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

A R T I C L E  I N F O
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

A B S T R A C T
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 (PNP) [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-
uous 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 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 insertions and deletions are remodelled using a fragment library. 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. 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].
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 \( < 4.0 \) indicating poor quality homology models. The later, on the other hand, combines several protein parameters including:

---

**Table 1**

| Warhead       | Nucleophile-Electrophile Chemistry | Example Structure Name | Target                                      |
|---------------|-----------------------------------|------------------------|---------------------------------------------|
| **β-Lactam**  |                                   |                        |                                             |
| **Carbamate** |                                   |                        |                                             |
| **Nitrile**   |                                   |                        |                                             |
| **Michael Acceptors** |                   |                        |                                             |
| **Nitro**     |                                   |                        |                                             |
| **Aromatic ester** |                           |                        |                                             |

---

Ma’mon M. Hatmal, O. Abuyaman and M. Taha Computational and Structural Biotechnology Journal 19 (2021) 4790–4824
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. 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 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 atom at distances ≤ 2.5 Å.

### 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 ≤ 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 atom at distances ≤ 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...
(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 Number | Name       | Structure | IC<sub>50</sub> or Ki (nM) | Warhead | Reference |
|-----------------|------------|-----------|-----------------------------|---------|-----------|
| 1               | Nafamostat | ![Structure 1](image1.png) | 100                          | Aromatic ester | [97]      |
| 2*              | Camostat   | ![Structure 2](image2.png) | 1000                         | Aromatic ester | [97]      |
| 3               | 5ce1-Hepsin| ![Structure 3](image3.png) | 100                          | None     | PDB code: 5CE1 |
| 4               |           | ![Structure 4](image4.png) | 8                            | None     | [90]      |
| 5*              |           | ![Structure 5](image5.png) | 19                           | None     | [90]      |
| 6               |           | ![Structure 6](image6.png) | 20                           | None     | [90]      |
| 7               |           | ![Structure 7](image7.png) | 19                           | None     | [90]      |

(continued on next page)
| Compound Number | Name     | Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|-----------------|----------|-----------|-----------------|---------|-----------|
| 8               | [54]ⁿ    |           | 45              | None    | [90]      |
| 9ⁿ              | [71]ⁿ    |           | 60              | None    | [90]      |
| 10ⁿ             | [31]ⁿ    |           | 160             | None    | [90]      |
| 11              | [69]ⁿ    |           | 220             | None    | [90]      |
| 12              | [92]ⁿ    |           | 0.9             | None    | [90]      |
| Compound Number | Name | Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|-----------------|------|-----------|----------------|---------|-----------|
| 13              | [93]a | ![Structure 13](image) | 4              | None    | [90]      |
| 14*             | [27]a | ![Structure 14](image) | 21             | None    | [90]      |
| 15              | [29]a | ![Structure 15](image) | 50             | None    | [90]      |
| 16              | Cefdinir | ![Structure 16](image) | Inactive**    | β-lactam | [97]      |
| 17              | Cefditoren | ![Structure 17](image) | Inactive**    | β-lactam | [97]      |

*(continued on next page)*
| Compound Number | Name            | IC₅₀ or Ki (nM) | Warhead   | Reference |
|-----------------|-----------------|-----------------|-----------|-----------|
| 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]      |
| Compound Number | Name         | IC₅₀ or Ki (nM) | Warhead         | Reference |
|-----------------|--------------|-----------------|-----------------|-----------|
| 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 | IC₅₀ or Ki (nM) | Warhead | Reference |
|--------------------|----------------|---------|-----------|
| **Inactive**       |                |         |           |
| Number 31* Ritonavir |                |         | [97]      |
| Number 32 Zafirlukast |                |         | [97]      |
| Number 33* Cobicistat |                |         | [97]      |
| Number 34 Doxazosin |                |         | [97]      |
| Number 35 Gefitinib |                |         | [97]      |
| Number 36 Pentamidine |                |         | [97]      |
| Number 37 Nafcillin |                |         | [97]      |
| Compound Name | Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|--------------|-----------|----------------|---------|-----------|
| Aztreonam    | ![Image](https://example.com/structure1.png) | Inactive** | β-lactam | [97]      |
| Carbenicillin| ![Image](https://example.com/structure2.png) | Inactive** | β-lactam | [97]      |
| Cefoperazone | ![Image](https://example.com/structure3.png) | Inactive** | β-lactam | [97]      |
| Dicloxacillin| ![Image](https://example.com/structure4.png) | Inactive** | β-lactam | [97]      |
| Ezetimibe    | ![Image](https://example.com/structure5.png) | Inactive** | β-lactam | [97]      |
| Meropenem    | ![Image](https://example.com/structure6.png) | Inactive** | β-lactam | [97]      |
| Entacapone   | ![Image](https://example.com/structure7.png) | Inactive** | Michael Acceptor | [97] |
| Compound Number | Name  | IC$_{50}$ or Ki (nM) | Warhead | Reference |
|-----------------|-------|----------------------|---------|-----------|
| 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] |
| Compound Number | Name        | Structure | IC₅₀ or Kᵢ (nM) | Warhead | Reference |
|-----------------|-------------|-----------|-----------------|---------|-----------|
| 53*             | Pinacidil   | ![Structure Image](image) | Inactive**     | Nitrile | [97]      |
| 54              | Febuxostat  | ![Structure Image](image) | Inactive**     | Nitrile | [97]      |
| 55              | Chloramphenicol | ![Structure Image](image) | Inactive**     | Nitro   | [97]      |
| 56*             | Flutamide   | ![Structure Image](image) | Inactive**     | Nitro   | [97]      |
| 57*             | Nifedipine  | ![Structure Image](image) | Inactive**     | Nitro   | [97]      |
| 58              | Nimodipine  | ![Structure Image](image) | Inactive**     | Nitro   | [97]      |
| 59              | Nisoldipine | ![Structure Image](image) | Inactive**     | Nitro   | [97]      |
| 60              | Nitazoxanide| ![Structure Image](image) | Inactive**     | Nitro   | [97]      |

(continued on next page)
| Compound Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|--------------------|----------------|---------|-----------|
| Nitrendipine       | Inactive⁺⁺     | Nitro   | [97]      |
| Nitrofurazone      | Inactive⁺⁺     | Nitro   | [97]      |
| Tinidazole         | Inactive⁺⁺     | Nitro   | [97]      |
| Tolcapone          | Inactive⁺⁺     | Nitro   | [97]      |
| Atazanavir         | Inactive⁺⁺     | Carbamate | [97] |
| Felbamate          | Inactive⁺⁺     | Carbamate | [97] |
| Fenspiride         | Inactive⁺⁺     | Carbamate | [97] |
| Linezolid          | Inactive⁺⁺     | Carbamate | [97] |
| Compound Number | Name          | IC₅₀ or Ki (nM) | Warhead        | Reference |
|-----------------|---------------|-----------------|----------------|-----------|
| 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)
| Compound Number | Name    | Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|-----------------|---------|-----------|----------------|---------|-----------|
| 78              | Loxapine| ![Structure](image1) | Inactive**     | None    | [97]      |
| 79              | Methazolamide | ![Structure](image2) | Inactive**     | None    | [97]      |
| 80              | Phenformin | ![Structure](image3) | Inactive**     | None    | [97]      |
| 81              | Pyrimethamine | ![Structure](image4) | Inactive**     | None    | [97]      |
| 82              | Quetiapine | ![Structure](image5) | Inactive**     | None    | [97]      |
| 83              | Sildenafil | ![Structure](image6) | Inactive**     | None    | [97]      |
| 84*             | Terazosin | ![Structure](image7) | Inactive**     | None    | [97]      |
| 85*             | Trimethoprim | ![Structure](image8) | Inactive**     | None    | [97]      |
| 86              | Alfuzosin | ![Structure](image9) | Inactive**     | None    | [97]      |
| Compound | Structure | IC<sub>50</sub> or Ki (nM) | Warhead | Reference |
|----------|-----------|--------------------------|---------|-----------|
| 87 Argatroban | ![Structure](image) | Inactive** | None | [97] |
| 88 Famotidine | ![Structure](image) | Inactive** | None | [97] |
| 89 Ranitidine | ![Structure](image) | Inactive** | None | [97] |
| 90 Dasatinib | ![Structure](image) | Inactive** | None | [97] |
| 91 Didanosine | ![Structure](image) | Inactive** | None | [97] |
| 92 Mercaptopurine | ![Structure](image) | Inactive** | None | [97] |
| 93* Diminazene | ![Structure](image) | Inactive** | None | [97] |
| 94 Erlotinib | ![Structure](image) | Inactive** | None | [97] |
| 95 Gabexate | ![Structure](image) | Inactive** | Aromaticester | [97] |
| 96 1p57-Hepsin | ![Structure](image) | 40,000 | None | [138] |

(continued on next page)
| Compound Number | Compound Name | Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|-----------------|---------------|-----------|----------------|---------|-----------|
| 97              | [53]a         | ![Compound 97 Structure](image) | 10,000         | None    | [90]      |
| 98              | [78]a         | ![Compound 98 Structure](image) | 20,000         | None    | [90]      |
| 99              | Sivelestat    | ![Compound 99 Structure](image) | Inactive**     | Aromatic ester | [97] |
| 100             | Telaprevir    | ![Compound 100 Structure](image) | Inactive**     | α-keto-amide | [97] |
| 101             | Azlocillin    | ![Compound 101 Structure](image) | Inactive**     | β-lactam | [97] |
| 102             | Doripenem     | ![Compound 102 Structure](image) | Inactive**     | β-lactam | [97] |
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).

Table 2 (continued)

| Compound Number | Structure | IC₅₀ or Ki (nM) | Warhead | Reference |
|-----------------|-----------|----------------|---------|-----------|
| 103* Famciclovir | ![Structure](image1) | Inactive** | None | [97] |
| 104 Mupirocin | ![Structure](image2) | Inactive** | Michael Acceptor | [97] |
| 105 Nizatidine | ![Structure](image3) | Inactive** | Nitro | [97] |
| 106 Darunavir | ![Structure](image4) | Inactive** | Carbamate | [97] |
| 107 Zanamivir | ![Structure](image5) | 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.
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).

\[
\text{Accuracy} = \frac{TP + TN}{N}
\]

(1)

\[
K = \frac{P_o + P_e}{1 - P_e}
\]

(2)

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

\[
K = \frac{P_o + P_e}{1 - P_e}
\]

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 TMRPRSS2-related ligands reported in the literature (active and inactive) [90,97]. Only untested compounds (155 molecules) were kept for subsequent virtual screening as potential TMRPRSS2 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 LRCCs 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.paramedchem.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 simulations were visualized using Discovery Studio (version 4.5, Biovia, USA).

2.7. Covalent docking
Capreomycin and aspoxicillin were covalently docked into TMRPRSS2 homology model using CovDock software [175] through Maestro Molecular Modeling Interface (Schrodinger 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 VSBG2.0 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 kcal/mol with a maximum number of poses = 200, the reactive residue is S478, docked ligand(s) is(are) confined to an enclosing box of \( < 20 \) 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., \( \beta \)-lactam addition and Michael acceptor). A single top-ranking docked pose was reported as output.

3. Results and discussion

3.1. Homology modeling of TMRPRSS2

Since TMRPRSS2 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 TMRPRSS2 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 data bank (https://www.rcsb.org/). However, we only opted for template proteins that have at least 65% sequence coverage with TMRPRSS2 and have their co-crystallized bound ligands successfully transferred by SWISS-MODEL to the proposed binding pocket of TMRPRSS2. 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 TMRPRSS2 in the proposed binding site is 77% with the catalytic serine (S478) being conserved. Fig. 3A compares TMRPRSS2 sequence to the template protein Hepsin.
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 released 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 an-carbons (Ca) and main chain carbons, respectively. Fig. 4 B 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 TMPRSS2 inhibitors. The search identified 1064 inhibitors that can be unequivocally classified as “active” (IC\textsubscript{50} \leq 1,000 nM) or “inactive” (IC\textsubscript{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.
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 TMPRSS2 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 ≤ 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 demarcate 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 learners, 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 ≥ 8 on training and testing poses (i.e., 8 or 9 of the scoring functions ranked the selected poses

### Table 3

| Learner | Descriptors | Accuracy | Cohen’s Kappa |
|---------|-------------|----------|---------------|
|         |             | L20%Out* | Testingb | Re-substitutionc | L20%out* | Testingc | Re-substitutionc |
| Xgboost | LRCFs and Scoring Functions | 0.96 | 0.90 | 1.00 | 0.95 | 0.73 | 1.00 |
| LG | LRCFs | 0.91 | 0.83 | 0.97 | 0.63 | 0.53 | 0.89 |
| C-SVM | Scoring | 0.95 | 0.88 | 1.00 | 0.80 | 0.68 | 1.00 |
|     | LRCFs and Scoring Functions | 0.94 | 0.89 | 0.95 | 0.73 | 0.71 | 0.80 |
|     | LRCFs | 0.86 | 0.87 | 0.87 | 0.22 | 0.21 | 0.32 |
|     | Scoring Functions | 0.84 | 0.72 | 1.00 | 0.00 | 0.00 | 1.00 |
| nu-SVM | 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 | 0.95c | 0.90c | 0.96 | 0.79c | 0.72 | 0.98 |
|     | LRCFs | 0.90c | 0.81 | 0.96 | 0.57c | 0.46 | 0.83 |
|     | Scoring Functions | 0.95c | 0.89 | 1.00 | 0.78c | 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 | 0.63 | 0.66 | 0.68 |
|     | Scoring Functions | 0.92 | 0.88 | 0.93 | 0.00 | 0.00 | 0.00 |
| Naive Bayesian | 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 | 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 | 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 |

* L20%Out: Leave 20% out cross-validation for accuracy and Cohen’s Kappa.
* Testing: Accuracy and Cohen’s Kappa determined by applying the particular ML model to predict the same training compounds used to build the model.
* 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.
* Performed using optimized nu value of 0.1.
* Performed using optimized k value of 6 (number of neighbors).
* 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), 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 camostatin 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

| Compoundsa | Predicted number of active and inactive docked poses for testing set compounds. |
|------------|--------------------------------------------------------------------------------|
| Numbers    | Anti-TMPRSS2 Activity | Predicted number of active and inactive docked poses | Xgboost | Percent Active Poses | Nu-SVM | RF | Percent Active Poses | Active | Inactive | Active | Inactive | Percent Active Poses | Active | Inactive | Percent Active Poses |
|------------|-----------------------|---------------------------------------------------|---------|---------------------|--------|----|---------------------|--------|----------|--------|----------|---------------------|--------|----------|---------------------|
| 5          | Active                | 9704                                              | 3904    | 94.2               | 94.0   | 92.8 | 1008                | 2.0    | 98.1     |
| 9          | Active                | 9704                                              | 3904    | 94.2               | 94.0   | 92.8 | 1008                | 2.0    | 98.1     |
| 10         | Active                | 9704                                              | 3904    | 94.2               | 94.0   | 92.8 | 1008                | 2.0    | 98.1     |
| 14         | Active                | 9704                                              | 3904    | 94.2               | 94.0   | 92.8 | 1008                | 2.0    | 98.1     |
| 3          | Active                | 9704                                              | 3904    | 94.2               | 94.0   | 92.8 | 1008                | 2.0    | 98.1     |
| 101        | Active                | 9704                                              | 3904    | 94.2               | 94.0   | 92.8 | 1008                | 2.0    | 98.1     |
| 21         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 24         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 26         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 31         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 33         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 39         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 53         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 56         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 57         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 62         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 65         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 69         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 70         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 84         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 85         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 93         | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |
| 103        | Inactive              | 4205                                              | 4205    | 95.0               | 95.0   | 95.0 | 5000                | 5.0    | 95.0     |

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).

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 | Predicted number of active and inactive docked poses |
|-----------|-----------------------------------------------------|
| Xgboost   | Active Poses | Inactive Poses | %Active Poses |
| nu-SVM    | Active Poses | Inactive Poses | %Active Poses |
| RF        | Active Poses | Inactive Poses | %Active Poses |
| 108 5-Iodo-Sunitinib | 1 | 218 | 0.5 | 0 | 219 | 0 |
| 109 5-CH3O-Sunitinib | 2 | 253 | 0.8 | 2 | 253 | 0.8 |
| 110 Aenocoumarol | 1 | 10 | 9.1 | 0 | 10 | 0 |
| 111 Acrivastine | 0 | 24 | 0 | 0 | 24 | 0 |
| 112 Alectinib | 0 | 38 | 0 | 0 | 38 | 0 |
| 113 Alogliptin | 0 | 32 | 0 | 0 | 32 | 0 |
| 114 Apallicin | 4 | 284 | 1.4 | 5 | 283 | 1.7 |
| 115 Apraclonidine | 0 | 1 | 0 | 0 | 1 | 0 |
| 116 Aspoxicillin | 89 | 121 | 42.4 | 75 | 135 | 35.7 |
| 117 Ast-1306 | 1 | 523 | 0.2 | 12 | 512 | 2.3 |
| 118 Bacampicillin | 34 | 157 | 17.8 | 25 | 166 | 13.1 |
| 119 Bambuterol | 0 | 42 | 0 | 1 | 41 | 2.4 |
| 120 Baricitinib | 0 | 25 | 0 | 0 | 25 | 0 |
| 121 Belinostat | 0 | 78 | 0 | 0 | 78 | 0 |
| 122 Benznidazole | 0 | 12 | 0 | 0 | 12 | 0 |
| 123 Biapenem | 1 | 29 | 3.3 | 2 | 28 | 6.7 |
| 124 Brodimoprim | 0 | 1 | 0 | 0 | 1 | 0 |
| 125 Canertinib | 10 | 1284 | 0.8 | 19 | 1275 | 1.5 |
| 126 Capreomycin1 | 1183 | 931 | 56.9 | 1203 | 911 | 56.9 |
| 127 Carisoprodol | 0 | 8 | 0 | 0 | 8 | 0 |
| 128 Carumonam | 0 | 11 | 0 | 0 | 11 | 0 |
| 129 Cefadroxil | 0 | 87 | 0 | 11 | 76 | 12.6 |
| 130 Cefamandole | 0 | 1 | 0 | 0 | 1 | 0 |
| 131 Cefatrizine | 2 | 575 | 0.3 | 17 | 560 | 2.9 |
| 132 Cefazezone | 0 | 57 | 0 | 0 | 57 | 0 |
| 133 Cefazolin | 0 | 51 | 0 | 0 | 51 | 0 |
| 134 Cefbuperazone | 0 | 66 | 0 | 0 | 66 | 0 |
| 135 Cefcapene | 0 | 115 | 0 | 0 | 115 | 0 |
| 136 Cefclidin | 0 | 181 | 0 | 1 | 180 | 0.6 |
| 137 Cefpirim | 2 | 105 | 1.9 | 0 | 107 | 0 |
| 138 Cefnetamet | 0 | 52 | 0 | 0 | 52 | 0 |
| 139 Cefixime | 0 | 4 | 0 | 0 | 4 | 0 |
| 140 Cefmenoxime | 0 | 176 | 0 | 0 | 176 | 0 |
| 141 Cefmetazol | 0 | 25 | 0 | 0 | 25 | 0 |
| 142 Cefmindex | 0 | 80 | 0 | 4 | 76 | 5 |
| 143 Cefodizime | 0 | 43 | 0 | 0 | 43 | 0 |
| 144 Cefonicid | 0 | 94 | 0 | 0 | 94 | 0 |
| 145 Ceforanide | 1 | 166 | 0.6 | 2 | 165 | 1.2 |
| 146 Cefotaxime | 0 | 112 | 0 | 0 | 112 | 0 |
| 147 Cefotetan | 0 | 79 | 0 | 0 | 79 | 0 |
| 148 Cefotiam | 0 | 183 | 0 | 1 | 182 | 0.5 |
| 149 Cefoxitin | 2 | 33 | 5.7 | 1 | 34 | 2.9 |
| 150 Cefozogran | 0 | 270 | 0 | 1 | 269 | 0.4 |
| 151 Cefpinolce | 0 | 179 | 0 | 1 | 178 | 0.6 |
| 152 Cefpramid | 3 | 309 | 1 | 0 | 312 | 0 |
| 153 Cefpirime | 1 | 157 | 0.6 | 0 | 158 | 0 |
| 154 Cefpodoxime | 0 | 74 | 0 | 0 | 74 | 0 |
| 155 Cefprozil | 0 | 3 | 0 | 0 | 3 | 0 |
| 156 Cefroxidine | 0 | 58 | 0 | 0 | 58 | 0 |
| 157 Ceferam | 0 | 16 | 0 | 0 | 16 | 0 |
| 158 Ceferzole | 0 | 38 | 0 | 0 | 38 | 0 |
| 159 Celfibutin | 0 | 61 | 0 | 0 | 61 | 0 |
| 160 Cefizoxime | 1 | 42 | 2.3 | 2 | 41 | 4.7 |
| 161 Cefixione | 0 | 186 | 0 | 0 | 186 | 0 |
| 162 Cefuroxime | 0 | 59 | 0 | 1 | 58 | 1.7 |
| 163 Cefuzonam | 0 | 168 | 0 | 0 | 168 | 0 |
| 164 Cephacetrile | 0 | 2 | 0 | 0 | 2 | 0 |
| 165 Cephaloridine | 0 | 23 | 0 | 1 | 22 | 4.3 |
| 166 Cephalothin | 0 | 19 | 0 | 0 | 19 | 0 |
| 167 Cephapirin | 0 | 38 | 0 | 2 | 36 | 5.3 |
| 168 Cephradine | 1 | 44 | 2.2 | 0 | 45 | 0 |
| 169 Cilastatin | 0 | 113 | 0 | 3 | 110 | 2.7 |
| 170 Cinsanelin | 0 | 188 | 0 | 0 | 188 | 0 |
| 171 Citalopram | 0 | 9 | 0 | 0 | 9 | 0 |
| 172 Clohenprobit | 0 | 138 | 0 | 0 | 138 | 0 |
| 173 Clonazepam | 0 | 5 | 0 | 0 | 5 | 0 |
| 174 Clonizane | 0 | 41 | 0 | 0 | 41 | 0 |
| 175 Cyclacillin | 0 | 23 | 0 | 2 | 21 | 8.7 |
| 176 Cypermethrin | 0 | 57 | 0 | 0 | 57 | 0 |

(continued on next page)
### Table 5 (continued)

| Compounds | Predicted number of active and inactive docked poses | Xgboost | nu-SVM | RF |
|-----------|------------------------------------------------------|---------|--------|----|
|           |                                                      | Active  | Inactive | %Active Poses |
|           |                                                      | Active  | Inactive | %Active Poses |
|           |                                                      | Active  | Inactive | %Active Poses |
| 177       | Dacomitinib                                         | 0       | 430      | 0             |
| 178       | Dantrolene                                          | 0       | 91       | 0             |
| 179       | Debrisoquin                                         | 0       | 1        | 0             |
| 180       | Demecarium                                          | 1       | 12       | 7.7           |
| 181       | Difenoxin                                            | 0       | 26       | 0             |
| 182       | Diphenoxylate                                        | 1       | 83       | 1.2           |
| 183       | Enzalutamide                                         | 0       | 16       | 0             |
| 184       | Etraperine                                           | 0       | 21       | 0             |
| 185       | Eszopiclone                                          | 0       | 41       | 0             |
| 186       | Ethacryn_Acid                                        | 0       | 1        | 0             |
| 187       | Etonitazene                                          | 0       | 55       | 0             |
| 188       | Etozolin                                             | 0       | 5        | 0             |
| 189       | Etreterine                                           | 0       | 7        | 0             |
| 190       | Ezogabine                                            | 0       | 144      | 0             |
| 191       | Famitnib                                             | 1       | 323      | 0.3           |
| 192       | Flubanilate                                          | 0       | 11       | 0             |
| 193       | Fluocoxacillin                                       | 0       | 10       | 0             |
| 194       | Flumizadole                                          | 0       | 9        | 0             |
| 195       | Flutrimazepam                                        | 0       | 2        | 0             |
| 196       | Fosamiprenavir                                       | 22      | 43       | 33.8          |
| 197       | Guanadrel                                            | 0       | 8        | 0             |
| 198       | Guanethidin                                          | 1       | 4        | 20            |
| 199       | Guanfacine                                           | 0       | 10       | 0             |
| 200       | Guanisquin                                            | 0       | 3        | 0             |
| 201       | Guanoxan                                             | 0       | 12       | 0             |
| 202       | Guanoxifen                                           | 0       | 11       | 0             |
| 203       | Henatinb                                             | 0       | 1        | 0             |
| 204       | Hetacillin                                            | 0       | 21       | 1             |
| 205       | Hydroxybutylaminic                                    | 0       | 2        | 0             |
| 206       | Ibrutinib                                             | 1       | 387      | 0.3           |
| 207       | Imipenem                                             | 0       | 16       | 0             |
| 208       | Jodenguan                                             | 0       | 5        | 0             |
| 209       | Irideotecan                                          | 1       | 24       | 4             |
| 210       | Iosulfazecin                                         | 1       | 93       | 1.1           |
| 211       | Leuprolide                                            | 0       | 2        | 0             |
| 212       | Levobemastine                                         | 0       | 8        | 0             |
| 213       | Levofuraltadone                                       | 0       | 29       | 0             |
| 214       | Levopropylcillin                                      | 0       | 2        | 0             |
| 215       | Lodoxamide                                           | 0       | 27       | 0             |
| 216       | Loracarbef                                            | 0       | 55       | 0             |
| 217       | Methenazolone                                         | 0       | 82       | 0             |
| 218       | Methylpiperidone                                      | 0       | 7        | 0             |
| 219       | Mefoxanin                                             | 0       | 8        | 0             |
| 220       | Mometopotin                                          | 0       | 125      | 0             |
| 221       | Moricizine                                            | 0       | 219      | 0             |
| 222       | Naquotinib                                            | 4       | 330      | 1.2           |
| 223       | Neratinb                                             | 7       | 324      | 2.1           |
| 224       | Nitosamide                                            | 0       | 16       | 0             |
| 225       | Nifurtimox                                            | 0       | 3        | 0             |
| 226       | Nilutamide                                            | 0       | 2        | 0             |
| 227       | Nisutinone                                            | 0       | 6        | 0             |
| 228       | Nitrozepam                                            | 0       | 6        | 0             |
| 229       | Nitrofurantoin                                        | 0       | 35       | 0             |
| 230       | Octrocyline                                           | 0       | 12       | 0             |
| 231       | OctyMethoxynicnic                                    | 0       | 16       | 0             |
| 232       | Olmutinib                                             | 8       | 566      | 1.4           |
| 233       | Ortalinb                                             | 1       | 31       | 3.1           |
| 234       | Osimertinib                                           | 34      | 1031     | 3.2           |
| 235       | Oxaclidin                                             | 0       | 16       | 0             |
| 236       | Oxamnique                                             | 0       | 26       | 0             |
| 237       | Peptolinib                                            | 1       | 229      | 0.4           |
| 238       | Penicillin G                                          | 0       | 12       | 0             |
| 239       | Penicillin V                                          | 0       | 20       | 0             |
| 240       | Pentagastroin                                         | 5       | 566      | 0.9           |
| 241       | Pericyazine                                           | 0       | 41       | 0             |
| 242       | Piritrexin                                            | 31      | 547      | 5.4           |
| 243       | Poziotinib                                            | 0       | 167      | 0             |
| 244       | Propanil                                             | 0       | 11       | 0             |
| 245       | Pyrotinin                                             | 4       | 143      | 2.7           |
| 246       | Rescinnamine                                          | 1       | 7        | 12.5          |
| 247       | Rocletinib                                            | 7       | 910      | 0.8           |
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. 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 site. These results nicely correlate with the predictions of our ML models despite being generated for rather diverse set of compounds.

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 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 site. 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 threshold line (magenta 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 warhead to reside for some time, e.g., 20 to 80 ns [177,178], at close proximity to the nucleophilic residue.
**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.)
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 Biointf 2009;10(5):579–91.
[2] Pantsar T, Poso A. Binding Affinity vs 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] Pirzi L, Rastelli G. Molecular Docking: Shifting Paradigms in Drug Discovery. Int J Mol Sci. 2019;20(18):4331.
[5] Ashitawy HM, Mahapatra NR. Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment. J Chem Inf Model 2017;57(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, Regnman 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 TJ, Makino S, Stillman AG, Kuntz ID. J Comput Aided Mol Des 2001;15(5):411–28.
International Journal of Pure and Applied Mathematics. 2018;118(15):389–405.

[72] Heikamp K, Bajrath J. Support vector machines for drug discovery. Expert Opin Drug Discov. 2013;8(1):93–104.

[73] Jayaraj PB, Jain S. Ligand based virtual screening using SVM on GPU. Comput Biol Chem. 2019;83:107143.

[74] Hagen M, Buhl R. A probabilistic neural network approach for modeling and classification of bacterial growth/no-growth data. J Microbiol Methods. 2002;52(1):217–26.

[75] Khosravani A, Ayat S. Diagnosing Breast Cancer Type by Using Probabilistic Neural Network Decision Support System. International Journal of Knowledge Engineering. 2016;21(1):73–6.

[76] Tran DH, Ng AWM, Pereira BJ, Burn S, Davis P. Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. Urban Water J. 2007;4(2):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 coal stover boilers. Computer Aided Chemical Engineering: Elsevier; 2011. p. 1784–8.

[82] Tsamardinos I, Ge菖sidou E, Borboudakis G. Boosting the out-of-sample predictions for efficient and accurate cross-validation. Machine learning. 2010;79(1):125–52.

[83] Sahiner B, Chan H-P, Hadjiiskii 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 C, Drummond GR, Vowler SL. Making with what we have: use your bootstraps. The Journal of physiology. 2012;590(15):3403–6.

[85] Hilgenfeld R, Peris 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, Krüger N, Herrler T, Erischen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;182(1):271–80.e6.

[87] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. A Novel Coronavirus (SARS-CoV-2) 2019: A Systematic Drug Repurposing Approach to Identify Promising Inhibitors Against 3C-like Protease and 2'-O-Ribose Methyltransferase. American Chemical Society (ACS) 2020.

[88] Laporte M, Naesens L. Airway proteases: an emerging drug target for cystic fibrosis. Trends in Pharmacological Sciences. 2008;29(12):639–45.

[89] Vacher J, Loh R, Powell T, et al. A plant pathogen virulence factor inhibits the eukaryotic proteasome by a novel mechanism. Nature 2008;452(7188):755–8.

[90] Niseng CF, Braeke S. The oxa-Michael reaction: from recent developments to applications. Chem Soc Rev. 2020;49(9):2617–87.

[91] Gehringer M, Laufer SA. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors. Angew Chem. 2020;132(34):14183–95.

[92] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Peptidyl-peptide Fragment: Novel Strategy for Inhibiting the Peptidyl-peptidease Activity of the Main Protease of SARS-CoV-2. Bioactive Carbon Compounds. 2020;12:6473–724.

[93] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Peptidyl-peptide Fragment: Novel Strategy for Inhibiting the Peptidyl-peptidease Activity of the Main Protease of SARS-CoV-2. Bioactive Carbon Compounds. 2020;12:6473–724.

[94] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Peptidyl-peptide Fragment: Novel Strategy for Inhibiting the Peptidyl-peptidease Activity of the Main Protease of SARS-CoV-2. Bioactive Carbon Compounds. 2020;12:6473–724.

[95] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Peptidyl-peptide Fragment: Novel Strategy for Inhibiting the Peptidyl-peptidease Activity of the Main Protease of SARS-CoV-2. Bioactive Carbon Compounds. 2020;12:6473–724.

[96] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Peptidyl-peptide Fragment: Novel Strategy for Inhibiting the Peptidyl-peptidease Activity of the Main Protease of SARS-CoV-2. Bioactive Carbon Compounds. 2020;12:6473–724.

[97] Ray S, Kreitler DF, Gulick AM, Murkin AS. The Nitro Group as a Masked Peptidyl-peptide Fragment: Novel Strategy for Inhibiting the Peptidyl-peptidease Activity of the Main Protease of SARS-CoV-2. Bioactive Carbon Compounds. 2020;12:6473–724.
