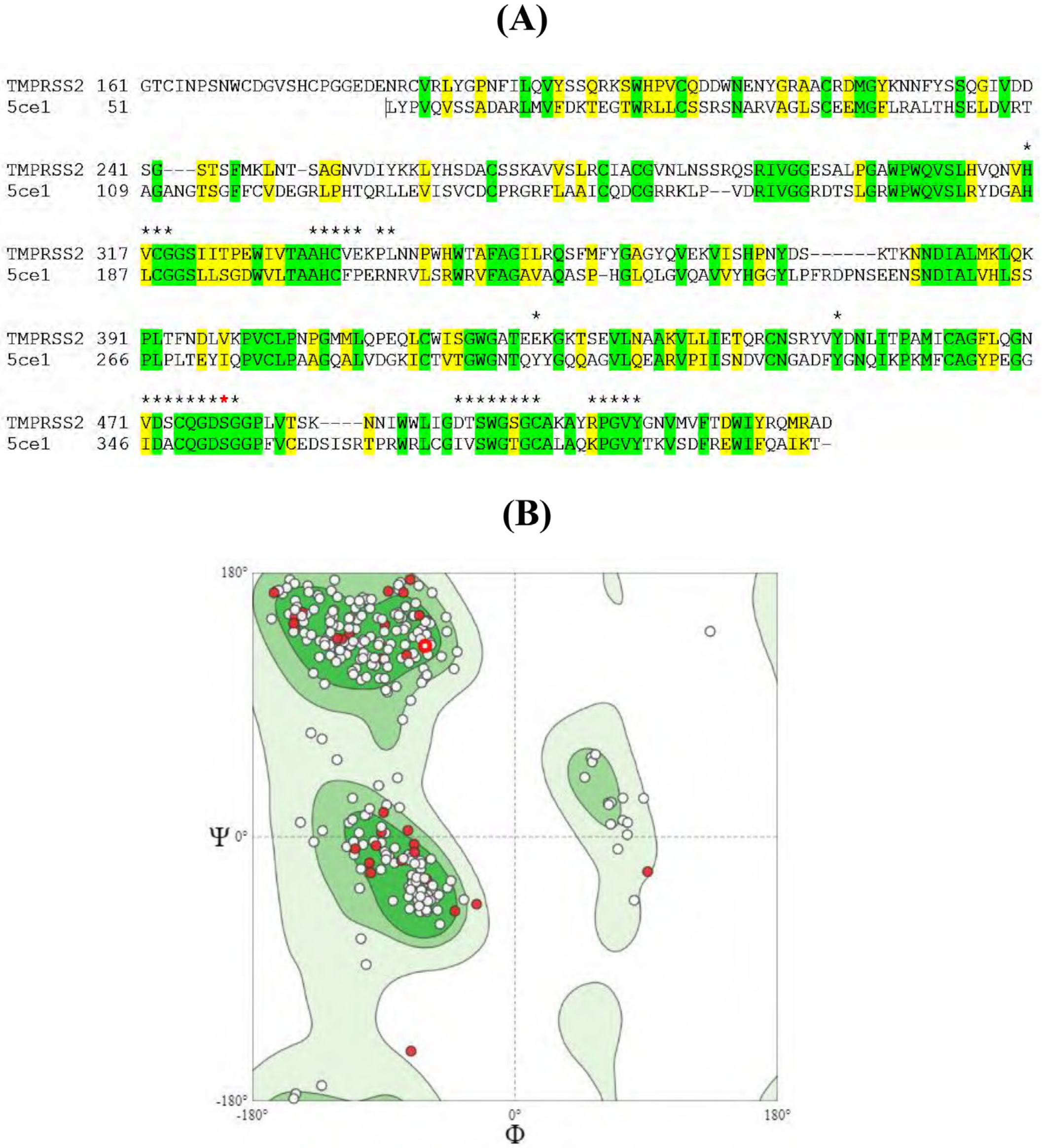
Capreomycin and aspoxicillin were covalently docked into TMPRSS2 homology model using CovDock software [175] through Maestro Molecular Modeling Interface (Schroedingers Inc., USA). CovDock begins with Glide docking to a receptor with the reactive residue trimmed to alanine. The reactive residue is then added and sampled to form a covalent bond with the ligand in different poses. Covalent complexes are minimized using the Prime VSGB2.0 energy model to score the top covalent complexes. An apparent affinity score, based on the Glide score of pre-reactive and post-reactive poses, is also calculated to estimate binding energies for use in virtual screening. The following docking settings were implemented: The docking mode was set to Pose Prediction (thorough), Energy cutoff to retain poses for further refinement = 2.5 k cal/mol with a maximum number of poses = 200, the reactive residue is S478, docked ligand(s) is(are) confined to an enclosing box of $\leq$ 20 Å, the center of the enclosing box is set to the hydroxyl of the S478. In each docked ligand, the corresponding reaction type was selected from the available drop-list (i.e., β-lactam addition and Michael acceptor). A single top-ranking docked pose was reported as output.

3. Results and discussion

3.1. Homology modeling of TMPRSS2

Since TMPRSS2 had no known crystallographic structure at the time of preparing this manuscript, it was necessary to build an appropriate homology model for this target. Sequence data for TMPRSS2 was obtained from Pubmed (GenBank accession number NP_001128571). Then BLAST and HHBlits search (141, 142) for pairwise sequence-to-sequence alignment was performed to search and identify close template structure(s) in the protein databank (https://www.rcsb.org/). However, we only opted for template proteins that have at least 65% sequence coverage with TMPRSS2 and have their co-crystallized bound ligands successfully transferred by SWISS-MODEL to the proposed binding pocket of TMPRSS2. In SWISS-MODEL, for a template bound ligand to be transferred to the corresponding homology model, the ligand should be biologically related to the modelled protein, has favorable interactions and no clashing contacts within model atoms, and the contacting residues are conserved between the target and the template. Subsequent homology modelling was performed using SWISS-MODEL on the resulting alignments [143,144,156–158].

Hepsin crystallographic structure 5ce1 was selected as template as it scored 66% sequence coverage, had its bound ligand successfully transferred to the homology model, and achieved the highest SWISS-MODEL Global Model Quality Estimate (GMQE = 0.49), despite an overall sequence identity of 33.43%. Out of 35 binding site amino acids in the homology model, 23 are identical (66%) to their counterparts in the template, while 4 amino acids were similar (11%). Overall, amino acid homology with TMPRSS2 in the proposed binding site is 77% with the catalytic serine (S478) being conserved. Fig. 3A compares TMPRSS2 sequence to the template protein Hepsin.



(A)

```
                                                                                   *
TMPRSS2 161 GTCINPSNWCDGVSHCPGGEDENRCVRLYGPNFILQVYSSQRKSWHPVCQDDWNENYGRAACRDMGYKNNFYSSQGIVDD
5ce1     51                      |LYPVQVSSADARLMVFDKTEGTWRLLCSSRSNARVAGLSCEEMGFLRALTHSELDVRT

                                                                                   *
TMPRSS2 241 SG---STSFMKLNT-SAGNVDIYKKLYHSDACSSKAVVSLRCIACGVNLNSSRQSRIVGGESALPGAWPWQVSLHVQNVH
5ce1    109 AGANGTSGFFCVDEGRLPHTQRLLEVISVCDCPRGRFLAAICQDCGRRKLP--VDRIVGGRDTSLGRWPWQVSLRYDGAH

            ***            ***** **
TMPRSS2 317 VCGGSIITPEWIVTAAHCVEKPLNNPWHWTAFAGILRQSFMFYGAGYQVEKVISHPNYDS------KTKNNDIALMKLQK
5ce1    187 LCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAVAQASP-HGLQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSS

                                               *                          *
TMPRSS2 391 PLTFNDLVKPVCLPNPGMMLQPEQLCWISGWGATEEKGKTSEVLNAAKVLLIETQRCNSRYVYDNLITPAMICAGFLQGN
5ce1    266 PLPLTEYIQPVCLPAAGQALVDGKICTVTGWGNTQYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGYPEGG

            *********                   ********    *****
TMPRSS2 471 VDSCQGDSGGPLVTSK----NNIWWLIGDTSWGSGCAKAYRPGVYGNVMVFTDWIYRQMRAD
5ce1    346 IDACQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIFQAIKT-
```

(B)

Fig. 3. Criteria for the resulting TMPRSS2 homology model. **(A)** Alignment of TMPRSS2 (GenBank: NP_001128571) and Hepsin crystallographic template (PDB code: 5ce1, resolution 2.5 Å) yielding homology model. Gaps are shown as (–), identical and similar residues are highlighted in green and yellow, respectively. Binding site amino acids are indicated with asterisks, the catalytic S478 is marked with red asterisk. (B) Ramachandran plot of the homology model amino acids. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The resulting homology structure was evaluated employing structure assessment tools within SWISS-MODEL. The results are as follows (in brackets): QMEAN (-1.48), MolProbity score (1.91), clash score (6.0), Ramachandran Favored (92.17%), Ramachandran Outliers (1.16%) (See Fig. 3B), rotamer outliers (1.35%), C-Beta deviations (reporting geometry problems around alpha-carbons:7 amino acids), bad angles (37 amino acids) and no bad bonds.

Interestingly, all binding site amino acids were devoid of any significant modelling-related artefacts as in Fig. 4A. Thankfully, the binding site of a recently release crystallographic structure of TMPRSS2 (7 meq, released 2021-04-21) fitted closely the binding site of our homology model within a radius of 12 Å surrounding the co-crystallized ligand. The alignment scored RMSD values of 1.76 Å and 1.74 Å based on α-carbons (Ca) and main chain carbons, respectively. Fig. 4B shows the alignment between the binding sites of the homology model and corresponding X-ray structure of TMPRSS2. Still, the side of chain of the critical binding amino acid Gln438 in the X-ray structure (corresponding to Gln475 in the homology model) is missing including the terminal amide moiety. This major artefact undermines the validity of docking studies using the crystallographic structure and adds merits to the homology model.

3.2. Data collection, Docking, scoring and Machine Learning.

The literature was carefully searched for TMRPSS2 inhibitors. The search identified 1064 inhibitors that can be unequivocally classified as "active" ($IC_{50} \leq$ 1,000 nM) or "inactive" (IC_{50}-$\geq$ 10,000 nM) as in the **Data Collection** section under **Experimental**. However, we opted to select a subset of inhibitors for performing this study partially to minimize the computational cost, but mainly, to assess the possibility of exploiting limited number of ligands for building successful and predictive machine learning models. Limitations related to number of available ligands is often encountered in drug discovery projects particularly those involving new biotargets.

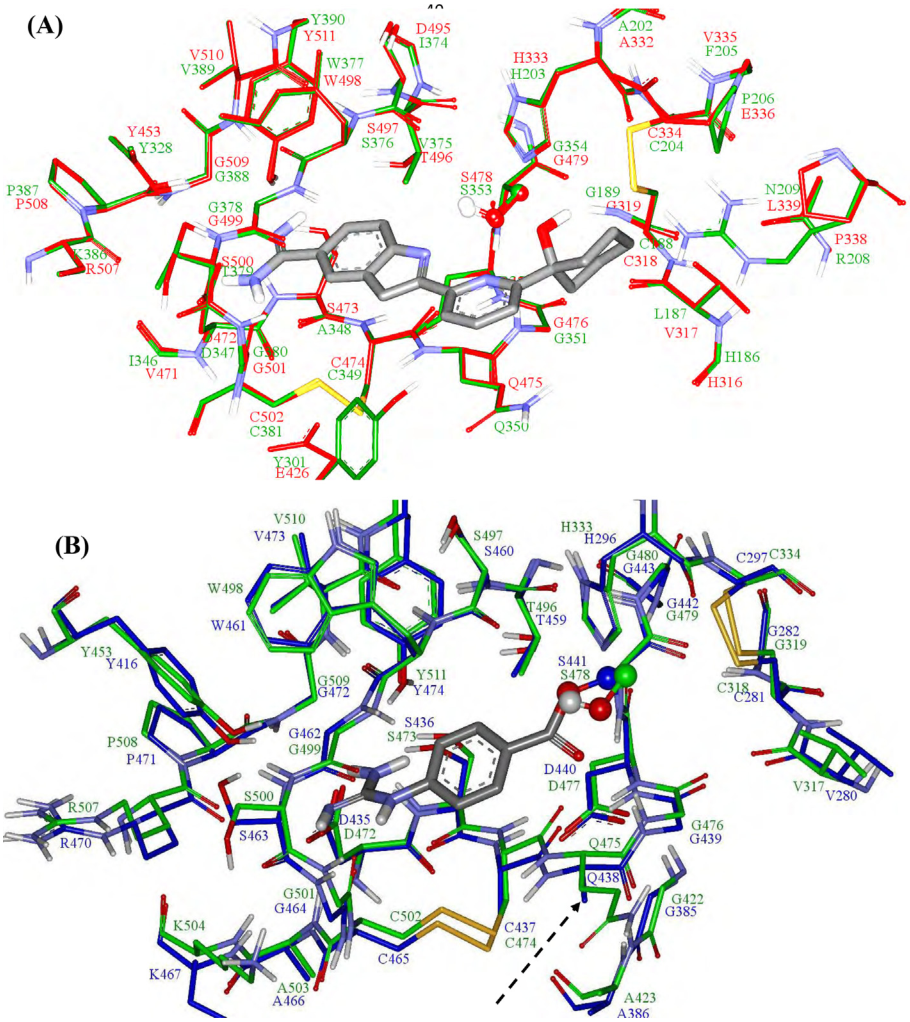


Fig. 4. TMPRSS2 binding site **(A)** Detailed view of many TMPRSS2 homology model binding site residues (green backbone) compared to their corresponding counter parts in template serine protease Hepsin (PDB ID: 5ce1, red backbone) X-ray structure including the crystallographic bound pose of hepsin inhibitor 2-[6-(1-hydroxycyclohexyl) pyridin-2-yl]-1H-indole-5-carboximidamide (compound **3** in Table 2). Most binding site amino acids are correctly aligned (**B**) Main binding site residues of TMPRSS2 homology model (green backbone) aligned to their counterparts in a recently released TMPRSS2 X-ray crystalographic structure (PDB ID: 7 meq, blue backbone) including the bound fragment of nafamostat (compound **1** in Table 2). The dotted arrow points to the missing side chain of Gln438 (corresponding to Gln475 in the homology model). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To select diverse representative molecules from the collected compounds, it was decided to use principal component analysis (PCA) combined with visual selection, as in Fig. 2. Eventually, 107 compounds were selected for modelling, including: 15 actives (of which two are irreversible) and 92 inactives (of which 64 exhibit covalent warheads). Table 2 shows the chemical structures of the modeled compounds and their bioactivities (90, 97, 137, 140). Incidentally, we attempted to use automatic means for selecting diverse subset employing the Find Diverse Molecules protocol within Discovery Studio (it attempts to select diverse sets based on a maximum dissimilarity algorithm using Tanimoto distance and molecular fingerprints), however, the selected list had limited number of actives (only two molecules). Moreover, it is of comparable diversity criteria to our visually-selected subset (as in Supplementary Table SM-1).

The modeled compounds were appropriately ionized and tautomerized yielding 338 tautomeric forms ready for docking.

The structures were then docked into the binding pocket of TMPPRSS2 homology model using three docking engines, namely, LibDock (26,147), CDOCKER (149) and LigandFit (148). This yielded 65,941 docked conformers/poses. The resulting docked poses were then scored by 9 docking scoring functions. Docked poses of each compound that scored within top 20% of at least 4 scoring functions (i.e., consensus among 4 scoring functions) were retained for subsequent processing and the rest were discarded resulting in 12,457 docked poses/conformers. The reason for this restriction is to limit ML modelling to high-quality docked poses vis-à-vis binding enthalpies in order to minimize ML noise caused by low quality poses.

The veracities of the implemented docking settings were evaluated by two means:

(i) Looking at the docked poses of co-crystallized hepsin ligand (PDB code: 5ce1, compound **3** in Table 2) and nafamostat (compound **1** in Table 2). In the former case we were interested to see if the docked pose(s) approximate the experimental pose of the ligand despite being generated for another protein (i.e., hepsin). The close analogy between hepsin and TMPRSS2 binding sites (as in Fig. 4) should correct this discrepancy. On the other hand, the docked poses of nafamostat should probe the docking veracity by looking at the distance between nafamostat's warhead (the ester carbon atom) and the nucleophilic oxygen of Ser478. Proximity between these two atoms below 3.3 Å should allow successful covalent bond formation (see section **Molecular Dynamics Simulation and Covalent Docking** below), and therefore demonstrates the veracity of the docking settings. Needless to say, that nafamostat covalently binds TMPRSS2 Ser478 hydroxyl as in Fig. 4B.

Fig. 5 shows the best docked poses in the two cases. In compound **3** case, 2 and 14, out of 28 docked poses are within RMSD 2.0 and 4.0 Å, respectively, from the crystallographic pose, while one of the top-ranking docked poses (i.e., 4th based on consensus score and 1st based on 6 out of 9 scoring functions) achieved RMSD of 1.18 Å from the crystallographic pose, as in Fig. 5A. Moreover, upon docking **3** into its native protein, hepsin (PDB code: 5ce1), the same docking-scoring settings generated 89 docked poses, all of which were within RMSD of 1.2 Å from the crystallographic bound pose, as in Supplementary Fig. SM-1. Needless to say, success to reproduce the crystallographic pose (i.e., RMSD $\leq$ 2.0 Å) among top-ranked solutions is considered sufficient validation for certain docking-scoring settings [182–184].

Unfortunately, it is not possible to perform similar analysis in nafamostat (compound **1**) case because the corresponding complex (PDB code: 7 meq) includes only fragment of the bound ligand following covalent bond formation with Ser478, i.e., it is not possible to compare whole docked nafamostat with partial crystallographic fragment in 7 meq. As an alternative we opted to gauge the success of the docking-scoring settings by measuring the distance separating the warhead of docked nafamostat poses from the hydroxyl of Ser478 (as in Fig. 5B). Interestingly, 4 out of 285 docked poses were positioned within the binding site such that the warhead ester carbon of nafamostat occurs within 3.3 Å from the nucleophilic hydroxyl of Ser478. Two poses are shown in Fig. 5B. Such close proximity should allow successful covalent bond formation (177), and therefore, further validates our docking settings.

(ii) By assessing the difference between docking-scoring values calculated for docked poses of active versus inactive compounds (training and testing alike). T-test analysis indicates that 8 out of 9 docking-scoring functions (namely, LigScore1, LigScore2, -PLP1, -PLP2, Jain, -PMF, -PMF04, -Cdocker Interaction Energy) have generated statistically significant scoring values for docked poses of active compounds compared to inactive counterparts (training and testing alike), as in Supplementary Table SM-7. These results further support the notion that the implemented docking/scoring

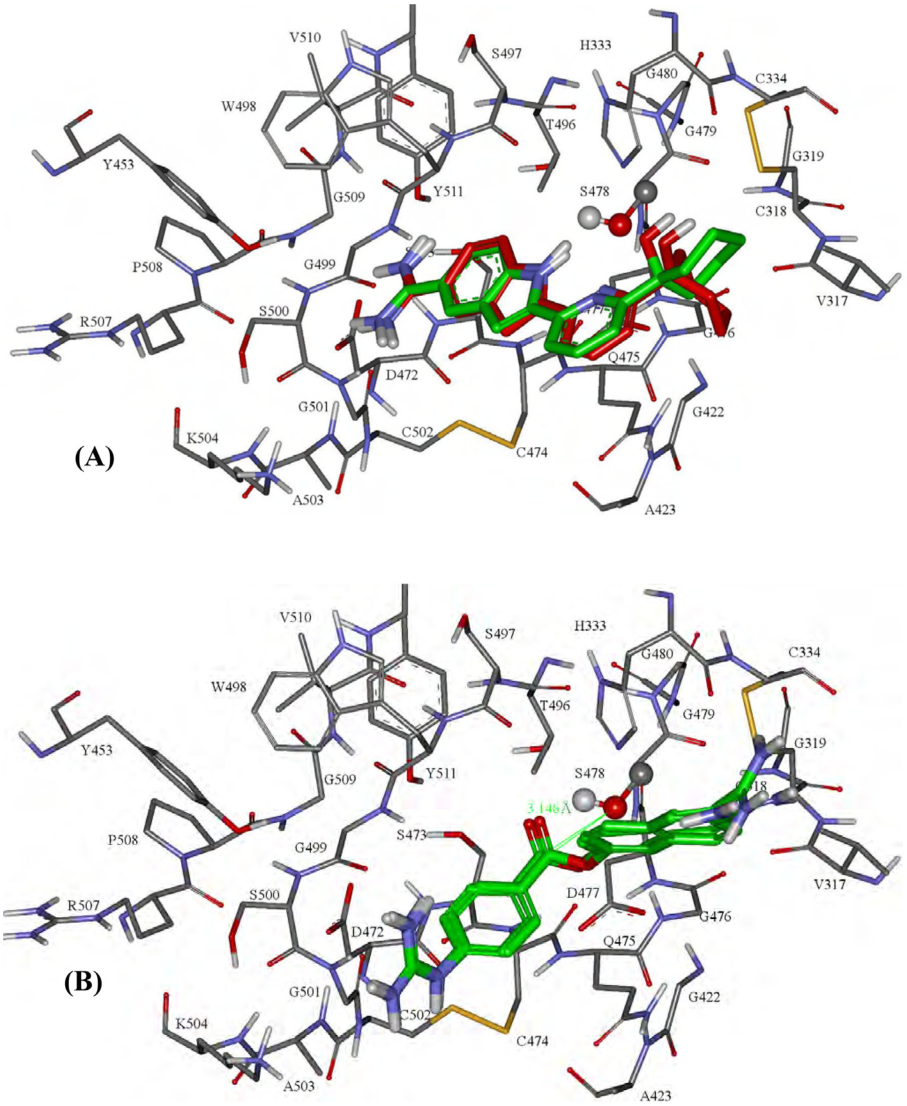


Fig. 5. Assessment of docking veracity. **(A)** The co-crystallized pose of **3** (Table 2, Green skeleton) within hepsin (PDB code: 5ce1) extracted into the binding site of TMPRSS2 homology model and compared with high ranking docked pose of the same molecule within TMPRSS2 binding site (red skeleton). **(B)** Two high ranking docked poses of nafamostat (compound **1** in Table 2) with their ester carbons at close proximities (*ca.* 3.1 Å) to the nucleophilic hydroxyl of Ser478. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conditions segregate active compounds from inactives by allocating them distinct binding poses and regions within the binding site.

We hypothesize that convergence of particularly high-quality docked poses, corresponding to active ligands, on certain unique binding site contacts (i.e., not contacted by high quality docked poses of inactive ligands) highlights the significance of those binding site points as activity discriminators. Fig. 6 illustrates this point: Clearly, the optimal docked pose of active inhibitor **12** (Table 2) fills distinct space within the binding pocket from that occupied by the best docked pose of the analogous inactive inhibitor **97**. This trend is also apparent upon comparing multiple high-ranking docked poses for the same compounds albeit less obvious to the human eye necessitating ML usage. Moreover, this consistency between docked poses and bioactivities further supports the veracity of docking settings.

To eliminate repetitions in the docked poses/conformers list, which might emerge due to the use of multiple docking engines on the same set of ligands, it was decided to filter out docked poses of RMSD <2.0. This step reduced the number of poses for subsequent ML to 12073.

In summary, it can be concluded that the overall effect of docking/scoring/RMSD filtering is enriching the ML list from just 108 observations (i.e., TMPRSS2 ligands) to 12,073 observations (corresponding docked poses) in a process reminiscent of statistical bootstrapping albeit the multiple docked poses represent realistic permutations rather than simple repetitive sampling. Subsequently, the docked poses were used to generate LRCFs such that binding site atoms that are positioned within 2.5 Å from docked poses are given a binary code of 1.0, otherwise they are annotated as zeros. Additionally, scoring values of docked poses (that survived the RMSD filter) calculated by all 9 scoring functions were

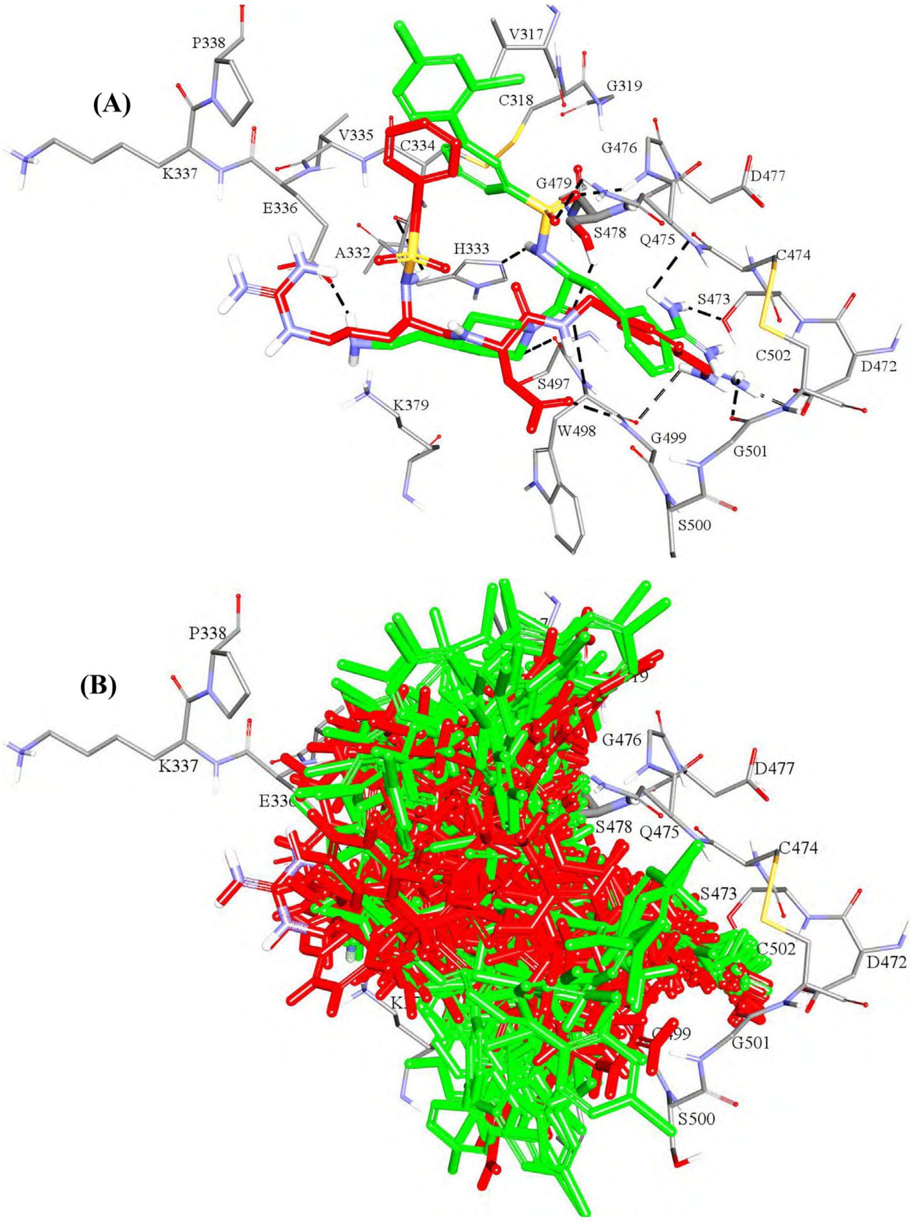


Fig. 6. Docking/bioactivity consistency. **(A)** Highest-ranking docked poses of compounds 12 (Table 2, IC_{50} = 0.9 nM, green skeleton) and 97 (Table 2, IC_{50} = 10,000 nM, red skeleton). **(B)** Docked poses of the same compounds as were used for generating LRCFs and subsequent ML. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

also added as descriptors. Thereafter, the modeled list was randomly split into training (~82%, 9890 poses corresponding to 85 compounds, see Table 2) and testing (~18%, 2183 poses corresponding to 23 compounds, see Table 2) sets. The splitting was performed based on the modelled compounds (prior to docking, Table 2). Approximately 16% and 84% of the training poses correspond to active and inactive ligands, respectively. On the other hand, the testing list included ~28% and ~72% poses corresponding to active and inactive ligands, respectively.

The apparent imbalance between poses corresponding to active and inactive ligands is attributed to the fact that known "active" TMPRSS2 inhibitors are rather limited in number and diversity, while their inactive counterparts are quite numerous and very diverse allowing them to generate significantly more docked poses. This imbalance is not necessarily disadvantageous as the excess of inactive poses should better demark the "forbidden" regions within the binding site. Nevertheless, such imbalance necessitates the use of Cohen's kappa [172,176] as additional tool for evaluating the success/failure of a particular ML.

Seven MLs were evaluated against the training and testing set, namely, Xgboost, SVM, RF, PNN, NB, kNN, and MLP. LRCFs and/or docking-scoring values were evaluated as input explanatory descriptors, as in Table 3.

Clearly from Table 3, all learners achieved excellent accuracies against testing and training sets regardless of using LRCFs and/or scoring functions values as input descriptors. This indicates that the data is self-consistent. Still, in most cases the combination of both descriptor classes generally yielded better accuracies compared to either class alone (i.e., LRCF or scoring functions).

Intriguingly, Cohen's Kappa unveiled significant differences between learners not seen from accuracy values: In three leaners, the use of a single class of descriptors completely failed to yield significant Kappa, i.e., C-SVM, nu-SVM, PNN. In the particular case of NB, combining both classes, or using LRCFs alone, failed to yield significant Cohen's Kappa values. On the other hand, in kNN and MLP cases, the combined use of both descriptor classes (LRCFs and scoring functions) failed to have any significant additive advantage over the use of one class only. However, in RF and Xgboost cases, the use of any of the two descriptor classes was successful and the combination yielded even better Cohen's Kappa values. Overall, the varying performances of different learners vis-à-vis success of their ML models upon using differing combinations of descriptor classes points to the orthogonality of these learners and that their ML models inherently different.

From the above and by comparing the behavior of different ML models in Table 3 we concluded that the best models to be used for predicting the bioactivity of screened compounds are Xgboost, RF and nu-SVM. Moreover, the orthogonality of these ML approaches prompted us to stack them in a *meta*-learner were each individual learner casts an equivalent vote for predicting the anti-TMPRSS2 bioactivity of screened ligand docked poses.

To further exclude the possibility of chance correlation we performed y-scrambling validation experiment. In this test, 100 random bioactivity data are generated. Subsequently, each learner is challenged to use these random data to generate ML models using scoring functions and LRCFs descriptors [180]. Supporting Tables SM-4 to SM-6 show the results of the experiments employing Leave-20%-Out or Out-of-Bag cross-validations. The results show that the original non-randomized data yielded ML models of significantly superior accuracy and Cohen's Kappa values compared to all randomized trials. The effect is particularly evident in Cohen's Kappa values. Overall, the results support the validity of our ML models.

Incidentally, upon using more stringent pose-selection criteria through imposing consensus score $\geq$ 8 on training and testing poses (i.e., 8 or 9 of the scoring functions ranked the selected poses

Table 3
Accuracy and Cohen's Kappa values for ML models developed using different ML learners combined with LRCFs and/or scoring function values as descriptors.

Learner	Descriptors	Accuracy			Cohen's Kappa		
		L20%out[a]	Testing[b]	Re-substitution[c]	L20%out[a]	Testing[b]	Re-substitution[c]
Xgboost	LRCFs and Scoring Functions	0.96	0.90	1.00	0.85	0.73	1.00
	LRCFs	0.91	0.83	0.97	0.63	0.53	0.89
	Scoring	0.95	0.88	1.00	0.80	0.68	1.00
C-SVM	LRCFs and Scoring Functions	0.94	0.89	0.95	0.73	0.71	0.80
	LRCFs	0.86	0.76	0.87	0.22	0.21	0.32
	Scoring Functions	0.84	0.72	1.00	0.00	0.00	1.00
nu-SVM[d]	LRCFs and Scoring Functions	0.95	0.90	0.99	0.80	0.72	0.95
	LRCFs	0.87	0.80	0.94	0.51	0.45	0.77
	Scoring Functions	0.84	0.72	1.00	0.02	0.00	1.00
RF	LRCFs and Scoring Functions	0.95[g]	0.90	1.00	0.79[g]	0.72	0.98
	LRCFs	0.90[g]	0.81	0.96	0.57[g]	0.46	0.83
	Scoring Functions	0.95[g]	0.89	1.00	0.78[g]	0.69	1.00
PNN	LRCFs and Scoring Functions	0.91	0.88	0.92	0.62	0.66	0.67
	LRCFs	0.84	0.72	0.84	0.00	0.00	0.00
	Scoring Functions	0.92	0.88	0.93	0.63	0.66	0.68
Naïve Bayasian	LRCFs and Scoring Functions	0.84	0.72	0.84	0.01	0.00	0.00
	LRCFs	0.84	0.72	0.84	0.00	0.00	0.00
	Scoring Functions	0.87	0.81	0.87	0.49	0.48	0.48
kNN[e]	LRCFs and Scoring Functions	0.93	0.87	0.94	0.69	0.64	0.76
	LRCFs	0.88	0.80	0.90	0.43	0.37	0.52
	Scoring Functions	0.93	0.87	0.94	0.68	0.64	0.75
MLP[f]	LRCFs and Scoring Functions	0.89	0.82	0.99	0.56	0.49	0.96
	LRCFs	0.89	0.82	0.99	0.57	0.50	0.96
	Scoring Functions	0.87	0.79	0.88	0.42	0.38	0.44

[a] L20%out: Leave 20% out cross-validation for accuracy and Cohen's Kappa.
[b] Testing: Accuracy and Cohen's Kappa determined against the testing set (marked with asterisks in Table 2).
[c] Re-substitution: Accuracy and Cohen's Kappa determined by applying the particular ML model to predict the same training compounds used to build the model.
[d] Performed using optimized nu value of 0.1.
[e] Performed using optimized k value of 6 (number of neighbors).
[f] Performed using 3 hidden layers, 100 neuron per layer, and 100 iterations (epochs).
[g] RF ML models were validated by Out-of-Bag validation instead of Leave 20% out cross-validation for accuracy and Cohen's Kappa.

within their top 20%), the best three MLs (Xgboost, RF and nu-SVM) performed slightly worse (see Supplementary Table SM-3). It is noteworthy to mention that under such restrictions (imposing consensus score ≥ 8), the number of training compounds fell from 85 to 42, whereas testing compounds decreased from 23 to 13. This is because not all screened compounds have such high ranking docked poses (consensus score ≥ 8). This reduction in training and testing data should restrict the applicability domain of the respective ML model [181]. Moreover, imposing docking consensus score ≥ 8 will undoubtedly lower the number of docking hits captured in a virtual screening campaign because many potential hits fail to reach high quality docked poses.

3.3. Virtual screening and prediction of hits' anti-TMPRSS2 bioactivities

To utilize optimal ML models to search for potential anti-TMPRSS2 inhibitors within FDA-approved drugs, it was decided to screen 155 drug molecules of reactive warheads (Supplementary Table SM-2). All screened molecules were not tested before against TMPRSS2.

Virtual screening commenced by properly ionizing and tautomerizing screened molecules. Subsequently, they were docked, scored and RMSD-filtered utilizing the same settings implemented for the training and testing sets. The resulting docked poses were then used to generate corresponding LRCFs in exactly the same manner as in the training and testing sets. Subsequently, the resulting LRCFs were substituted in the best ML models, namely, Xgboost, nu-SVM and RF, to predict the activity label of each docked pose/conformer. This resulted in a situation where each screened compound yielded a set of poses that are assigned either "active" or "inactive" labels. This prompted us to define a threshold by which to consider certain screened molecule as being promising or not, i.e., as anti-TMPRSS2, based on the ratio of docked poses/conformers predicted to be "active" compared to those predicted to be "inactive". We decided that the most reasonable way to define such a threshold is to evaluate the active/inactive ratios within the testing set. It can be reasonably assumed the least active-to-inactive ratio of unequivocally documented active inhibitor represents an acceptable threshold for identifying potentially new active hits. Table 4 shows the percentages of active poses of testing compounds as predicted by the top three ML (i.e., Xgboost, nu-SVM and RF).

Clearly from Table 4, the anti-TMPRSS2 inhibitor **2** (camostate, Ki = 1000 nM, Table 2) shows the smallest ratios of predicted active-to-inactive docked poses among other inhibitors in the testing set, and therefore, it can be used to discriminate actives among screened compounds (Supplementary Table SM-2). Still, careful assessment of Table 4 shows that two inactives, namely, **31** (Ritonavir) and **65** (Atazanavir), to have predicted active-inactive ratios above than the proposed thresholds. Nevertheless, the total number of docked poses of these outliers are rather low (**31** has 10 docked poses, while **65** has only 2 docked poses) and way from the number of docked poses of active compounds (average = 123.6, minimum = 102). On the other hand, it is noteworthy to mention that camostate has the highest number of docked poses among all active testing compounds (204 poses), which further highlights the importance of this point.

Table 5 shows the predicted active-to-inactive ratios of the screened compounds. Clearly all three top MLs agreed on three drugs to exhibit active-to-inactive ratios exceeding the corresponding thresholds of camostat in Table 4. The three compounds are: **116** (aspoxicillin), **126** (capreomycin), and **196** (fosamprenavir). Moreover, these compounds have considerable number of docked poses, namely, 4404 for the two forms of capreomycin, 210 for aspoxicillin and 65 for fosamprenavir. Still, only capreomycin and aspoxicillin have their count of docked poses exceeding the minimum of docked poses of active testing compounds (i.e., **102**), suggesting the lesser propensity of fosamprenavir as valid TMPRSS2 inhibitor. To validate our overall computational

Table 4
Predicted active and inactive docked poses for testing set compounds.

Compounds[a]		Predicted number of active and inactive docked poses								
Numbers	Anti-TMPRSS2 Activity	Xgboost			nu-SVM			RF		
		Active Poses	Inactive Poses	Percent Active Poses[b]	Active Poses	Inactive Poses	Percent Active Poses[b]	Active Poses	Inactive Poses	Percent Active Poses[b]
5	Active	100	4	96.2	94	10	90.4	102	2	98.1
9	Active	97	5	95.1	97	5	95.1	99	3	97.1
10	Active	90	14	86.5	79	25	76	84	20	80.8
14	Active	103	1	99	99	5	95.2	102	2	98.1
2[c]	Active	39	165	19.1	43	161	21.1	26	178	12.7
19	Inactive	2	45	4.3	6	41	12.8	0	47	0
21	Inactive	0	8	0	0	8	0	3	5	37.5
24	Inactive	0	141	0	1	140	0.7	0	141	0
26	Inactive	2	954	0.2	3	953	0.3	1	955	0.1
31	Inactive	5	5	50	2	8	20	3	7	30
33	Inactive	18	12	60	1	29	3.3	7	23	23.3
39	Inactive	0	18	0	0	18	0	0	18	0
53	Inactive	0	7	0	0	7	0	0	7	0
56	Inactive	0	2	0	0	2	0	0	2	0
57	Inactive	0	13	0	0	13	0	0	13	0
62	Inactive	0	2	0	0	2	0	0	2	0
65	Inactive	1	1	50	1	1	50	1	1	50
69	Inactive	0	2	0	0	2	0	0	2	0
70	Inactive	0	13	0	0	13	0	0	13	0
84	Inactive	0	92	0	1	91	1.1	0	92	0
85	Inactive	4	163	2.4	1	166	0.6	0	167	0
93	Inactive	2	43	4.4	5	40	11.1	3	42	6.7
103	Inactive	0	10	0	1	9	10	0	10	0

[a] Compounds' numbers and bioactivities are as in Table 2.

[b] Determined by dividing the number of active poses by the total number poses (active + inactive).

[c] The percent active poses of this compound (camostat, Ki = 1000 nM) were used as threshold to classify screened compounds into potential active and inactive TMPRSS2 inhibitors.

Table 5
Screened drug molecules and counts of their predicted "active" and "inactive" docked poses as predicted by the three top MLs.

Compounds[a]		Predicted number of active and inactive docked poses								
		Xgboost			nu-SVM			RF		
		Active Poses	Inactive Poses	%Active Poses[b]	Active Poses	Inactive Poses	%Active Poses[b]	Active Poses	Inactive Poses	%Active Poses[b]
108	5-Iodo-Sunitinib	1	218	0.5	0	219	0	0	219	0
109	5-CH$_3$O-Sunitinib	2	253	0.8	2	253	0.8	1	254	0.4
110	Acenocoumarol	1	10	9.1	0	11	0	0	11	0
111	Acrivastine	0	24	0	0	24	0	0	24	0
112	Alectinib	0	38	0	0	38	0	0	38	0
113	Alogliptin	0	32	0	0	32	0	0	32	0
114	Apalcillin	4	284	1.4	5	283	1.7	1	287	0.3
115	Apraclonidine	0	1	0	0	1	0	0	1	0
116	Aspoxicillin[c]	89	121	42.4	75	135	35.7	97	113	46.2
117	Ast-1306	1	523	0.2	12	512	2.3	0	524	0
118	Bacampicillin	34	157	17.8	25	166	13.1	21	170	11
119	Bambuterol	0	42	0	1	41	2.4	0	42	0
120	Baricitinib	0	25	0	0	25	0	0	25	0
121	Belinostat	0	78	0	0	78	0	0	78	0
122	Benznidazole	0	12	0	0	12	0	0	12	0
123	Biapenem	0	7	0	0	7	0	0	7	0
124	Brodimoprim	1	29	3.3	2	28	6.7	2	28	6.7
125	Canertinib	10	1284	0.8	19	1275	1.5	0	1294	0
126	Capreomycin1[c,d]	1183	931	56	1203	911	56.9	1590	524	75.2
	Capreomycin2[c,d]	1313	977	57.3	1269	1021	55.4	1725	565	75.3
127	Carisoprodol	0	8	0	0	8	0	0	8	0
128	Carumonam	0	11	0	0	11	0	1	10	9.1
129	Cefadroxil	0	87	0	11	76	12.6	0	87	0
130	Cefamandole	0	71	0	0	71	0	0	71	0
131	Cefatrizine	2	575	0.3	17	560	2.9	2	575	0.3
132	Cefazedone	0	57	0	0	57	0	0	57	0
133	Cefazolin	0	51	0	0	51	0	0	51	0
134	Cefbuperazone	0	66	0	0	66	0	0	66	0
135	Cefcapene	0	115	0	0	115	0	0	115	0
136	Cefclidin	0	181	0	0	181	0	1	180	0.6
137	Cefepime	2	105	1.9	0	107	0	0	107	0
138	Cefetamet	0	52	0	0	52	0	0	52	0
139	Cefixime	0	4	0	0	4	0	0	4	0
140	Cefmenoxime	0	176	0	0	176	0	0	176	0
141	Cefmetazole	0	25	0	0	25	0	0	25	0
142	Cefminox	0	80	0	4	76	5	0	80	0
143	Cefodizime	0	43	0	0	43	0	0	43	0
144	Cefonicid	0	94	0	0	94	0	0	94	0
145	Ceforanide	1	166	0.6	0	167	0	2	165	1.2
146	Cefotaxime	0	112	0	0	112	0	0	112	0
147	Cefotetan	0	79	0	0	79	0	0	79	0
148	Cefotiam	0	183	0	1	182	0.5	0	183	0
149	Cefoxitin	2	33	5.7	1	34	2.9	0	35	0
150	Cefozopran	0	270	0	1	269	0.4	0	270	0
151	Cefpimizole	0	179	0	1	178	0.6	1	178	0.6
152	Cefpiramide	3	309	1	0	312	0	1	311	0.3
153	Cefpirome	1	157	0.6	0	158	0	0	158	0
154	Cefpodoxime	0	74	0	0	74	0	0	74	0
155	Cefprozil	0	3	0	0	3	0	0	3	0
156	Cefroxadine	0	58	0	0	58	0	0	58	0
157	Cefteram	0	16	0	0	16	0	0	16	0
158	Ceftezole	0	38	0	0	38	0	0	38	0
159	Ceftibuten	0	61	0	0	61	0	0	61	0
160	Ceftizoxime	1	42	2.3	2	41	4.7	0	43	0
161	Ceftriaxone	0	186	0	0	186	0	0	186	0
162	Cefuroxime	0	59	0	1	58	1.7	0	59	0
163	Cefuzonam	0	168	0	0	168	0	0	168	0
164	Cephacetrile	0	2	0	0	2	0	0	2	0
165	Cephaloridine	0	23	0	1	22	4.3	0	23	0
166	Cephalothin	0	19	0	0	19	0	0	19	0
167	Cephapirin	0	38	0	2	36	5.3	0	38	0
168	Cephradine	1	44	2.2	0	45	0	0	45	0
169	Cilastatin	0	113	0	3	110	2.7	1	112	0.9
170	Cinanserin	0	188	0	0	188	0	0	188	0
171	Citalopram	0	9	0	0	9	0	0	9	0
172	Clobenprobit	0	138	0	0	138	0	0	138	0
173	Clonazepam	0	5	0	0	5	0	0	5	0
174	Clonitazene	0	41	0	0	41	0	0	41	0
175	Cyclacillin	0	23	0	2	21	8.7	0	23	0
176	Cypermethrin	0	57	0	0	57	0	0	57	0

(*continued on next page*)

Table 5 (*continued*)

Compounds[a]		Predicted number of active and inactive docked poses								
		Xgboost			nu-SVM			RF		
		Active Poses	Inactive Poses	%Active Poses[b]	Active Poses	Inactive Poses	%Active Poses[b]	Active Poses	Inactive Poses	%Active Poses[b]
177	Dacomitinib	0	430	0	3	427	0.7	0	430	0
178	Dantrolene	0	91	0	0	91	0	0	91	0
179	Debrisoquin	0	1	0	0	1	0	0	1	0
180	Demecarium	1	12	7.7	3	10	23.1	0	13	0
181	Difenoxin	0	26	0	0	26	0	0	26	0
182	Diphenoxylate	1	83	1.2	2	82	2.4	0	84	0
183	Enzalutamide	0	16	0	1	15	6.3	0	16	0
184	Ertapenem	0	21	0	0	21	0	1	20	4.8
185	Eszopiclone	0	41	0	0	41	0	0	41	0
186	Ethacrynic_Acid	0	1	0	0	1	0	0	1	0
187	Etonitazene	0	55	0	0	55	0	0	55	0
188	Etozolin	0	5	0	0	5	0	0	5	0
189	Etretinate	0	7	0	0	7	0	1	6	14.3
190	Ezogabine	0	144	0	0	144	0	0	144	0
191	Famitinib	1	323	0.3	1	323	0.3	0	324	0
192	Flubanilate	0	11	0	0	11	0	0	11	0
193	Flucloxacillin	0	10	0	0	10	0	0	10	0
194	Flunidazole	0	9	0	0	9	0	0	9	0
195	Flunitrazepam	0	2	0	0	2	0	0	2	0
196	Fosamprenavir[c]	22	43	33.8	30	35	46.2	20	45	30.8
197	Guanadrel	0	8	0	1	7	12.5	0	8	0
198	Guanethidine	1	4	20	1	4	20	0	5	0
199	Guanfacine	0	10	0	0	10	0	0	10	0
200	Guanisoquin	0	3	0	0	3	0	0	3	0
201	Guanoxan	0	12	0	0	12	0	0	12	0
202	Guanoxyfen	0	11	0	0	11	0	0	11	0
203	Henatinib	0	1	0	0	1	0	0	1	0
204	Hetacillin	0	21	0	1	20	4.8	0	21	0
205	Hydroxystilbamidine	0	2	0	0	2	0	0	2	0
206	Ibrutinib	1	387	0.3	2	386	0.5	0	388	0
207	Imipenem	0	16	0	0	16	0	0	16	0
208	Iobenguane	0	5	0	0	5	0	0	5	0
209	Irinotecan	1	24	4	0	25	0	2	23	8
210	Isosulfazecin	1	93	1.1	5	89	5.3	1	93	1.1
211	Leuprolide	0	2	0	0	2	0	2	0	100
212	Levocabastine	0	8	0	1	7	12.5	0	8	0
213	Levofuraltadone	0	29	0	0	29	0	0	29	0
214	Levopropylcillin	0	2	0	0	2	0	0	2	0
215	Lodoxamide	0	27	0	0	27	0	0	27	0
216	Loracarbef	0	55	0	5	50	9.1	0	55	0
217	Mebendazole	0	82	0	0	82	0	0	82	0
218	Methicillin	0	7	0	0	7	0	0	7	0
219	Mitomycin	0	8	0	0	8	0	0	8	0
220	Momelotinib	0	125	0	0	125	0	0	125	0
221	Moricizine	0	219	0	0	219	0	0	219	0
222	Naquotinib	4	330	1.2	5	329	1.5	4	330	1.2
223	Neratinib	7	324	2.1	13	318	3.9	13	318	3.9
224	Niclosamide	0	16	0	0	16	0	0	16	0
225	Nifurtimox	0	3	0	0	3	0	0	3	0
226	Nilutamide	0	2	0	0	2	0	0	2	0
227	Nitisinone	0	6	0	0	6	0	0	6	0
228	Nitrazepam	0	6	0	0	6	0	0	6	0
229	Nitrofurantoin	0	35	0	0	35	0	0	35	0
230	Octocrylene	0	12	0	0	12	0	0	12	0
231	OctylMethoxycinnamic	0	16	0	0	16	0	0	16	0
232	Olmutinib	8	566	1.4	12	562	2.1	6	568	1
233	Orantinib	1	31	3.1	1	31	3.1	0	32	0
234	Osimertinib	34	1031	3.2	48	1017	4.5	22	1043	2.1
235	Oxacilin	0	16	0	2	14	12.5	0	16	0
236	Oxamniquine	0	26	0	0	26	0	0	26	0
237	Pelitinib	1	229	0.4	2	228	0.9	1	229	0.4
238	Penicillin G	0	12	0	0	12	0	0	12	0
239	Penicillin V	0	20	0	0	20	0	0	20	0
240	Pentagastrin	5	566	0.9	5	566	0.9	89	482	15.6
241	Pericyazine	0	41	0	1	40	2.4	0	41	0
242	Piritrexim	31	547	5.4	24	554	4.2	3	575	0.5
243	Poziotinib	0	167	0	1	166	0.6	0	167	0
244	Proguanil	0	11	0	0	11	0	0	11	0
245	Pyrotinib	4	143	2.7	0	147	0	3	144	2
246	Rescinnamine	1	7	12.5	0	8	0	0	8	0
247	Rociletinib	7	910	0.8	4	913	0.4	9	908	1

Table 5 (*continued*)

Compounds[a]		Predicted number of active and inactive docked poses								
		Xgboost			nu-SVM			RF		
		Active Poses	Inactive Poses	%Active Poses[b]	Active Poses	Inactive Poses	%Active Poses[b]	Active Poses	Inactive Poses	%Active Poses[b]
248	Romidepsin	0	3	0	0	3	0	0	3	0
249	Siguazodan	1	53	1.9	3	51	5.6	0	54	0
250	Spebrutinib	0	880	0	5	875	0.6	0	880	0
251	Streptomycin	2	97	2	4	95	4	18	81	18.2
252	Tedizolid-phosphate	0	2	0	0	2	0	0	2	0
253	Tegaserod	0	68	0	0	68	0	0	68	0
254	Tepotinib	1	88	1.1	1	88	1.1	0	89	0
255	Ticarcillin	0	26	0	0	26	0	0	26	0
256	Vasopressin	0	2	0	0	2	0	0	2	0
257	Verapamil	0	79	0	0	79	0	0	79	0
258	Vilazodone	0	99	0	0	99	0	1	98	1
259	Vorapaxar	1	35	2.8	1	35	2.8	0	36	0
260	Yn-968d1	0	86	0	1	85	1.2	0	86	0
261	Zaleplon	0	3	0	0	3	0	0	3	0
262	Zopiclone	0	37	0	1	36	2.7	0	37	0

[a] Compounds are as in Supplementary Table SM-2.

[b] Determined by dividing the number of active poses by the total number poses (active + inactive).

[c] The percent active poses of these compounds exceeded the threshold in Table 2 (compound **2** or camostat) and therefore are predicted to be active anti-TMPRSS2 inhibitors.

method, we evaluated the three promising hits by molecular dynamics simulation and covalent docking.

3.4. Molecular dynamics simulation and covalent docking

As a prerequisite for successful covalent bond formation between the ligand's warhead and the targeted nucleophilic center within the binding site (hydroxyl of Ser478), it is necessary for the warhead to reside for some time, e.g., 20 to 80 ns [177,178], at close proximity to the nucleophilic residue. It can be assumed that the sum of van der Waals' radii of oxygen and carbon atoms (3.3 Å, representing the hydroxyl oxygen of Ser478 and carbon atoms of electrophilic warheads, magenta line in Fig. 7) is a reasonable distance threshold for probable subsequent nucleophilic attack [177].

To evaluate the residence time and distance separating the reactive warheads of capreomycin, aspoxicillin and fosamprenavir from the nucleophilic hydroxyl of Ser478, we opted to perform 200-ns MD simulations for the three hits as they dock within the binding pocket. The starting poses were selected to have the highest consensus score among other docked poses in each case. MD trajectories show fosamprenavir to leave the binding pocket after approximately 60 ns. Moreover, it failed to have close encounters with the nucleophilic Ser478 OH (within distance threshold) as can be seen in Fig. 7A where the distance separating the electrophilic carbamate atom of fosamprenavir and the nucleophilic Ser478 OH never crossed the threshold magenta line. Accordingly, we were prompted to discard fosamprenavir from subsequent covalent docking. However, the other two hits (capreomycin and aspoxicillin) remained during MD simulation within the active site (Fig. 7A). However, aspoxicillin seems to have better chances to form covalent bond with Ser478 as it crosses/approaches the distance theshold line (magneta line in Fig. 7A) more frequently.

Covalent docking shows both aspoxicillin and capreomycin to successfully form covalent bonds with the nucleophilic hydroxyl of Ser478, as in Fig. 8. In aspoxicillin case, the nucleophilic hydroxyl attacks and ring-opens the β-lactam ring (Fig. 8A and 8B), while in capreomycin case an oxa-Michael addition reaction takes place whereby the nucleophilic hydroxyl adds to the Michael acceptor feature within capreomycin (Fig. 8C and D). Moreover, covalent docking shows that each docked compound exhibits additional reversible binding interactions: Aspoxicillin forms hydrogen bonding interactions with Glu426, Ser497 and Gly476 together with hydrophobic interactions Cys502 and His333 (Fig. 8A and B). Similarly, docked capreomycin is involved in hydrogen bonding interactions with Gln475, His333, Ser497, Gly476, His316, and Gly428. Additionally, the guanidine of capereomycin is stacked against the imidazole side chain of His333.

Covalent docking, however, is only a decorative tool to suggest how a covalent–bond forming drug might fit into the binding pocket in case it succeeds in forming covalent bond. However, we believe the main supportive tool in our case is MD simulations. To enhance confidence in MD simulations as success gauge for our ML models, we performed MD simulations (200 ns) for further 4 compounds (starting with high-quality docked poses of consensus score of 9), namely, camostat (2), cobicistat (**33**), ibrutinib (**206**) and piritrexim (**242**). The former two belong to the testing set, while the latter two belong to the screened set. Camostat is the only active compound among the collection and it was correctly predicted by our ML models to be active. While the other three were all predicted to be inactive by our models, however one of them, i.e., cobicistat, is experimentally proven to be inactive. This collection should allow us to adequately probe the correlation between our ML predictions and MD simulation data. Fig. 7B shows the results of the MD study. Clearly, only camostat persisted within the vicinity of the binding site, while the rest, whether experimentally inactive or predicted to be inactive, were quickly ejected from the binding site. These results nicely correlate with the predictions of our ML models despite being generated for rather diverse set of compounds.

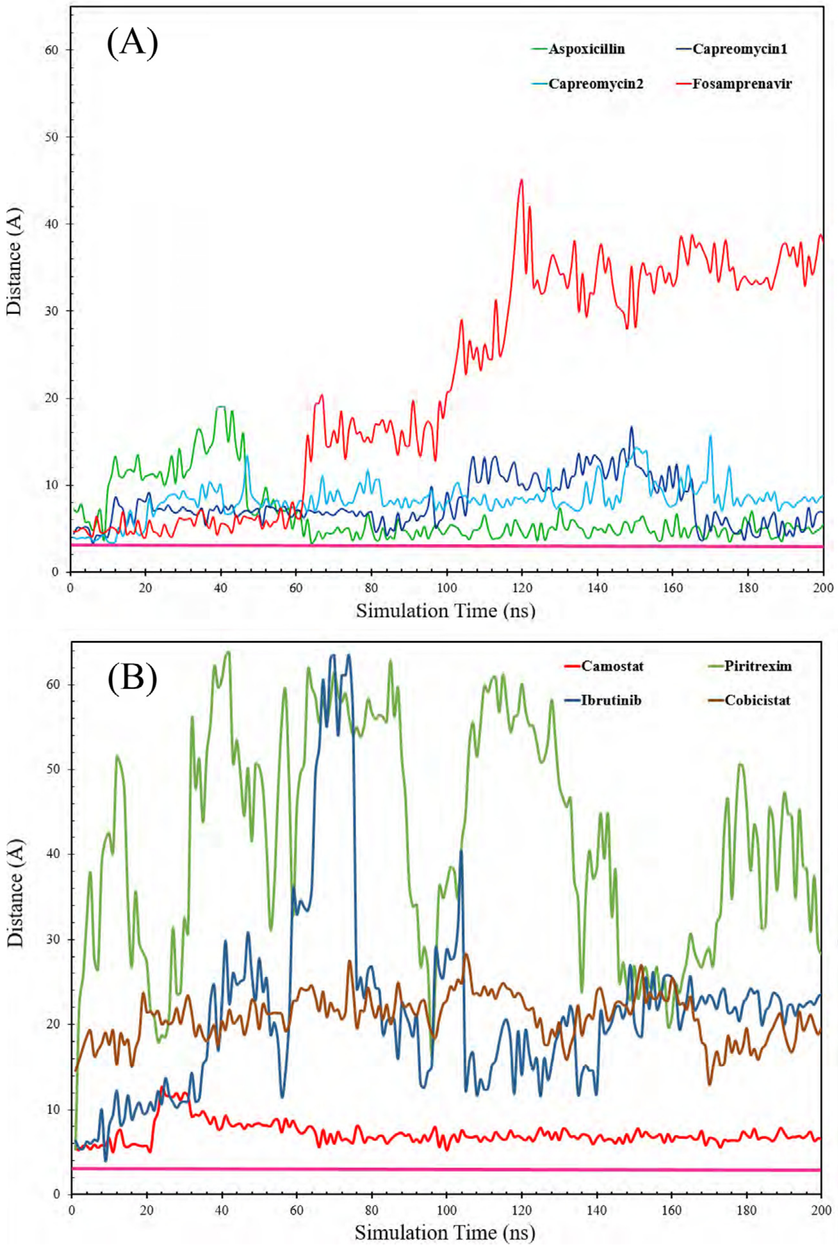


Fig. 7. MD simulations distance probes **(A)** Distances separating the reactive warheads of candidate hits: capreomycin (**126**), aspoxicillin (**116**) and fosamprenavir (**196**) from the nucleophilic hydroxy of Ser478 during 200 ns of MD trajectories. **(B)** Distances separating the reactive warheads of camostat (**2**, active testing compound), ibrutinib (**206,** predicted inactive) or cobicistat (**33**, inactive testing compound) or the central atom of piritrexim (**242,** predicted inactive) from the nucleophilic hydroxy of Ser478 during 200 ns of MD trajectories. Each time step represents 1.0 ns. The magenta lines represent the minimum distance between the reactive warhead atoms of the hits and Ser478 hydroxyl oxygen atom in a covalent-bond productive encounter (the distance represents the summation of van der Waals' radii of carbon and oxygen atoms). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

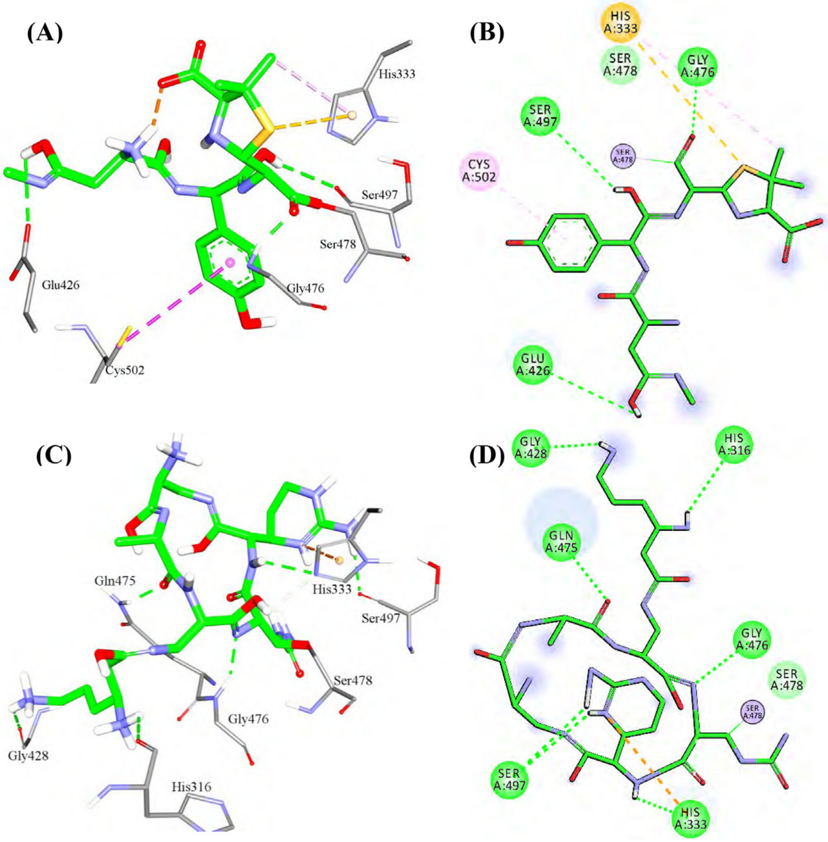


Fig. 8. The 3-D and 2-D structure of covalent docked complexes tying the reactive warheads (A) and (B) Aspoxicilin, (C) and (D) Capreomycin 1 with the hydroxy of Ser478.

4. Conclusions

In the present work we introduce computational bootstrapping of machine learning QSAR modelling using multiple high-quality docked poses. Ligand-receptor contact fingerprints and scoring function values were used as descriptors, while several MLs were scanned. We implemented this method for the discovery of potential inhibitors for the serine protease enzyme TMPRSS2 involved in the infectivity of coronaviruses. Three hits were identified. Subsequent molecular dynamic simulation and covalent docking supported the results of the new computational approach.

CRediT authorship contribution statement

Ma'mon M. Hatmal: Resources, Formal analysis, Investigation, Writing - review & editing. **Omar Abuyaman:** Formal analysis, Investigation, Resources, Writing – original draft, Writing - review & editing. **Mutasem Taha:** Conceptualization, Methodology, Supervision, Investigation, Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Dr. Eyad Qunaibi and Mr. Ramzi Saeed for their technical assistance. The authors also thank the Deanships of Academic Research at the University of Jordan and the Hashemite University for funding this project. The authors are also indebted for the Science Research Fund/Jordan Ministry of Higher Education for their generous grant.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.08.023.

References

[1] Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. Briefings Bioinf 2009;10(5):579–91.

[2] Pantsar T, Poso A. Binding Affinity via Docking: Fact and Fiction. Molecules (Basel, Switzerland). 2018;23(8):1899.

[3] Jorgensen WL. Efficient drug lead discovery and optimization. Acc Chem Res 2009;42(6):724–33.

[4] Pinzi L, Rastelli G. Molecular Docking: Shifting Paradigms in Drug Discovery. Int J Mol Sci. 2019;20(18):4331.

[5] Ashtawy HM, Mahapatra NR. Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. J Chem Inf Model 2017;58(1):119–33.

[6] Toviwek B, Gleeson D, Gleeson M. QM/MM and molecular dynamics investigation of the mechanism of covalent inhibition of TAK1 kinase. Org Biomol Chem 2021;19(6):1412–25.

[7] Bissantz C, Folkers G, Rognan D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. J Med Chem 2000;43(25):4759–67.

[8] Cavasotto CN, Aucar MG. High-Throughput Docking Using Quantum Mechanical Scoring. Front Chem. 2020;8:246.

[9] Cosconati S, Forli S, Perryman AL, Harris R, Goodsell DS, Olson AJ. Virtual Screening with AutoDock: Theory and Practice. Expert Opin Drug Discov. 2010;5(6):597–607.

[10] Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. J Comput Aided Mol Des 1997;11(5):425–45.

[11] Ewing TJA, Makino S, Skillman AG, Kuntz ID. J Comput Aided Mol Des 2001;15(5):411–28.

[12] Gehlhaar DK, Bouzida D, Rejto PA. Reduced Dimensionality in Ligand—Protein Structure Prediction: Covalent Inhibitors of Serine Proteases and Design of Site-Directed Combinatorial Libraries. ACS Symposium Series: American Chemical Society 1999:292–311.
[13] Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295(2):337–56.
[14] Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J Med Chem 2004;47(7):1750–9.
[15] Jain AN. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 1996;10(5):427–40.
[16] Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. LigScore: a novel scoring function for predicting binding affinities. J Mol Graph Model 2005;23(5):395–407.
[17] Lans I, Palacio-Rodríguez K, Cavasotto CN, Cossio P. Flexi-pharma: a molecule-ranking strategy for virtual screening using pharmacophores from ligand-free conformational ensembles. J Comput Aided Mol Des 2020;34(10):1063–77.
[18] Li J, Fu A, Zhang L. An Overview of Scoring Functions Used for Protein-Ligand Interactions in Molecular Docking. Interdisciplinary Sciences: Computational Life Sciences. 2019;11(2):320–8.
[19] Liu K, Kokubo H. Prediction of ligand binding mode among multiple cross-docking poses by molecular dynamics simulations. J Comput Aided Mol Des 2020;34(11):1195–205.
[20] Michel J, Essex JW. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. J Comput Aided Mol Des 2010;24(8):639–58.
[21] Muegge I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. Virtual Screening: An Alternative or Complement to High Throughput Screening?: Kluwer Academic Publishers. p. 99-114.
[22] Muegge I. Effect of ligand volume correction on PMF scoring. J Comput Chem 2001;22(4):418–25.
[23] Muegge I. PMF Scoring Revisited. J Med Chem 2006;49(20):5895–902.
[24] Muegge I, Martin YC. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. J Med Chem 1999;42(5):791–804.
[25] Pujadas G, Vaque M, Ardevol A, Blade C, Salvado M, Blay M, et al. Protein-ligand Docking: A Review of Recent Advances and Future Perspectives. Curr Pharm Anal 2008;4(1):1–19.
[26] Rao SN, Head MS, Kulkarni A, LaLonde JM. Validation Studies of the Site-Directed Docking Program LibDock. J Chem Inf Model 2007;47(6):2159–71.
[27] Rarey M, Kramer B, Lengauer T, Klebe G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. J Mol Biol 1996;261(3):470–89.
[28] Sakano T, Mahamood MI, Yamashita T, Fujitani H. Molecular dynamics analysis to evaluate docking pose prediction. Biophys Physicobiol. 2016;13:181–94.
[29] Sulimov VB, Kutov DC, Sulimov AV. Advances in Docking. Curr Med Chem 2020;26(42):7555–80.
[30] Wang R, Liu L, Lai L, Tang Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. J Mol Model 1998;4(12):379–94.
[31] Bohm HJ. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des. 1998;12(4):309–23.
[32] Rajamani R, Good AC. Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. Curr Opin Drug Discov Devel. 2007;10(3):308–15.
[33] Boyd S. FlexX suite. Chem World-Uk 2007. p.:72-.
[34] Andrew RL, Brian KS, Catherine EP. Prediction of Protein−Ligand Interactions. Docking and Scoring: Successes and Gaps. J Med Chem 2006;49:5851–5.
[35] Krovat EM, Langer T. Impact of Scoring Functions on Enrichment in Docking-Based Virtual Screening: An Application Study on Renin Inhibitors†. J Chem Inf Comput Sci 2004;44(3):1123–9.
[36] Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug Discovery Today 2006;11(13–14):580–94.
[37] Krissinel E. Crystal contacts as nature's docking solutions. J Comput Chem 2010;31(1):133–43.
[38] Taha MO, Habash M, Al-Hadidi Z, Al-Bakri A, Younis K, Sisan S. Docking-Based Comparative Intermolecular Contacts Analysis as New 3-D QSAR Concept for Validating Docking Studies and in Silico Screening: NMT and GP Inhibitors as Case Studies. J Chem Inf Model 2011;51(3):647–69.
[39] Jaradat NJ, Khanfar MA, Habash M, Taha MO. Combining docking-based comparative intermolecular contacts analysis and k-nearest neighbor correlation for the discovery of new check point kinase 1 inhibitors. J Comput Aided Mol Des 2015;29(6):561–81.
[40] Taha MO, Habash M, Khanfar MA. The use of docking-based comparative intermolecular contacts analysis to identify optimal docking conditions within glucokinase and to discover of new GK activators. J Comput Aided Mol Des 2014;28(5):509–47.
[41] Al-Sha'er MA, Taha MO. Application of docking-based comparative intermolecular contacts analysis to validate Hsp90α docking studies and subsequent in silico screening for inhibitors. J Mol Model 2012;18(11):4843–63.
[42] MmM Hatmal, Jaber S, Taha MO. Combining molecular dynamics simulation and ligand-receptor contacts analysis as a new approach for pharmacophore modeling: beta-secretase 1 and check point kinase 1 as case studies. J Comput Aided Mol Des 2016;30(12):1149–63.
[43] MmM Hatmal, Taha MO. Simulated annealing molecular dynamics and ligand–receptor contacts analysis for pharmacophore modeling. Future. Med Chem 2017;9(11):1141–59.
[44] MmM Hatmal, Taha MO. Combining Stochastic Deformation/Relaxation and Intermolecular Contacts Analysis for Extracting Pharmacophores from Ligand-Receptor Complexes. J Chem Inf Model 2018;58(4):879–93.
[45] Ballester PJ, Schreyer A, Blundell TL. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? J Chem Inf Model 2014;54(3):944–55.
[46] Da C, Kireev D. Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. J Chem Inf Model 2014;54(9):2555–61.
[47] Deng Z, Chuaqui C, Singh J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. J Med Chem 2004;47(2):337–44.
[48] Kelly MD, Mancera RL. Expanded Interaction Fingerprint Method for Analyzing Ligand Binding Modes in Docking and Structure-Based Drug Design. J Chem Inf Comput Sci 2004;44(6):1942–51.
[49] Lenselink EB, Jespers W, van Vlijmen HWT, Ijzerman AP, van Westen GJP. Interacting with GPCRs: Using Interaction Fingerprints for Virtual Screening. J Chem Inf Model 2016;56(10):2053–60.
[50] Mpamhanga CP, Chen B, McLay IM, Willett P. Knowledge-Based Interaction Fingerprint Scoring: A Simple Method for Improving the Effectiveness of Fast Scoring Functions. J Chem Inf Model 2006;46(2):686–98.
[51] Pérez-Nueno VI, Rabal O, Borrell JI, Teixidó J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. J Chem Inf Model 2009;49(5):1245–60.
[52] Marcou G, Rognan D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. J Chem Inf Model 2006;47(1):195–207.
[53] Rácz A, Bajusz D, Héberger K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. Journal of cheminformatics. 2018;10(1):48-.
[54] Bajusz D, Ferenczy GG, Keserű GM. Discovery of Subtype Selective Janus Kinase (JAK) Inhibitors by Structure-Based Virtual Screening. J Chem Inf Model 2015;56(1):234–47.
[55] Bishop CM. Model-based machine learning. Philos Trans A Math Phys Eng Sci. 2012;371(1984):20120222-.
[56] Duan Y, Edwards JS, Dwivedi YK. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. Int J Inf Manage 2019;48:63–71.
[57] Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects. Drug Discovery Today 2019;24(10):2017–32.
[58] Lee J-G, Jun S, Cho Y-W, Lee H, Kim GB, Seo JB, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol. 2017;18(4):570–84.
[59] Naz K, Naz A, Ashraf ST, Rizwan M, Ahmad J, Baumbach J, et al. PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. BMC Bioinf 2019;20(1):123.
[60] Ong E, Wang H, Wong MU, Seetharaman M, Valdez N, He Y. Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. Bioinformatics 2020;36(10):3185–91.
[61] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
[62] Zhong F, Xing J, Li X, Liu X, Fu Z, Xiong Z, et al. Artificial intelligence in drug design. Science China Life Sciences. 2018;61(10):1191–204.
[63] Carpenter KA, Cohen DS, Jarrell JT, Huang X. Deep learning and virtual drug screening. Future Med Chem 2018;10(21):2557–67.
[64] Chao T, Hongbo P, Yansheng L, Zhengrou Z. Unsupervised Spectral-Spatial Feature Learning With Stacked Sparse Autoencoder for Hyperspectral Imagery Classification. IEEE Geosci Remote Sens Lett 2015;12(12):2438–42.
[65] Cloutier L, Sirois S. Bayesian versus Frequentist statistical modeling: A debate for hit selection from HTS campaigns. Drug Discovery Today 2008;13(11–12):536–42.
[66] Karthiga B, Rekha M. Feature extraction and I-NB classification of CT images for early lung cancer detection. Mater Today: Proc 2020;33:3334–41.
[67] Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today 2015;20(3):318–31.
[68] Wickramasinghe I, Kalutarage H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Comput 2020.
[69] Ren X, Guo H, Li S, Wang S, Li J. In: A Novel Image Classification Method with CNN-XGBoost Model. Digital Forensics and Watermarking. Springer International Publishing; 2017. p. 378–90.
[70] Rozinajová V, Ezzeddine AB, Lóderer M, Loebl J, Magyar R, Vrablecová P. Computational Intelligence in Smart Grid Environment. In: Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications. Elsevier; 2018. p. 23–59.
[71] Umamaheswari C, Bhavani, R. R., & Thirunadana Sikamani, K. A comparative study on various data mining classification methods: KNN, PNN and ANN for tiles defect detection. International Journal of Pure and Applied Mathematics.

International Journal of Pure and Applied Mathematics. 2018;118(Special Issue 9):389–405.
[72] Heikamp K, Bajorath J. Support vector machines for drug discovery. Expert Opin Drug Discov 2013;9(1):93–104.
[73] Jayaraj PB, Jain S. Ligand based virtual screening using SVM on GPU. Comput Biol Chem 2019;83:107143.
[74] Hajmeer M, Basheer I. A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data. J Microbiol Methods 2002;51(2):217–26.
[75] Khosravanian A, Ayat S. Diagnosing Breast Cancer Type by Using Probabilistic Neural Network in Decision Support System. International Journal of Knowledge Engineering. 2016;2(1):73–6.
[76] Tran DH, Ng AWM, Perera BJC, Burn S, Davis P. Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. Urban Water J 2006;3(3):175–84.
[77] Wang J, Hu X, Zhu D. In: Applications of Data Mining in the Healthcare Industry. Encyclopedia of Healthcare Information Systems. IGI Global; 2008. p. 68–73.
[78] Koh HC, Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management : JHIM. 2005;19(2):64–72.
[79] Wasserman PD. Advanced methods in neural computing: John Wiley & Sons, Inc.; 1993.
[80] Gupta P, Sinha NK. In: Neural Networks for Identification of Nonlinear Systems: An Overview. Soft Computing and Intelligent Systems: Elsevier; 2000. p. 337–56.
[81] Sainlez M, Heyen G. In: Recurrent neural network prediction of steam production in a Kraft recovery boiler. Computer Aided Chemical Engineering: Elsevier; 2011. p. 1784–8.
[82] Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Machine learning. 2018;107(12):1895–922.
[83] Sahiner B, Chan H-P, Hadjiiski L. Classifier performance estimation under the constraint of a finite sample size: resampling schemes applied to neural network classifiers. Neural Netw. 2008;21(2–3):476–83.
[84] Calmettes G, Drummond GB, Vowler SL. Making do with what we have: use your bootstraps. The Journal of physiology. 2012;590(15):3403–6.
[85] Hilgenfeld R, Peiris M. From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. Antiviral Res. 2013;100(1):286–95.
[86] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181(2):271-80.e8.
[87] Khan RJ, Jha R, Amera GM, Jain M, Singh E, Pathak A, et al. Targeting Novel Coronavirus 2019: A Systematic Drug Repurposing Approach to Identify Promising Inhibitors Against 3C-like Proteinase and 2'-O-Ribose Methyltransferase. American Chemical Society (ACS) 2020.
[88] Laporte M, Naesens L. Airway proteases: an emerging drug target for influenza and other respiratory virus infections. Curr Opin Virol. 2017;24:16–24.
[89] Li Q, Wang Z, Zheng Q, Liu S. Potential clinical drugs as covalent inhibitors of the priming proteases of the spike protein of SARS-CoV-2. Comput Struct Biotechnol J. 2020;18:2200–8.
[90] Meyer D, Sielaff F, Hammami M, Böttcher-Friebertshäuser E, Garten W, Steinmetzer T. Identification of the first synthetic inhibitors of the type II transmembrane serine protease TMPRSS2 suitable for inhibition of influenza virus activation. Biochem J 2013;452(2):331–43.
[91] Redka D, Kurji N, Shahani V, Windemuth A, MacKinnon S, Landon M. PolypharmDB, a deep learning-based resource, quickly identifies repurposed drug candidates for COVID-19. American Chemical Society (ACS) 2020.
[92] Rensi S, Altman B, Liu R, Lo T, McInnes Y-C, derry a G, et al. Homology Modeling of TMPRSS2 Yields Candidate Drugs That May Inhibit Entry of SARS-CoV-2 into Human Cells. American Chemical Society (ACS) 2020.
[93] Thunders M, Delahunt B. Gene of the month: TMPRSS2 (transmembrane serine protease 2). J Clin Pathol. 2020;73(12):773–6.
[94] . Coronavirus Cases. Worldometer 2020..
[95] Barile E, Baggio C, Gambini L, Shiryaev SA, Strongin AY, Pellecchia M. Potential Therapeutic Targeting of Coronavirus Spike Glycoprotein Priming. Molecules (Basel, Switzerland). 2020;25(10):2424.
[96] Damalanka VC, Janetka JW. Recent progress on inhibitors of the type II transmembrane serine proteases, hepsin, matriptase and matriptase-2. Future Med Chem 2019;11(7):743–69.
[97] Yamamoto M, Matsuyama S, Li X, Takeda M, Kawaguchi Y, Inoue J-I, et al. Identification of Nafamostat as a Potent Inhibitor of Middle East Respiratory Syndrome Coronavirus S Protein-Mediated Membrane Fusion Using the Split-Protein-Based Cell-Cell Fusion Assay. Antimicrob Agents Chemother. 2016;60(11):6532–9.
[98] Di Cera E. Serine proteases. IUBMB Life 2009;61(5):510–5.
[99] Hoffmann M, Schroeder S, Kleine-Weber H, Müller MA, Drosten C, Pöhlmann S. Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19. Antimicrob Agents Chemother. 2020;64(6):e00754–e820.
[100] Yamamoto M, Kiso M, Sakai-Tagawa Y, Iwatsuki-Horimoto K, Imai M, Takeda M, et al. The anticoagulant nafamostat potently inhibits SARS-CoV-2 infection in vitro: an existing drug with multiple possible therapeutic effects. Cold Spring Harbor Laboratory; 2020.
[101] ClinicalTrials.gov. Identifier NCT04352400, Efficacy of Nafamostat in Covid-19 Patients (RACONA Study) (RACONA). In: (US) BMNLoM, editor. 2020.
[102] ClinicalTrials.gov. Identifier NCT04470544, CamostatMesilate Treating Patients With Hospitalized Patients With COVID-19 (RECOVER). In: (US) BMNLoM, editor. 2020.
[103] Baillie TA. Targeted Covalent Inhibitors for Drug Design. Angew Chem Int Ed 2016;55(43):13408–21.
[104] Blay V, Pei D. Serine proteases: how did chemists tease out their catalytic mechanism? ChemTexts. 2019;5(4).
[105] Kalgutkar AS, Dalvie DK. Drug discovery for a new generation of covalent drugs. Expert Opin Drug Discov. 2012;7(7):561–81.
[106] Powers JC, Asgian JL, Ekici ÖD, James KE. Irreversible Inhibitors of Serine, Cysteine, and Threonine Proteases. Chem Rev 2002;102(12):4639–750.
[107] Kitahata S, Yakushiji F, Ichikawa S. Impact of the structures of macrocyclic Michael acceptors on covalent proteasome inhibition. Chem Sci. 2017;8(10):6959–63.
[108] Lanning BR, Whitby LR, Dix MM, Douhan J, Gilbert AM, Hett EC, et al. A road map to evaluate the proteome-wide selectivity of covalent kinase inhibitors. Nat Chem Biol. 2014;10(9):760–7.
[109] Faucher F, Bennett JM, Bogyo M, Lovell S. Strategies for Tuning the Selectivity of Chemical Probes that Target Serine Hydrolases. Cell Chem Biol. 2020;27(8):937–52.
[110] Darvesh S, Darvesh KV, McDonald RS, Mataija D, Walsh R, Mothana S, et al. Carbamates with Differential Mechanism of Inhibition Toward Acetylcholinesterase and Butyrylcholinesterase. J Med Chem 2008;51(14):4200–12.
[111] Ghosh AK, Brindisi M. Organic carbamates in drug design and medicinal chemistry. J Med Chem 2015;58(7):2895–940.
[112] Bachovchin DA, Cravatt BF. The pharmacological landscape and therapeutic potential of serine hydrolases. Nat Rev Drug Discov. 2012;11(1):52–68.
[113] Bandyopadhyay A, Gao J. Targeting biomolecules with reversible covalent chemistry. Curr Opin Chem Biol. 2016;34:110–6.
[114] Berger JP, SinhaRoy R, Pocai A, Kelly TM, Scapin G, Gao Y-D, et al. A comparative study of the binding properties, dipeptidyl peptidase-4 (DPP-4) inhibitory activity and glucose-lowering efficacy of the DPP-4 inhibitors alogliptin, linagliptin, saxagliptin, sitagliptin and vildagliptin in mice. Endocrinol Diabetes Metab. 2017;1(1):e00002-e.
[115] Nabeno M, Akahoshi F, Kishida H, Miyaguchi I, Tanaka Y, Ishii S, et al. A comparative study of the binding modes of recently launched dipeptidyl peptidase IV inhibitors in the active site. Biochem Biophys Res Commun 2013;434(2):191–6.
[116] Abdeldayem A, Raouf YS, Constantinescu SN, Moriggl R, Gunning PT. Advances in covalent kinase inhibitors. Chem Soc Rev 2020;49(9):2617–87.
[117] Clerc J, Groll M, Illich DJ, Bachmann AS, Huber R, Schellenberg B, et al. Synthetic and structural studies on syringolin A and B reveal critical determinants of selectivity and potency of proteasome inhibition. Proc Natl Acad Sci U S A. 2009;106(16):6507–12.
[118] Groll M, Schellenberg B, Bachmann AS, Archer CR, Huber R, Powell TK, et al. A plant pathogen virulence factor inhibits the eukaryotic proteasome by a novel mechanism. Nature 2008;452(7188):755–8.
[119] Nising CF, Bräse S. The oxa-Michael reaction: from recent developments to applications in natural product synthesis. Chem Soc Rev 2008;37(6):1218.
[120] Barak DS, Dahatonde DJ, Batra S. Microwave-Assisted Metal-Free Decarboxylative Iodination/Bromination of Isoxazole-4-carboxylic Acids. Asian J Org Chem 2019;8(11):2149–54.
[121] Chegaev K, Lazzarato L, Tamboli Y, Boschi D, Blangetti M, Scozzafava A, et al. Furazan and furoxan sulfonamides are strong α-carbonic anhydrase inhibitors and potential antiglaucoma agents. Bioorg Med Chem 2014;22(15):3913–21.
[122] Conti P, Dallanoce C, De Amici M, De Micheli C, Ebert B. Synthesis and binding affinity of new muscarinic ligands structurally related to oxotremorine. Bioorg Med Chem Lett 1997;7(8):1033–6.
[123] Dighe SU, Mukhopadhyay S, Kolle S, Kanojiya S, Batra S. Synthesis of 3,4,5-Trisubstituted Isoxazoles from Morita-Baylis-Hillman Acetates by an NaNO2/I2-Mediated Domino Reaction. Angew Chem 2015;127(37):11076–80.
[124] Fernandes GFdS, de Souza PC, Marino LB, Chegaev K, Guglielmo S, Lazzarato L, et al. Synthesis and biological activity of furoxan derivatives against Mycobacterium tuberculosis. Eur J Med Chem 2016;123:523–31.
[125] Gehringer M, Laufer SA. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. J Med Chem 2018;62(12):5673–724.
[126] Kulikov AS, Larin AA, Fershtat LL, Anikina LV, Pukhov SA, Klochkov SG, et al. Synthesis, structural characterization and cytotoxic activity of heterocyclic compounds containing the furoxan ring. Arkivoc. 2017;2017(3):250–68.
[127] Nepali K, Lee H-Y, Liou J-P. Nitro-Group-Containing Drugs. J Med Chem 2018;62(6):2851–93.
[128] Olender D, Żwawiak J, Zaprutko L. Multidirectional Efficacy of Biologically Active Nitro Compounds Included in Medicines. Pharmaceuticals (Basel). 2018;11(2):54.
[129] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Electrophile in Covalent Enzyme Inhibition. ACS Chem Biol. 2018;13(6):1470–3.
[130] Ray S, Murkin AS. New Electrophiles and Strategies for Mechanism-Based and Targeted Covalent Inhibitor Design. Biochemistry 2019;58(52):5234–44.
[131] Shimadzu M, Ishikawa N, Yamamoto K, Tanaka A. Studies on furan derivatives.XIV. Nucleophilic substitution of methyl 5-nitro-2-furancarboxylate and 5-Nitrofuran-2-nitrile. J Heterocycl Chem 1986;23(4):1179–82.

[132] Shkineva TK, Vatsadze IA, Dalinger IL. A new general synthesis of functionally substituted pyrazolo[1,5-a]pyrimidines. Mendeleev Commun 2019;29 (4):429–31.

[133] Lei J, Zhou Y, Xie D, Zhang Y. Mechanistic insights into a classic wonder drug–aspirin. J Am Chem Soc. 2015;137(1):70–3.

[134] Rickert KW, Kelley P, Byrne NJ, Diehl RE, Hall DL, Montalvo AM, et al. Structure of human prostasin, a target for the regulation of hypertension. J Biol Chem. 2008;283(50):34864–72.

[135] Spraggon G, Hornsby M, Shipway A, Tully DC, Bursulaya B, Danahay H, et al. Active site conformational changes of prostasin provide a new mechanism of protease regulation by divalent cations. Protein Sci. 2009;18(5):1081–94.

[136] Sundermann TR, Benzin CV, Dražić T, Klein CD. Synthesis and structure-activity relationships of small-molecular di-basic esters, amides and carbamates as flaviviral protease inhibitors. Eur J Med Chem 2019;176:187–94.

[137] Somoza JR, Ho JD, Luong C, Ghate M, Sprengeler PA, Mortara K, et al. The Structure of the Extracellular Region of Human Hepsin Reveals a Serine Protease Domain and a Novel Scavenger Receptor Cysteine-Rich (SRCR) Domain. Structure. 2003;11(9):1123–31.

[138] Hempel T, Raich L, Olsson S, Azouz NP, Klingler AM, Hoffmann M, et al. Molecular mechanism of inhibiting the SARS-CoV-2 cell entry facilitator TMPRSS2 with camostat and nafamostat. Chem Sci 2021.

[139] Ramjee MK, Henderson IMJ, McLoughlin SB, Padova A. The Kinetic and Structural Characterization of the Reaction of Nafamostat with Bovine Pancreatic Trypsin. Thromb Res 2000;98(6):559–69.

[140] Rao KN, Anita RC, Sangeetha R, Anirudha L, Subramnay H. Crystal Structure of Serine protease Hepsin in complex with Inhibitor. Worldwide Protein Data Bank 2016.

[141] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421-..

[142] Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2011;9 (2):173–5.

[143] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–303.

[144] Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis 2009;30(S1):S162–73.

[145] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 2018;27(1):293–315.

[146] Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, et al. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins Struct Funct Bioinf 2003;50(3):437–50.

[147] Diller DJ, Merz KM. High throughput docking for library design and library prioritization. Proteins Struct Funct Genet 2001;43(2):113–24.

[148] Venkatachalam CM, Jiang X, Oldfield T, Waldman M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 2003;21(4):289–307.

[149] Wu G, Robertson DH, Brooks CL, Vieth M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER?A CHARMm-based MD docking algorithm. J Comput Chem 2003;24(13):1549–62.

[150] Wu G, Robertson DH, Brooks 3rd CL, Vieth M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. J Comput Chem. 2003;24(13):1549–62.

[151] Gehlhaar DK, Bouzida D, Rejto PA. Reduced Dimensionality in Ligand—Protein Structure Prediction: Covalent Inhibitors of Serine Proteases and Design of Site-Directed Combinatorial. Libraries. 1999.

[152] Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J Med Chem 1999;42 (5):791–804.

[153] Muegge I. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. In: Virtual Screening: An Alternative or Complement to High Throughput Screening? Springer; 2002. p. 99–114.

[154] Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20(4):281–95.

[155] Feher M. Consensus scoring for protein–ligand interactions. Drug Discovery Today 2006;11(9–10):421–8.

[156] Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. 2017;45(D1):D313–9.

[157] Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo-distance constraints applied on model quality estimation. Bioinformatics 2020;36(6):1765–71.

[158] Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. 2017;7(1):10480.

[159] Babajide Mustapha I, Saeed F. Bioactive Molecule Prediction Using Extreme Gradient Boosting. Molecules 2016;21(8):983.

[160] Prasath VB, Alfeilat HAA, Lasassmeh O, Hassanat A, Tarawneh AS (2017) Distance and similarity measures effect on the performance of k-nearest neighbor classifier—a review. arXiv preprint arXiv :1708.04321.

[161] Goh TC. Probabilistic neural network for evaluating seismic liquefaction potential. Can. Geotech. J. 2002;39:219–32.

[162] Wang S, Li X, Zhang S, Gui J, Huang D. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. Comput Biol Med 2010;40(2):179–89.

[163] Zhang H, Liu CT, Mao J, Shen C, Xie RL, Mu B. Development of novel in silico prediction model for drug-induced ototoxicity by using naive Bayes classifier approach. Toxicol In Vitro 2020;65:104812.

[164] Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today 2015;20(3):318–31.

[165] Wickramasinghe I, Kalutarage H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Comput 2020.

[166] Karthiga B, Rekha M. Feature extraction and I-NB classification of CT images for early lung cancer detection. Mater Today: Proc 2020.

[167] Riedmiller, M. and Braun, H., n.d. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. IEEE International Conference on Neural Networks,.

[168] Heikamp K, Bajorath J. Support vector machines for drug discovery. Expert Opin Drug Discov 2013;9(1):93–104.

[169] Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. J Med Chem 2005;48(7):2534–47.

[170] Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes? J Comput Aided Mol Des 2008;22(3–4):213–28.

[171] Wang X, Han W, Yan X, Zhang J, Yang M, Jiang P. Pharmacophore features for machine learning in pharmaceutical virtual screening. Mol Diversity 2020;24 (2):407–12.

[172] McHugh M. Interrater reliability: the kappa statistic. Biochemia Medica 2012:276–82.

[173] Kondeti PK, Ravi K, Mutheneni SR, Kadiri MR, Kumaraswamy S, Vadlamani R, et al. Applications of machine learning techniques to predict filariasis using socio-economic factors. Epidemiol Infect 2019;147:e260.

[174] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 2017;27 (5):1413–32.

[175] Zhu K, Borrelli KW, Greenwood JR, Day T, Abel R, Farid RS, Harder E. Docking covalent inhibitors: a parameter free approach to pose prediction and scoring. J Chem Inf Model. 2014, 28, 54(7):1932-40.

[176] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

[177] Chandar NB, Efremenko I, Silman I, Martin JML, Sussman JL. Molecular dynamics simulations of the interaction of Mouse and Torpedo acetylcholinesterase with covalent inhibitors explain their differential reactivity: Implications for drug design. Chem Biol Interact 2019;310:108715.

[178] Toviwek B, Gleeson D, Gleeson M. QM/MM and molecular dynamics investigation of the mechanism of covalent inhibition of TAK1 kinase. Org Biomol Chem 2021;19:1412–25.

[179] Fourches D, Ash J. 4D- quantitative structure–activity relationship modeling: making a comeback. Expert Opin Drug Discov 2019;14(12):1227–35.

[180] Lipiński PFJ, Szurmak P. SCRAMBLE 'N' GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models. Chem Pap 2017;71(11):2217–32.

[181] Kar S, Roy K, Leszczynski J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. Methods Mol Biol 2018;1800:141–69.

[182] Hevener K, Zhao W, Ball D, Babaoglu K, Qi J, White S, et al. Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. Journal Of Chemical Information And Modeling 2009;49:444–60.

[183] Taha M, AlDamen M. Effects of Variable Docking Conditions and Scoring Functions on Corresponding Protein-Aligned Comparative Molecular Field Analysis Models Constructed from Diverse Human Protein Tyrosine Phosphatase 1B Inhibitors. J Med Chem 2005;48:8016–34.

[184] Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev 2017;9:91–102.