Medicine in Drug Discovery 21 (2024) 100176

Contents lists available at ScienceDirect

Medicine in Drug Discovery

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

Revealing key structural features for developing new agonists targeting δ opioid receptor: Combined machine learning and molecular modeling perspective

Zeynab Fakhar [a], Ali Hosseinpouran [b], Orde Q. Munro [c], Sorena Sarmadi [d], Sajjad Gharaghani [a,*]

[a] Laboratory of Bioinformatics and Drug Design (LBD), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[b] Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran
[c] Molecular Sciences Institute, School of Chemistry, University of the Witwatersrand, PO, WITS 2050, Johannesburg, South Africa
[d] Department of Mathematics, University of Houston, Houston, TX 77204, USA

ARTICLE INFO

Keywords:
δ Opioid Receptor
Machine learning
QSAR
Ligand–Receptor Contacts
Non-Linear Regression
XGBOOST

ABSTRACT

Despite being the most widely prescribed and misused type of medication, opioids continue to function as robust pain relief agents; however, overdosing is a significant cause of fatalities among opioid users. The δ-opioid receptor (DOR) has immense promise in treating long-term pain by producing anxiolytic and antidepressant-like outcomes. Although DOR agonists play a crucial role, their clinical implementation is restricted because of the probable manifestation of severe, life-threatening complications. A Python-based machine learning approach was employed to develop a quantitative structure–activity relationship (QSAR) model in this study. To address this, 4217 compounds and their associated biological inhibition activities were retrieved from the gpcrdb database. The K-best features selection method revealed three key structural features such as SLOGPVSA2, Chi6ch, and S17 contributed significantly to the best model performance. Statistical analysis, K-fold cross-validation, applicability domain analysis, and external validation using 38 unseen FDA-approved drug data confirmed the robustness of the predictive model. A molecular docking study in along with Ligand–Receptor Contact Fingerprints (LRCFs) using the essential chemical interactions described for analog ligands releaved the key contact interactions of Asp 128, Tyr 129, Met 132, Trp 274, Ile 277, and Tyr 308 residues in the total binding affinities upon complexation. Our combinatorial study using regression QSAR and ligand–receptor Contact, analysis could serve in the design of more rational compounds for drug discovery targeting DOR.

1. Introduction

Opioids, which are highly potent pain-relieving agents, continue to be extensively prescribed and misused medications, with overdose being a primary cause of death among individuals who utilize them. In 2019, around 275 million individuals worldwide engaged in drug use at least once. Among this population, approximately 62 million individuals specifically used opioids, and out of those, about 36.3 million individuals experienced drug use disorders [1]. Despite the existence of effective treatment interventions for opioid dependence that can reduce the chances of overdose, fewer than 10 % of individuals requiring such treatment are currently accessing it [2]. As a result, healthcare has placed significant emphasis on prioritizing the development of new opioid painkillers, antitussives, antidepressants, and antipruritic therapies that carry a reduced risk of abuse and overdose. Despite its significant adverse effects, morphine, which serves as the primary component of opium, remains the most potent opiate and the most commonly utilized painkiller in modern medicine [3,4]. Heroin, an illicit substance derived from the Morphine family, carries substantial societal impact as a highly abused drug. Opioids exert their effects by binding to opioid receptors (ORs), specifically the Mu (μ) [5], Kappa (κ) [5,6], and Delta receptors (δ) [7,8] which belong to a crucial subfamily of G protein-coupled receptors (GPCRs). These receptors serve as significant protein targets for the treatment of both acute and chronic pain [9]. The Delta opioid receptors are composed of a solitary polypeptide chain that includes an N-terminal region on the extracellular side where glycosylation occurs (Fig. 1). The receptors also possess seven transmembrane alpha helices domains (TM) where ligands bind, namely

[*] Corresponding author.
E-mail address: s.gharaghani@ut.ac.ir (S. Gharaghani).

https://doi.org/10.1016/j.medidd.2024.100176
Received 12 October 2023; Received in revised form 2 January 2024; Accepted 12 January 2024
Available online 18 January 2024

TM1: Leu 48-Val 75; TM2: Ile 86-Leu 110; TM3: Ala 123-Val 144; TM4: Ile 168-Met 186; TM5: Trp 207-Leu 238; TM6: Met 262-Trp 284; TM7: Leu 300-Leu 321, and an intracellular C-terminal tail responsible for phosphorylation.

Claff *et al.* reported two agonist-bound crystal structures of thermostabilized DOP in an activated state and in complex with the peptide KGCHM07 (Ki = 5.17 ± 1.57 nM) and the small-molecule DPI-287 (Ki = 0.39 ± 0.12 nM) at 2.8 Å and 3.3 Å resolutions [10]. The transmembrane binding site of DOR is hydrophobic and consists of 10 key residues of Asp 128, Tyr 129, Met 132, Trp 274, Ile 277, His 278, Val 281, Trp 284, Leu 300, and Tyr 308 (Fig. 2) [10–12]. Among these binding site residues, Asp 128 plays a critical role in DOR activation and is considered to be the active site of the protein [13].

Nevertheless, researchers have extensively explored alternative opioid receptors as promising targets for the development of safer treatments for chronic pain. Among these receptors, the δ opioid receptor (DOR) has demonstrated significant potential for addressing chronic pain [14,15] displaying anxiolytic and antidepressant-like effects [16,17]. Apart from their crucial role in pain management, the use of DOR agonists has been explored for their potential in treating depression and alleviating spasms related to Parkinson's disease. However, the clinical application of DOR agonists is restricted due to the potential occurrence of severe and life-threatening side effects such as tolerance, convulsions, and seizures [18]. Despite numerous efforts, molecules that selectively target the delta opioid receptor have not been successfully translated into clinical applications [19]. This advantageous psychopharmacological profile with the undesired adverse effects puts DOR agonists at the forefront of the development of novel compounds and candidate drugs.

Computational intelligence techniques are utilized throughout the entire drug development process to enhance efficiency and automate research analysis. This approach aids in assessing risks, estimating costs, and expediting clinical trials in the field of drug discovery. Although the Delta opioid receptor (DOR) plays a crucial role in various diseases, such as migraine, alcohol use disorder, ischemia, and neurodegenerative diseases, there seems to be a scarcity of reported computational intelligence endeavors focused on DOR inhibitors within the realm of drug discovery research. Over the past few years, machine learning (ML) techniques have emerged as a prominent and powerful toolset in the pharmaceutical industry, enabling the extraction of valuable insights from vast datasets. Machine learning is widely applied in drug discovery to establish the correlation between the chemical properties of molecules and their biological activity. Recently, Podlewska *et al.*, [20] published a comparative case study of ORs on the application of ligand-based classification ML models and structure-based docking methods. In their study, the authors failed to interpret the importance of interaction fingerprint-derived docking analysis of ligands-ORs complexes. In their study three crystal structures of the three ORs were considered in which two crystals of δ and k are mutated in the binding site, but the μ receptor is wild type. This might be the main reason for the negative impact of ligand–protein interaction fingerprints in this study. In another study by Sakamuru *et al.*, [21] a virtual screening-based classification ML was performed on their in-house active compounds' library targeting ORs, and they developed classifier models to predict the OPR activity of small molecules. In the current project, we constructed a Python-based ML approach encompassing a regression quantitative structure–activity relationship [22,23] (QSAR) study using two widely applied machine learning algorithms in a Python environment, namely, XGBOOST and RF, to predict the inhibitory constant activities of 4,217 compounds against delta opioid receptors. This analysis identified the critical structural features that contribute to the agonist activity of the ligands and devised a novel template that closely resembles the highly potent bioactive compounds. A molecular docking study along with ligand–receptor Contact Fingerprints (LRCFs) [24] was conducted to elucidate the key contact interactions in the total binding affinities of ligands-DOR upon complexation. In this study, the regressor QSAR model along with ligand–receptor Contact analysis could serve as a robust predictive model for designing new inhibitors.

2. Methodologies

2.1. Data set collection

Out of total 7,154 compounds targeting DOR, a collection of 4,217 agonist ligands, filtered and downloaded from the G protein-coupled receptor database, was compile [11,25–27] (https://gpcrdb.org/). These agonist ligands were measured for their inhibition constant values using a standardized bioassay procedure, and the data was recorded in Ki format. An exploratory data analysis (EDA) approach was applied to refine SMILES notations, duplicate molecules, salt forms, heavy metals and fragments. Finally, the compounds were filtered on the basis of Lipinski's rule of five (RO5) [28].

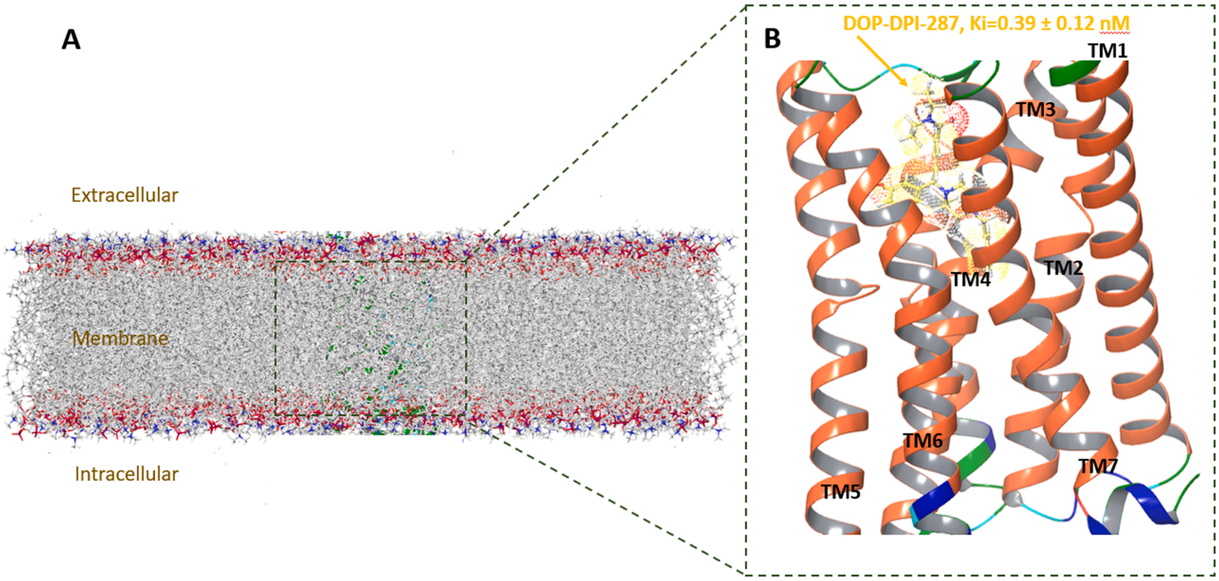

Fig. 1. A: 3D structural overview of δ opioid receptor (Chain-A) in complex with the DPI-287 agonist (PDB code: 6PT3); B: seven transmembrane (TM) alpha helices domains where DPI-287 ligand bind (yellow meshed surface).

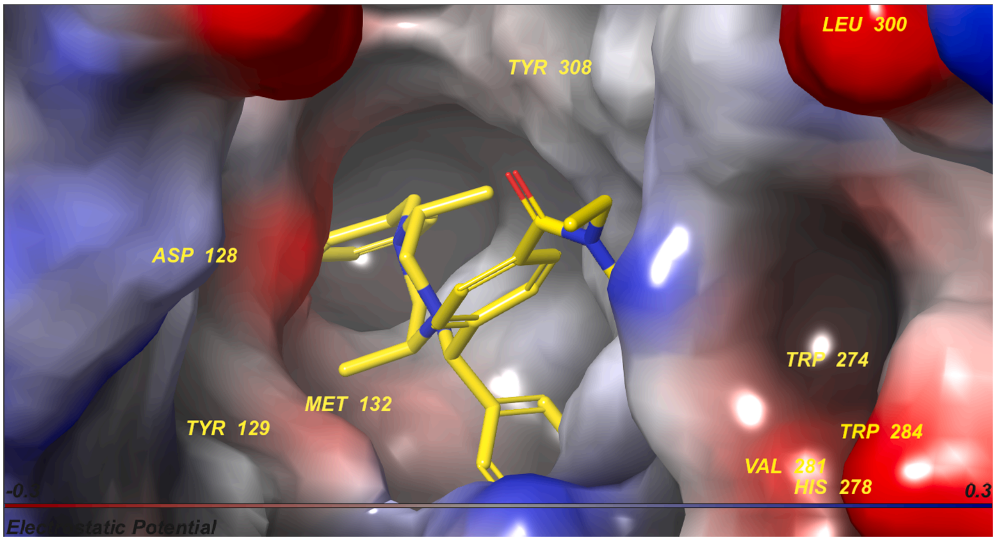

Fig. 2. Close-up view of the binding pocket of δ opioid receptors in complex with the DPI-287 agonist: PDB code: 6PT3.

The LigPrep module of the Schrodinger Suite was employed to perform geometric minimization of the compounds [29]. This process involved adding hydrogen atoms, adjusting bond lengths and angles to realistic values, correcting chiralities, ionization states, tautomers, stereochemistries, and ring conformations. The structures were assigned partial charges utilizing the OPLS-2005 force-field [30,31], and energy minimization was performed until the average root-mean-square deviation (RMSD) reached 0.001 Å. The ionization state at pH = 7 was determined using Epik's ionization tool. [32,33].

2.2. *Calculation and selection of the chemical descriptors*

Chemical descriptors can be described as numerical values or outcomes derived from mathematical operations that convert the encoded chemical information of a molecule into a meaningful representation. Alternatively, they can also be obtained through standardized experiments. Two-dimensional (2D) and three-dimensional (3D) molecular descriptors were calculated using PyBioMed [34] and PaDEL [35] libraries. The library PyBioMed is implemented in Python and is mainly based on toolkit RDKit (https://www.rdkit.org/) and Pybel [36] implementation, and depends on SciPy [37] and NumPy [38] python modules. Generally, each descriptor has different units and there are significant differences between different descriptors. The descriptors of all compounds in the training set were normalized and standardized using the scikit-learn ML library (https://scikit-learn.org/), and the results were transformed and appended to the internal test.

To identify the most pertinent descriptors and determine an appropriate number of features, highly correlated descriptors were initially eliminated using Pearson correlation analysis [39,40]. The ''select K-Best method'' [41] was employed as a supervised learning approach, utilizing the f-regressor function, to perform a univariate feature selection on the remaining descriptors. This method was utilized to reduce noise, eliminate redundancy, and effectively decrease the dimensions of the data. K-Best is an algorithm based on filtering that chooses prospective features based on a specific function σ (f, c), where f represents a feature and c represents a label.

2.3. *Nonlinear machine learning algorithms*

From the multitude of modeling machine learning approaches available, random forest (RF) [42] and extreme gradient boosting (XGB) [43,44] have been chosen for constructing the models in this study due to their proven effectiveness, robustness, and extensive utilization in QSAR modelling [43,45–48]. RF, introduced by Leo Breiman and Adele Cutler [42], is an extensively employed algorithm in machine learning for drug discovery tasks, regardless of the specific problem at hand. While it is challenging to designate a single model as the absolute best for all problem types, RF stands out for its exceptional performance, speed, and generalizability [21,22,49,50]. This algorithm utilizes multiple decision trees to train and predict samples. RF works by creating numerous decision trees during the training phase and then determining the most common class across all the trees to produce the final output [51]. This approach combines the concept of "bagging" with random feature selection to create an ensemble of decision trees that exhibit controlled variation [52]. Chen and Guestrin introduced XGBoost [44] a tree-boosting system that is widely used and effective in machine learning. The method builds upon the tree-based ensemble methods, but with the addition of a boosting step that improves the trees step-by-step by minimizing errors. Friedman [53] established the roots of this method. XGBoost's success can be attributed to its scalability on multiple levels and faster computation speed compared to other solutions. It is currently one of the best algorithms for solving a wide range of problems, particularly with larger datasets. The method's usefulness is demonstrated by its successful applications in machine learning and data mining challenges [43,54]. Currently, the algorithm produces state-of-the-art solutions for a wide range of problems, especially in larger datasets. Its advantages are clearly emphasized by the fruitful applications of the method in machine learning and data mining challenges. The RF and XGBOOST algorithms are highly renowned and extensively employed machine learning techniques that have found widespread application in resolving classification and regression problems [55–58].

2.4. *The evaluation of prediction regression model*

To assess the performance and goodness-of-fit of the models, a statistical analysis of various metrics including the regression coefficient (R), determination coefficient (R2), mean squared error (MSE), and root mean squared error (RMSE) was conducted for each machine learning algorithm. The hyperparameter model selection was determined based on the algorithm that yielded the lowest root mean squared error (RMSE):

$$R = 1 - \frac{\sum (y_i - \widehat{y}_i)}{\sum (y_i - \overline{y}_i)} \tag{1}$$

$$R^2 = 1 - \frac{\sum (y_i - \widehat{y}_i)^2}{\sum (y_i - \overline{y}_i)^2} \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2} \quad (4)$$

where y_i, $\widehat{y}_i$, $\overline{y_i}$ and n are the actual values, predicted values, actual average values, and the size of the dataset, respectively.

2.5. Applicability domain

The assessment of the applicability domain (AD) of the QSAR model is a crucial step in ensuring the reliability of predictions within the specific chemical space for which the model was developed [59–61]. The leverage approach was employed to establish the applicability domain for a specific compound [62–65]. In this context, our QSAR model was utilized to forecast the compound's activity:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \quad (5)$$

Where, x_i denotes the descriptor row vector for the compound in question, while X represents the descriptor matrix for the compounds in the training set. The warning leverage, which serves as an indicator of potential extrapolation, was also considered:

$$h^* = \frac{(p+1)}{n} \quad (6)$$

In this equation, n is the number of training compounds used for model fitting and p is the number of descriptors. The Williams plot [66] is used to interpret the AD inside a squared area within the leverage threshold (hi < h*) and ± 3.0 standard residual deviations ($\pm\delta$).

2.6. Ligand-target interaction fingerprint

The ligand-receptor molecular docking experiments of 1,793 ligands inside the active site of DOR (PDB code: 6PT3) [10] were performed using the Glide module implemented in the Schrödinger suite [67,68]. The enzyme structure was preprocessed, minimized, and refined using the Protein Preparation Wizard [69] in Schrödinger Suite [70]. To prepare the receptor structure at pH = 7, several actions were taken. These included removing any crystallographic waters, inserting any missing hydrogen or side chain atoms, and determining the correct charge and protonation state of the structure. Additionally, the ionization states of acidic and basic amino acid residues were taken into account. Furthermore, the protein structure went through an energy minimization process using the OPLS-2005 force-field This was done to alleviate any steric clashes that may have arisen due to the addition of hydrogen atoms among closely-spaced residues. A root mean square deviation (RMSD) cut-off value of 0.30 Å was used for this purpose. Preparation of the reference inhibitors (DPI-287: K_i = 0.39 nM; KGCHM07: Ki = 5.17 ± 1.57 nM) [10] were accomplished by applying the LigPrep module from the Schrodinger Suite [29]. This process involved adding hydrogen atoms, fine-tuning bond lengths and angles to a realistic configuration, addressing chirality concerns, determining ionization states, adjusting tautomers, stereochemistries, and optimizing ring conformations. Additionally, partial charges were assigned, and the ionization state was set to pH = 7 [33,71]. Out of the six available crystal structures for the δ opioid receptor with PDB codes 6PT3, 6PT2, 4RWA, 4RWD, 4N6H, and 4EJ4, only the 6PT3 and 6PT2 structures are in the active and native forms. In contrast, the 4RWA, 4RWD, 4N6H, and 4EJ4 structures are crystallized in mutated and inactive forms. As a result, this study has specifically chosen to focus on the 6PT3 and 6PT2 crystal structures along with their co-crystallized ligands, the small-molecule DPI-287 and the peptide KGCHM07.

In the next step, protein coordinates were extracted from the crystal structure of DOR bound to the selective agonist DPI-287 (PDB code: 6PT3). The appropriate receptor grid was generated based on a set of center coordinates (X = 4.22, Y = -40.91, Z = -46.61) using two cubical boxes having a common centroid to organize the calculations: a larger enclosing box and a smaller binding box with dimensions of 12 × 12 × 12 Å and 31 × 31 × 31 Å, respectively. The grid box was centered on the centroid of the crystallized inhibitor as a reference in the complex covering the binding site of DOR, which was sufficiently large to explore a larger region of the enzyme structure. The Glide extra-precision [67] (XP) mode was employed to obtain poses that fit the known pharmacophore of the DPI-287 and KGCHM07 ligands as well as 1,793 compounds. The best docking pose for each compound was the pose with the best docking scoring index that complied with the essential chemical interactions described for analog ligands (ECIDALs) [72,73]. The most obvious essential chemical interactions are the presence of interactions between the active residue of His278 and the co-crystal ligand (DPI-287) in the crystal structure of DOR (6PT3) [10] or interactions between Tyr 129 and Trp 284 and the co-crystal ligand (Peptide agonist KGCHM07) in the crystal structure of DOR (6PT2) [10] forming anπ-π stacking interaction.

In the second path, the selected poses from the previous step with docking scores lower than −5.0 kcalmol^{-1} were considered for the Structural Interaction Fingerprints [74,75] (SIFt) analysis. The Schrödinger Suite's Maestro module incorporates SIFt, which provides data regarding the ligand's proximity to receptor residues and the specific types of residues it interacts with [76]. The presence of various chemical interactions between ligands and the binding site residues of the target receptor is quantified using "bits." To establish the binding site, distance cut-offs are employed, and the interacting set consists of residues with atoms falling within the specified cut-off distance from ligand atoms. These bits are then recorded in an interaction matrix, which documents the specific chemical interactions between each ligand and each interacting residue within the receptor. Finally, the residual decomposition interaction energies were calculated to determine the key interacting residues involved in the binding affinities of the selected docked complexes.

3. Results

3.1. Exploratory data set analysis

Exploratory data analysis was applied to the data retrieved from G protein-coupled receptor database [11,25–27] (https://gpcrdb.org/). Of 4,217 molecules, 1,793 unique molecules were passed. As per the Rule of Five (RO5), compounds are more inclined towards oral absorption if they possess fewer than 5 hydrogen bond donors, fewer than 10 hydrogen bond acceptors, a molecular weight below 500 Daltons, and a logarithm of the partition coefficient (log P) less than 5.0. To eliminate the influence of magnitude, we converted the biological activities of the compounds to the logarithmic scale ($pK_i = -\log(K_i)$) and used them as dependent variables. The range of pK_i values was from 4 to 10 in the dataset. The SMILES notations of 1,793 compounds and their corresponding pK_i values are listed in csv format in the Supplementary Material. This dataset was randomly split into two sets of training, including cross validation sets and an external set (test set), with a ratio of 70:30, containing 1,255 training and 538 internal test sets. The training set was used for model generation, and the test set for model evaluation. Another 38 FDA-approved drugs targeting DOR were collected from the CHEMBL database (https://www.ebi.ac.uk/chembl/) [77], as an unseen external validation test set to further evaluate the selected predictive model (Data available as csv format in the Supplementary Materials).

3.2. Descriptor selection

A number of 1,211 two-dimensional (2D) and three-dimensional (3D) molecular descriptors of Basak, Burden, Molecular Constitutional, Geary Autocorrelation, Moran Autocorrelation, Moreau-Broto Autocorrelation, CATS2D, Charge, Molecular Connectivity Indices, Electro-topological State indices, Topological, MOE-type, Autocorrelation, Charged partial surface area (CPSA), Radial Distribution Function (RDF), Petitjean shape index, and Weighted Holistic Invariant Molecular (WHIM) descriptors were derived using PyBioMed [34] and PaDEL [35] libraries, Table 1.

To scale the total input descriptors with different units, the variables of the training set were normalized and standardized, and the results were transformed and appended to the internal test. The mean squared error (RMSE) was calculated using two ML regression models of RF (original = 0.688, normalized = 0.681, standardized = 0.680) and XGBOOST (original = 0.693, normalized = 0.685, standardized = 0.683). Finally, negligible RMSE variations were observed between the raw descriptors and the normalized/standardized values. Therefore, in this study, the raw descriptors were selected for further feature selection methods.

In the first stage of the feature selection method, from the total number of 1,211 descriptors, the highly correlated features with the a of 0.8 threshold were removed using Pearson correlation [39,40,78], which resulted in 520 variables. One strategy for enhancing QSAR models involves eliminating or reducing uninformative descriptors. Therefore, this process can enhance both the precision and resilience of the model. In the next stage, to evaluate the contribution of each 520 descriptors on the regression model performance and generalization ability, the K-best feature selection method-based f-regressor from the scikit-learn [79] Python module was used. According to the statistical analysis of the regression coefficient (R), determination coefficient for training (R^2), and determination coefficient for test (Q^2) from our models using ML algorithms, the ranking list of the most contributing descriptors is given in Supplementary Table S1.

Among the obtained results, K-best = 3 showed that the three features SLOGPVSA2, Chi6ch, and S17 from MOE-type descriptors through the contributions of SLogP and surface area as well as connectivity [80,81], and electrotopological state [82] from the 2D-descriptors category contributed significantly to the model performance (Fig. 3).

Table 1
2D and 3D-descriptors generated for 1,793 compounds using PyBioMed and PaDEL libraries.

Descriptors and Fingerprints	Number of Features	Type of Descriptors	Libraries
Basak	21	2D	PyBioMed
Burden	64	2D	PyBioMed
Molecular Constitutional	30	2D	PyBioMed
Geary Autocorrelation	32	2D	PyBioMed
Moran Autocorrelation	32	2D	PyBioMed
Moreau-Broto Autocorrelation	32	2D	PyBioMed
CATS2D	150	2D	PyBioMed
Charge	25	2D	PyBioMed
Molecular Connectivity Indices	44	2D	PyBioMed
Electrotopological State Indices	237	2D	PyBioMed
Topological	35	2D	PyBioMed
MOE-type	60	2D	PyBioMed
Autocorrelation	70	3D	PaDEL
CPSA	29	3D	PaDEL
RDF	18	3D	PaDEL
Petitjeanshapeindex	3	3D	PaDEL
WHIM	92	3D	PaDEL

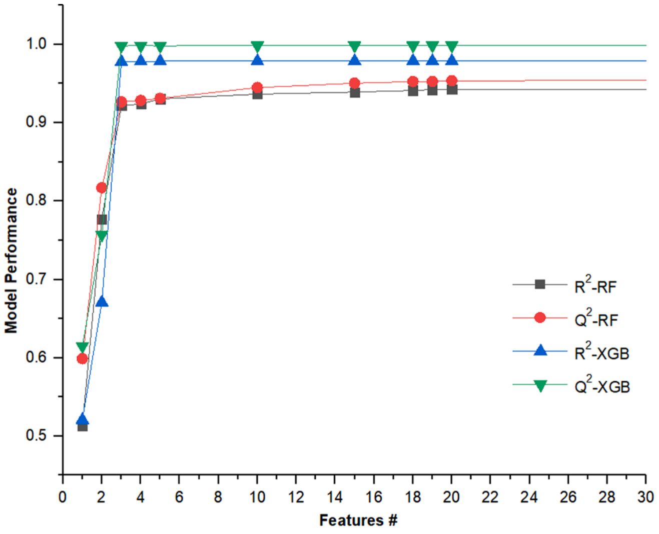

Fig. 3. Comparative model performance based on the number of features.

3.3. Model construction using ML algorithms

Two machine learning methods including RF and XGB, were employed for the regression model's construction for different numbers of features based on the K-best selection method. In this prediction, the RF parameters were set to n_estimators = 100 and the XGB parameters were defined with boosting rounds parameter values of 1000, maximum depth 7 and eta 0.1. Some statistical parameters, regression coefficient (R), determination coefficient for training (R^2), and determination coefficient for test (Q^2), as well as root mean squared error (RMSE) of five-fold cross validation (CV) from the training data were used to evaluate the performance of the regression models.

At the outset of the modeling process, 30 % of the dataset was chosen randomly and set aside as the hold-out test dataset, separate from the training phase. Following the execution of each algorithm run employing the SLOGPVSA2, Chi6ch, and S17 descriptors, the ultimate algorithm model was chosen by considering the R^2 value for the training set and the cross-validation root mean square error (CV_RMSE). The statistical analysis of the QSAR models based on three key features derived from the K-best feature selection method is presented in Table 2.

The model that was created using the training dataset of compounds was employed to forecast the (pKi) activity of the testing dataset of compounds. According to the obtained statistical parameters in Table 2, the XGB model was selected as the final algorithm with high activity-descriptor relationship efficiency based on the determination coefficient of the train set (R^2 = 0.978), determination coefficient for test (Q^2 = 0.998), regression coefficient for train set (R_train = 0.989), regression coefficient for test set (R_test = 0.999) and Root Mean Squared Error of fivefold cross validation (CV_RMSE = 0.461). The robustness of the best-generated model using the XGBOOST algorithm is depicted via the activity interactive graph that presents the predicted against experimental (pK_i) activity, as shown in Fig. 4.

Knowing the high predictive and descriptive ability, the generated model was considered to be highly robust in predicting the agonist activity of these compounds against Delta opioid receptor (DOR). The predicted activities of the studied compounds against DOR by the built QSAR model are provided in csv format in the Supplementary Materials. The model's capacity to establish a correlation between activity and structure is demonstrated by the smaller residual values observed in both the training and internal testing sets, as provided in the supplementary material in CSV format. The correlation between the experimental activity and the predicted activity according to the model was highly significant as determined by statistical analysis. The closeness of regression-coefficient (R_train and R_test) to 1.0 indicates that the

Table 2
Statistical parameters for train, test, cross validation, and unseen external test sets for both considered algorithms presented.

Statistical analysis	R^2	Q^2	R_train	R_test	CV_RMSE	R^2_Test^{ext}	R_Test^{ext}
RF	0.922	0.927	0.965	0.749	0.514	0.841	0.949
XGB	0.978	0.998	0.989	0.999	0.461	0.986	0.993

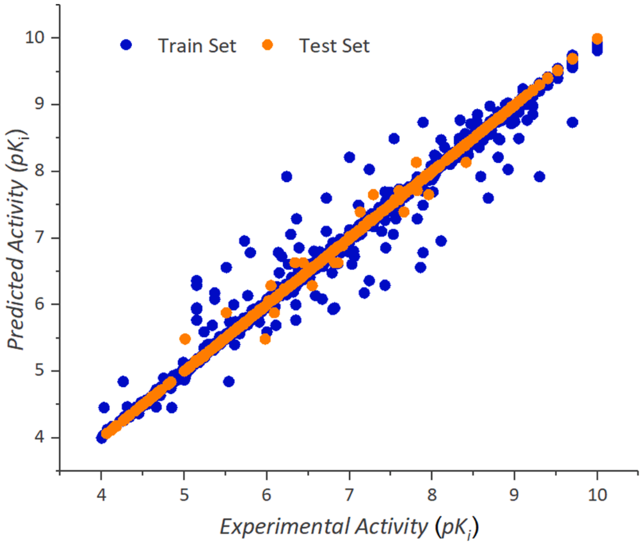

Fig. 4. The graphical representation of the predicted activities versus experimental activities of both the train and test sets for the best predictive xgboost model.

developed model elaborated a large portion of the descriptor-variation large enough for a good QSAR model, indicating that the model is highly predictive and excellent. The high R test value in the developed model signifies that the model is capable of delivering reliable and accurate predictions for novel compounds. A good and acceptable QSAR model must obey the following criteria: regression-coefficient for train and test sets (R_train and R_test) values close to one and the lowest RMSE value for cross validation close to zero [39,58,65,83]. The generated QSAR model met the criteria and was therefore statistically acceptable. To further validate the predictive model verification, 38 FDA-approved drugs as an unseen external test set were given to our studied models. Determination of coefficient (R^2_Test^{ext} = 0.986) and regression coefficient (R_Test^{ext} = 0.993) of the XGB model reconfirmed the robustness of our proposed model (Table 2). The predicted activity values of 38 drugs in csv format are presented in the Supplementary Materials.

3.4. Applicability domain analysis

An essential aspect of confirming the validity of the QSAR model's predictions is the establishment of its applicability domain (AD). Reliable predictions can only be made by a QSAR model for chemicals that fall within its applicable domain (AD), and not through extrapolation beyond this domain. Various techniques exist to establish the applicability domain (AD) of QSAR models [59], with the most prevalent approach being the calculation of leverage values for each compound [66]. To visualize the applicability domain, a Williams plot, which represents the standardized residuals against the leverage (hi), was employed [84]. A molecule with a leverage value (hi) greater than h* has an impact on the performance of the QSAR model and can be excluded from AD. Additionally, the range within ± 3 δ (standard deviations) of standardized residuals is commonly adopted as a threshold for confirming predictions regarding a molecule. This is because data points falling within ± 3 δ of standardized residuals from the mean encompass 99 % of the data in a normal distribution [85].

From the Williams plot (Fig. 5), thirteen test compounds were found to be out of the warning leverage (h*) criteria, which indicates potential model extrapolation (the warning leverage limit is 0.010).

This plot placed the AD within ± 3.0 standard residual deviations and the warning leverage (0.010) (Fig. 5). thirteen test compounds were found to be outside the warning leverage, indicating potential model extrapolation. Among the total number of 1,793 compounds, the insignificant numbers of 16 test sets and 38 train sets were observed as outlier compounds with standardized residuals with more/less than 3 δ /-3 δ standard deviation units. Notably, 98 % of the input compounds fell within the AD acceptable criteria, confirming the reliability of the model.

3.5. Ligand-target interaction contact analysis

A systematic and detailed analysis of all possible interactions involved in forming protein–ligand complexes between DOR and 1,793 compounds was investigated using Docking [67,68,86] calculations and the Interaction Fingerprints (IFPs) [87–89] panel of Maestro [76] implemented in the Schrodinger suite.

To verify the conducted docking simulations for subsequent analysis of interaction fingerprints (IFPs), we conducted docking of the co-crystal DOP-DPI-287and KGCHM07 inhibitors with the DOR protein. The most favorable docking orientation was determined based on factors such as binding energy and interactions between the ligand and receptor within the active site pocket. This optimal docking pose was then aligned with the co-crystallized structure of the DOP-DPI-287 inhibitor, resulting in a calculated RMSD of 0.102 Å, Fig. 6.

The value of IFPs lies in their capacity to encompass a wide range of interaction types that take place between a target protein and its ligands. Various categories of chemistries are considered in IFP calculations, including polar (P), hydrophobic (HP), hydrogen bond acceptor (HBA), where the residue accepts a hydrogen bond, hydrogen bond donor (HBD), where the residue donates a hydrogen bond, aromatic (Ar), and interactions with charged groups through electrostatic forces (CH). Information about contacts with backbone and side-chain groups is also provided in Figures S1, S2, S7, S8, S13, and S14. First, we calculated IFPs by considering two DOR–inhibitor complexes reported in PDB (PDB IDs 6PT2, and 6PT3) and presented in Figures S1-S12. Subsequently, we conducted a similar calculation, taking into account the complexes that were established through docking calculations involving our set of 1,793 compound structures (Figures S13-S20). We anticipate that there is a similarity between the IFPs of our docking poses and those of the DOR-inhibitor complexes reported in PDB format. The outcome of our docking calculations resulted in 10,857 poses for 1,793 compounds by considering 10 poses for each compound in our docking protocol. Among 10,875 poses, the best docking pose for each compound was selected based on the best docking scoring that complies with the essential chemical interactions described for analog ligands (ECIDALs) [72,73] These selection criteria filtered our docked poses to 1,051 complex structures. The IFP analysis applied to the two DOR–inhibitor complexes reported in PDB revealed 13 common DOR residual contacts with the co-crystalized Peptide agonists KGCHM07 and DPI-287 (Fig. 7).

These binding site residues include Asp 128, Tyr 129, and Met 132 of the TM3 region, Val 217 of TM5, Trp 274, Ile 277, His 278, Phe 280, Val 281, and Trp 284 of TM 6, Leu 300, Ile 304, and Tyr 308 of TM7.

Conversely, when the IFP analysis was used on the 1,051 complexes involving DOR and the optimal pose of each compound obtained through docking, it disclosed that there were 10 shared residues making

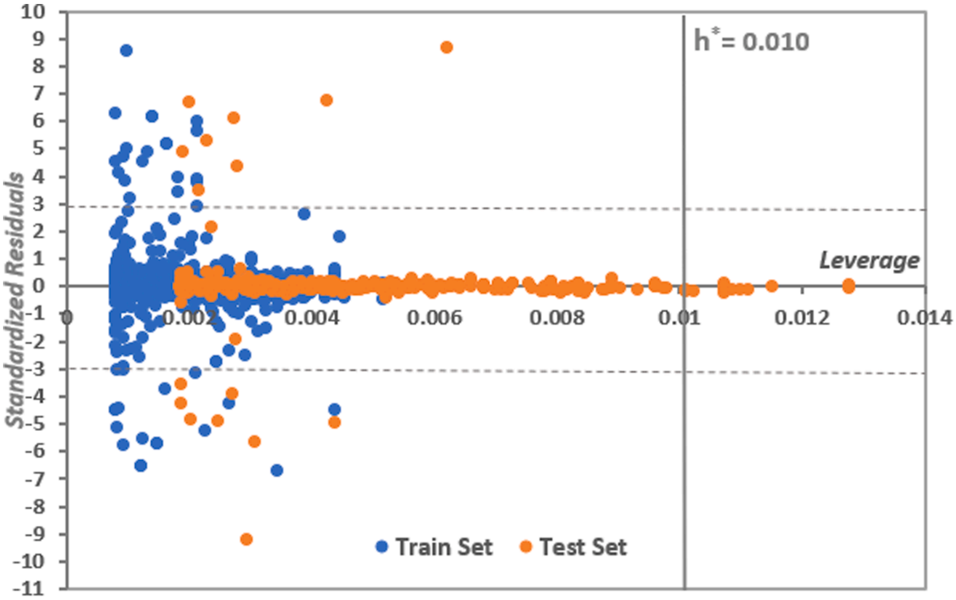

Fig. 5. Williams plot presented to evaluate the applicability domain of the best XGBOOST predictive model.

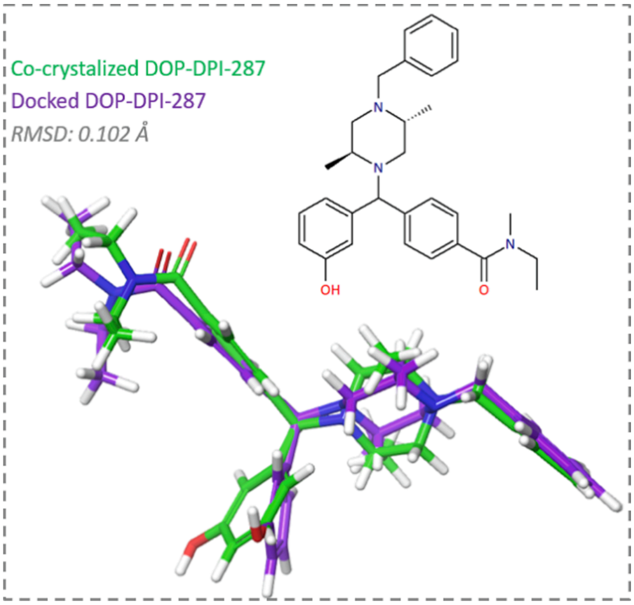

Fig. 6. The superposition of the co-crystallized DPI-287 and its best-docked pose based on the ECIDALs approach is presented.

contact with the co-crystallized ligands in the DOR complexes with the contributions of Asp 128 (100 %), Tyr 129 (80 %), Met 132 (90 %), Val 217 (50 %), Trp 274 (95 %), Ile 277 (90 %), Val 281 (70 %), Leu 300 (50 %), Ile 304 (90 %), and Tyr 308 (95 %) (Fig. 8).

Since the DOR binding site is predominantly hydrophobic, there were no instances of hydrogen bonding interactions observed when examining the occurrence of chemical contacts in the structures documented in PDB entries. The HBD and HBA showed negligible contributions in the docked structures, (Figures S17 and S18). The obtained IFP calculations indicated that our docking results conserve the main interactions observed for the available PDB structures: Asp 128 showed 100 % polar and electrostatic common contributions in the docked complexes and PDB structures (FiguresS6 and S12). The residues of Tyr 129, Met 132, Trp 274, Ile 277, Val 281, and Tyr 308 exhibited more than 70 % hydrophobic contribution, whereas Val 217 and Ile 304 exhibited > 50 % hydrophobic contribution in the docked complexes (Figure S19). Tyr 129, Met 132, Val 217, Trp 274, Ile 277, Phe 280, Val 281, Trp 284, Leu 300, Ile 304, and Tyr 304 residues are common binding site residues with 100 % hydrophobic contributions in the co-crystalized Peptide agonists KGCHM07 and DPI-287 bound to DOR (Figures S5 and S11). The IFPs used for both the PDB and docked structures validate the accuracy of our docking experiments, as they reveal that the residues present in the DOR binding site of the structures determined through X-ray crystallography are consistent with the ones identified in our docking poses.

To obtain detailed information about the key residues' contribution to the binding affinities of the docked complexes, the residual decomposition interaction energies were plotted. This analysis was considered for four docked complexes containing ligands with less potent inhibitory activities (low pKi value and higher docking score), such as CHEMBL3695269, CHEMBL3698762, CHEMBL3698760, and CHEMBL3698851, as well as four docked structures containing ligands with more potent inhibitory activities (high pKi value and more favorable docking score), namely, CHEMBL3965191, CHEMBL4587201, CHEMBL605293, and CHEMBL2370431.

As depicted in Fig. 9, the most significant interaction in all eight studied complexes is related to Asp 128 as a key active site residue responsible for DOR activation. The strength of this interaction is considerably higher for inhibitors with high pKi values and docking score. According to Fig. 9, Among the four inhibitors with stronger activity, CHEMBL2370431 exhibited significantly high affinity for the Asp 128 active site residue (Eint: -40.0 kcalmol^{-1}). Although, in all eight docked complexes, Tyr 129, Met 132, Trp 274, Ile 277, and Tyr 308 residues showed acceptable interactions with the ligands but these interactions are not comparable with the affinity magnitude of Asp 128. The depiction of the orientations of the docked complexes is acknowledged as a valuable insight into potent DOR inhibitors, which could be instrumental in the development of new, effective inhibitors.

4. Discussion

A thorough exploratory data analysis conducted on a GPCR (G Protein-Coupled Receptor) database, with the aim of developing a predictive model for agonist activity against the Delta opioid receptor (DOR). The analysis encompasses various aspects of the analysis, including dataset characteristics, selection criteria based on the Rule of

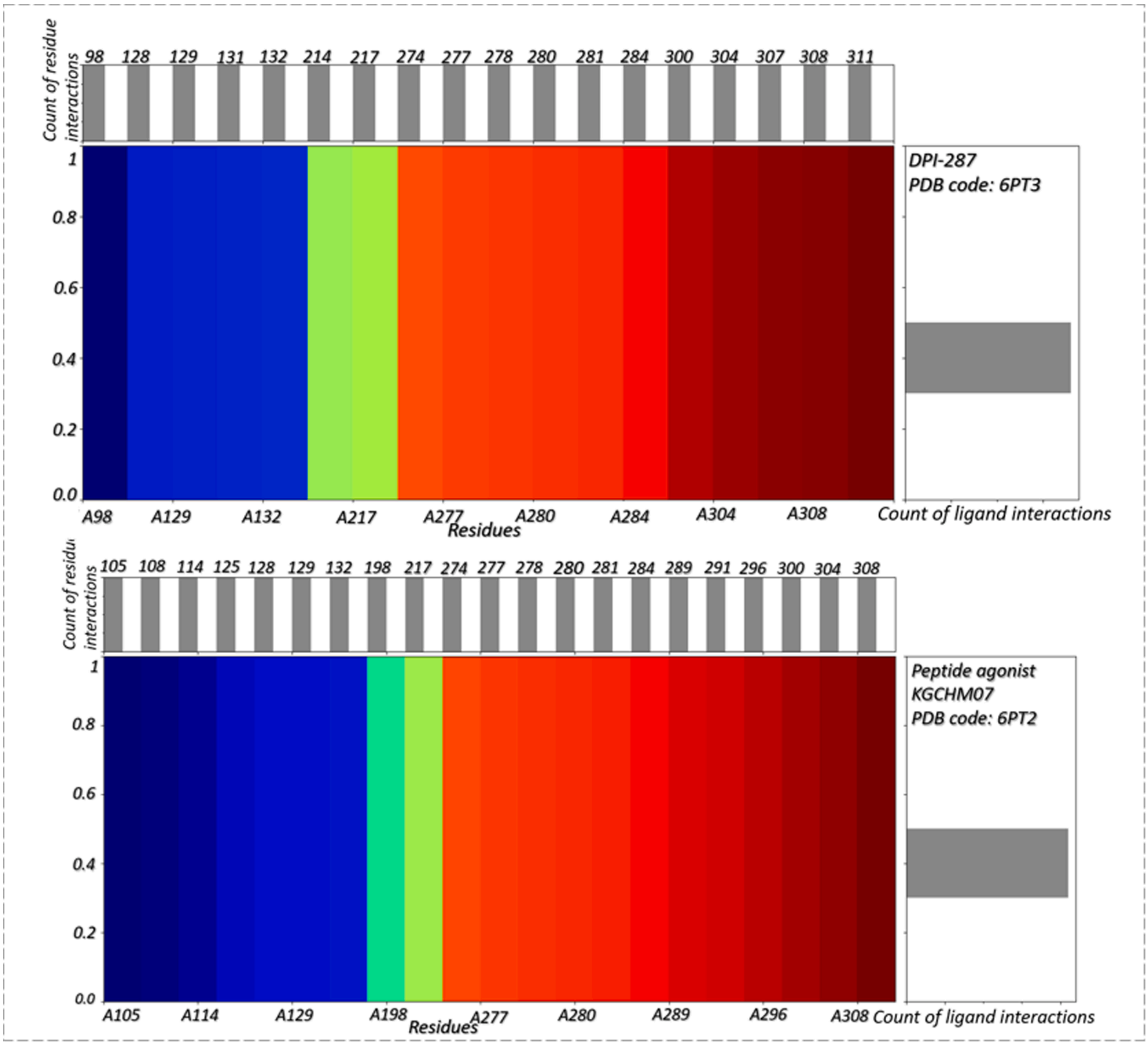

Fig. 7. Interaction map of all-in-contact residues for co-crystalized DPI-287 (PDB code:6PT3) and the peptide agonist KGCHM07 (PDB code: 6PT2) at the DOR–ligand binding interface.

Five (RO5) for oral absorption, conversion of biological activity to pKi values, division of the dataset into training, test, and external validation sets, feature selection process, identification of influential descriptors, construction of regression models using machine learning algorithms, validation of the models' predictive performance, determination of the applicability domain, and analysis of protein–ligand interactions between DOR and compounds. In the first step, a description of the dataset used for the analysis is provided, which consists of 4,217 molecules. Based on the criteria defined by the Rule of Five (RO5), a total of 1,793 unique molecules were generated. The conversion of biological activity to pKi values is explained to enable the use of a logarithmic scale. The dataset is then divided into a training set (70 %) and an internal test set (30 %) for model generation and evaluation, respectively. Additionally, a separate set of 38 FDA-approved drugs targeting DOR is used as an external validation test set. The second step focuses on the feature selection process and the identification of influential descriptors for the regression model. A wide range of 1,211 2D and 3D molecular descriptors from various libraries were derived, covering categories and properties related to molecular structure and properties. Normalization and standardization of the variables were performed to handle descriptors with different units. Notably, the performance of models using raw, normalized, or standardized descriptors showed negligible variations. Highly correlated features were removed early in the feature selection stage to enhance model accuracy and robustness. The remaining 520 descriptors were evaluated using the K-best feature selection method to determine their contribution to the model's performance and generalization ability. The most influential descriptors, namely SLOGPVSA2, Chi6ch, and S17, belonging to the MOE-type descriptor category, provided valuable information about octanol/water partition coefficient, surface area, connectivity, and electrotopological state, shedding light on their relevance in predicting the target properties.

Continuing the analysis, machine learning algorithms, specifically Random Forest (RF) and XGBoost (XGB), were utilized to construct regression models based on the selected features. The parameters of RF and XGB were defined, and statistical parameters such as regression coefficient (R), determination coefficients (R2 and Q2), and Root Mean Squared Error (RMSE) were employed to evaluate the models' performance. The XGB model was ultimately selected as the final algorithm due to its high efficiency in capturing the activity-descriptor relationship, supported by favorable determination coefficients and regression coefficients. The model's predictive capability and descriptive ability were demonstrated through an activity interactive graph, and its compliance with key criteria for a good QSAR (Quantitative Structure-Activity Relationship) model further confirmed its acceptability statistically. Validation of the model using an external test set of FDA-

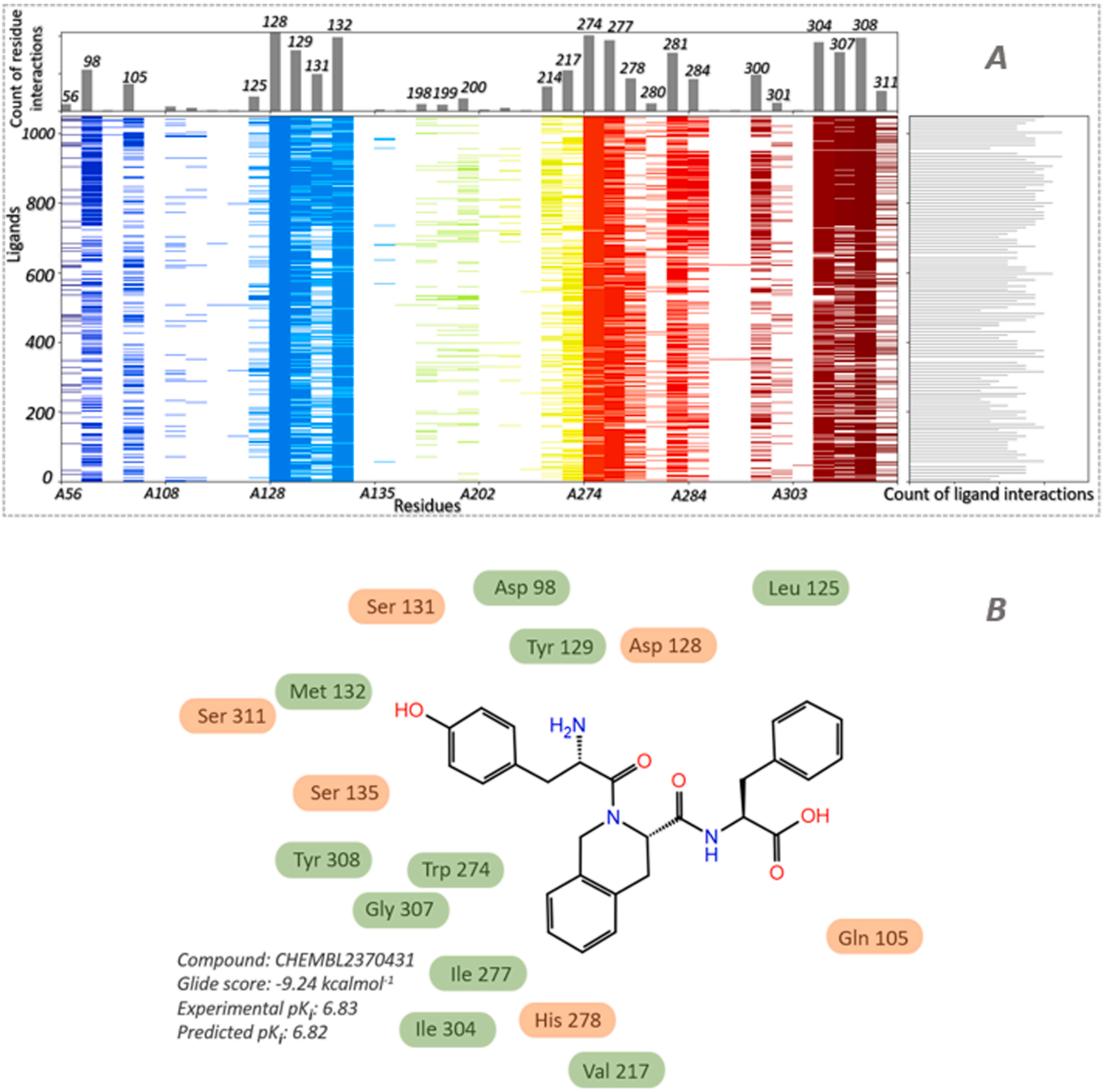

Fig. 8. A: Interaction map of all-in-contact residues for 1,050 docked complexes at the DOR–ligand binding interface; B: Close up of the residues contributing to the contact analysis for a docked compound with a higher Glide score and pK_i.

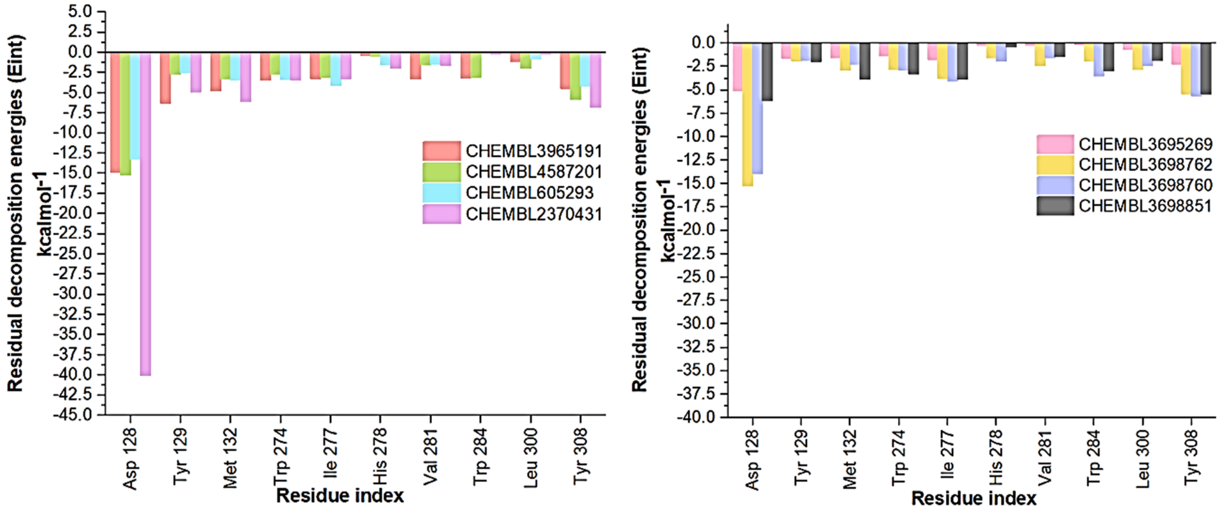

Fig. 9. Bar contribution of binding site residues to binding affinities of more potent docked inhibitors.

approved drugs strengthened its robustness and accuracy in predicting agonist activity against DOR. The next step of the analysis focused on the importance of validating the predicted QSAR model through the applicability domain (AD). The AD defines the range of compounds for which the model's predictions are reliable, avoiding extrapolations beyond the trained data. The Williams plot, which visualizes standardized residuals versus leverage values, was used to define the AD. Compounds exceeding the warning leverage threshold and those with standardized residuals outside the defined range were identified as potential outliers and typically excluded from the AD. The analysis revealed that a significant majority of the compounds fell within the acceptable criteria of the AD, confirming the model's reliability for most of the dataset.

Lastly, protein–ligand interactions between DOR and compounds were explored. Docking calculations and IFPs (Interaction Fingerprints) were employed for this analysis. The docking simulations were validated by comparing the docked pose of the co-crystallized inhibitors with the corresponding experimental structure, showing good alignment. IFPs were calculated to capture various types of interactions between DOR and ligands. The analysis revealed common binding site residues involved in the interactions and confirmed the reliability of the docking experiments. Important residues, particularly Asp 128, were identified, and their interactions with ligands were analyzed, providing insights for the design of novel DOR inhibitors.

Overall, the presented results provide a comprehensive analysis of the GPCR database, covering feature selection, model construction, validation, determination of the applicability domain, and analysis of protein–ligand interactions. This research serves as a foundation for the development of a robust predictive model for agonist activity against DOR, demonstrating high predictive performance and the potential for practical application in drug discovery.

5. Conclusion

In this study, a comprehensive QSAR model-based machine learning approach was constructed using 4,218 agonist compounds with inhibitory bioactivities measured using a consistent bioassay procedure in K_i format. An exploratory data analysis (EDA) approach was applied to refine SMILES notations, duplicate molecules, salt forms, heavy metals and fragments. Finally, the compounds were filtered based on Lipinski's rule of five (RO5). The compounds underwent geometric optimization, which entailed the addition of hydrogen atoms, fine-tuning bond lengths and angles to realistic values, rectifying chirality, addressing ionization states, optimizing tautomers, stereochemistry, and ring shapes. Subsequently, partial charges were assigned to the structures, followed by an energy minimization process at pH = 7. A total of 1,211 2D and 3D molecular descriptors were derived using PyBioMed and PaDEL libraries. The K-best features selection method revealed three key structural features such as SLOGPVSA2, Chi6ch, and S17 contributed significantly to the XGBOOST model performance. Statistical analysis, internal K-fold cross-validation, and external validation using 38 unseen FDA-approved drug data confirmed the robustness of the predictive model. Applicability domain (AD) analysis using William plot confirmed the reliability of the model by falling 98 % of the input compounds within its acceptable ± 3.0 standard residual deviations. A molecular docking study along with ligand–receptor contacts fingerprints (LRCFs) analysis revealed the key contact interactions of Asp 128, Tyr 129, Met 132, Trp 274, Ile 277, and Tyr 308 residues in the total binding affinities upon complexation of the ligands-DOR. Our study using regression QSAR along with ligand–receptor contact fingerprints analysis could serve in designing new agonist compounds to effectively target DOR.

Ethical statement

This work does not involve the use of humans or animals.

Author contributions

Zeynab Fakhar: Idea generation, Data organization, Systematic analysis, Research approach, Software tools, Verification, Initial manuscript composition and writing.

Ali Hosseinpouran: Reviewing the statistical analysis procedure and Methodologies.

Orde Q. Munro: Providing computational resources, reviewing and Editing the manuscript.

Sorena Sarmadi: Reviewing the statistical analysis procedure and Methodologies.

Sajjad Gharaghani: as the corresponding author conceived, designed, and supervised the research.

CRediT authorship contribution statement

Zeynab Fakhar: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ali Hosseinpouran:** Validation, Formal analysis. **Orde Q. Munro:** Software, Resources. **Sorena Sarmadi:** Validation, Formal analysis. **Sajjad Gharaghani:** Supervision, Investigation, Data Curation, 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

ZF would like to acknowledge the CHPC server based in Cape Town, South Africa, for providing computational resources. OQM acknowledges support from the National Research Foundation (NRF) grant No. 63799.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.medidd.2024.100176.

References

[1] WHO (2019). International Classification of Diseases for Mortality and Morbidity Statistics. Eleventh Revision. WHO (2019).

[2] Degenhardt L, et al. Estimating treatment coverage for people with substance use disorders: an analysis of data from the World Mental Health Surveys. World Psychiatry 2017;16:299–307.

[3] Ågmo A, Gómez M. Conditioned place preference produced by infusion of Met-enkephalin into the medial preoptic area. Brain Res 1991;550:343–6.

[4] Sauriyal DS, Jaggi AS, Singh N. Extending pharmacological spectrum of opioids beyond analgesia: Multifunctional aspects in different pathophysiological states. Neuropeptides 2011;45:175–88.

[5] Chen Y, Mestek A, Liu J, Yu L. Molecular cloning of a rat κ opioid receptor reveals sequence similarities to the μ and δ opioid receptors. Biochem J 1993;295:625–8.

[6] Minami M, et al. In situ hybridization study of κ-opioid receptor mRNA in the rat brain. Neurosci Lett 1993;162:161–4.

[7] Evans CJ, Keith DE, Morrison H, Magendzo K, Edwards RH. Cloning of a Delta Opioid Receptor by Functional Expression. Science 1992;1979(258):1952–5.

[8] Kieffer BL, Befort K, Gaveriaux-Ruff C, Hirth CG. The delta-opioid receptor: isolation of a cDNA by expression cloning and pharmacological characterization. Proceedings of the National Academy of Sciences 1992;89:12048–52.

[9] Stein C. Opioid Receptors. Annu Rev Med 2016;67:433–51.

[10] Claff T, et al. Elucidating the active δ-opioid receptor crystal structure with peptide and small-molecule agonists. Sci Adv 2022;5:eaax9115.

[11] Munk C, Harpsøe K, Hauser AS, Isberg V, Gloriam DE. Integrating structural and mutagenesis data to elucidate GPCR ligand binding. Curr Opin Pharmacol 2016;30:51–8.

[12] Fenalti G, et al. Molecular control of δ-opioid receptor signalling. Nature 2014;506:191–6.

[13] Collu F, Ceccarelli M, Ruggerone P. Exploring Binding Properties of Agonists Interacting with a δ-Opioid Receptor. PLoS One 2012;7:e52633-.
[14] Kieffer BL, Gavériaux-Ruff C. Exploring the opioid system by gene knockout. Prog Neurobiol 2002;66:285–306.
[15] Gavériaux-Ruff C, Kieffer BL. Delta opioid receptor analgesia: recent contributions from pharmacology and molecular approaches. Behav Pharmacol 2011;22.
[16] Gendron L, Cahill CM, von Zastrow M, Schiller PW, Pineyro G. Molecular Pharmacology of <em>δ</em>-Opioid Receptors. Pharmacol Rev 2016;68:631.
[17] Chung CS, P. & Kieffer, B. L.. Delta opioid receptors in brain function and diseases. Pharmacol Ther 2013;140:112–20.
[18] Mohamud AO, et al. Functional Characterization of Sodium Channel Inhibitors at the Delta-Opioid Receptor. ACS Omega 2022;7:16939–51.
[19] Meqbil YJ, van Rijn RM. Opportunities and Challenges for In Silico Drug Discovery at Delta Opioid Receptors. Pharmaceuticals 2022;15.
[20] Podlewska S, Kurczab R. Mutual Support of Ligand- and Structure-Based Approaches—To What Extent We Can Optimize the Power of Predictive Model? Case Study of Opioid Receptors *Molecules* 2021;26.
[21] Sakamuru S, et al. Predictive Models to Identify Small Molecule Activators and Inhibitors of Opioid Receptors. J Chem Inf Model 2021;61:2675–85.
[22] Carracedo-Reboredo P, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J 2021;19:4538–58.
[23] Fourches D, Ash J. 4D- quantitative structure–activity relationship modeling: making a comeback. Expert Opin Drug Discov 2019;14:1227–35.
[24] Taha MO, et al. 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;51:647–69.
[25] Pándy-Szekeres G, et al. The G protein database, GproteinDb. Nucleic Acids Res 2022;50:D518–25.
[26] Kooistra AJ, et al. GPCRdb in 2021: integrating GPCR sequence, structure and function. Nucleic Acids Res 2021;49:D335–43.
[27] Munk C, et al. GPCRdb: the G protein-coupled receptor database – an introduction. Br J Pharmacol 2016;173:2195–207.
[28] Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 2004;1:337–41.
[29] Schrödinger Release 2020-3: LigPrep, Schrödinger, LLC, New York, NY, 2020.
[30] Harder E, et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J Chem Theory Comput 2016;12:281–96.
[31] Shivakumar D, et al. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. J Chem Theory Comput 2010;6:1509–19.
[32] Greenwood JR, Calkins D, Sullivan AP, Shelley JC. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. J Comput Aided Mol Des 2010;24:591–604.
[33] Schrödinger Release 2020-3: Epik, Schrödinger, LLC, New York, NY, 2020.
[34] Dong J, et al. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. J Cheminform 2018;10:16.
[35] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011;32:1466–74.
[36] O'Boyle NM, Morley C, Hutchison GR. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. Chem Cent J 2008;2:5.
[37] Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020:17:261–272.
[38] Harris CR, et al. Array programming with NumPy. Nature 2020;585:357–62.
[39] Du Z, Wang D, Li Y. Comprehensive Evaluation and Comparison of Machine Learning Methods in QSAR Modeling of Antioxidant Tripeptides. ACS Omega 2022; 7:25760–71.
[40] Jhin C, Hwang KT. Adaptive Neuro-Fuzzy Inference System Applied QSAR with Quantum Chemical Descriptors for Predicting Radical Scavenging Activities of Carotenoids. PLoS One 2015;10:e0140154-.
[41] Ferri FJ, Pudil P, Hatef M, Kittler J. Comparative study of techniques for large-scale feature selection. in Machine Intelligence and Pattern. Recognition 1994;16: 403–13.
[42] Breiman L. Random Forests. Mach Learn 2001;45:5–32.
[43] Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. J Chem Inf Model 2016; 56:2353–60.
[44] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016: 785–794.
[45] Wu Z, et al. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinform 2021;22:bbaa321.
[46] Schwaighofer A, et al. Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. J Chem Inf Model 2007;47:407–24.
[47] Schroeter TS, et al. Predicting Lipophilicity of Drug-Discovery Molecules using Gaussian Process Models. ChemMedChem 2007;2:1265–7.
[48] Wu Z, et al. ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches. J Chem Inf Model 2019;59:4587–601.
[49] Burggraaff L, et al. Identification of novel small molecule inhibitors for solute carrier SGLT1 using proteochemometric modeling. J Cheminform 2019;11:15.
[50] Lee KH, et al. Toward Reducing hERG Affinities for DAT Inhibitors with a Combined Machine Learning and Molecular Modeling Approach. J Chem Inf Model 2021;61:4266–79.
[51] Brian Houston J, Carlile DJ. Prediction of Hepatic Clearance from Microsomes, Hepatocytes, and Liver Slices. Drug Metab Rev 1997;29:891–922.
[52] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. in Proceedings of the 14th International Joint Conference on Artificial Intelligence 1995:2:1137–1143.
[53] Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Ann Stat 2001;29:1189–232.
[54] Zhang D, Gong Y. The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure. IEEE Access 2020;8: 220990–1003.
[55] Mswahili ME, Martin GL, Woo J, Choi GJ, Jeong Y-S. Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum. Biomolecules 2021;11.
[56] Siramshetty VB, et al. Critical Assessment of Artificial Intelligence Methods for Prediction of hERG Channel Inhibition in the "Big Data" Era. J Chem Inf Model 2020;60:6007–19.
[57] Chen X, et al. Discovery of Dual FGFR4 and EGFR Inhibitors by Machine Learning and Biological Evaluation. J Chem Inf Model 2020;60:4640–52.
[58] Mamada H, Nomura Y, Uesawa Y. Novel QSAR Approach for a Regression Model of Clearance That Combines DeepSnap-Deep Learning and Conventional Machine Learning. ACS Omega 2022;7:17055–62.
[59] Lennart E, et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 2003;111:1361–75.
[60] Wang Z, Chen J, Hong H. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. Environ Sci Technol 2021;55:6857–66.
[61] Todeschini R, Consonni V, Gramatica P. 4.05 - Chemometrics in QSAR. Comprhenive Chemometrics 2009:129–72.
[62] Roy PP, Kovarich S, Gramatica P. QSAR model reproducibility and applicability: A case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-)triazoles. J Comput Chem 2011;32: 2386–96.
[63] Rakhimbekova A, et al. Comprehensive Analysis of Applicability Domains of QSPR Models for Chemical Reactions. Int J Mol Sci 2020;21:5542.
[64] Gajewicz A. How to judge whether QSAR/read-across predictions can be trusted: a novel approach for establishing a model's applicability domain. Environ Sci Nano 2018;5:408–21.
[65] Gharaghani S, Khayamian T, Ebrahimi M. Molecular dynamics simulation study and molecular docking descriptors in structure-based QSAR on acetylcholinesterase (AChE) inhibitors. SAR QSAR Environ Res 2013;24:773–94.
[66] Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci 2007;26:694–701.
[67] Friesner RA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J Med Chem 2006;49:6177–96.
[68] Halgren TA, Glide,, et al. A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J Med Chem 2004;47: 1750–9.
[69] Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des 2013;27:221–34.
[70] Schrödinger Release 2020-3: Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2020; Impact, Schrödinger, LLC, New York, NY; Prime, Schrödinger, LLC, New York, NY, 2020.
[71] Shelley JC, et al. Epik: a software program for pKaprediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 2007;21:681–91.
[72] Muñoz-Gutierrez C, Adasme-Carreño F, Fuentes E, Palomo I, Caballero J. Computational study of the binding orientation and affinity of PPARγ agonists: inclusion of ligand-induced fit by cross-docking. RSC Adv 2016;6:64756–68.
[73] Ramírez D, Caballero J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? Molecules 2018;23.
[74] 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;47:337–44.
[75] Singh J, Deng Z, Narale G, Chuaqui C. Structural Interaction Fingerprints: A New Approach to Organizing, Mining, Analyzing, and Designing Protein-Small Molecule Complexes. Chem Biol Drug Des 2006;67:5–12.
[76] Schrödinger Release 2020-3. Maestro. New York, NY: Schrödinger, LLC; 2020.
[77] Gaulton A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucl Acids Res 2012;40:D1100–7.
[78] Biesiada, J. & Duch, W. Feature Selection for High-Dimensional Data — A Pearson Redundancy Based Filter. Computer Recognition Systems 2;2007:242–249.
[79] Pedregosa, F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research; 2011:12:2825–2830.
[80] Hall LH, Kier LB. The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Relations. In: Boyd D, Lipkowitz K, editors. Reviews of Computational Chemistry. New York: VCH Publishers Inc.; 1991. p. 367–422.
[81] Kier LB, Hall LH. Molecular Connectivity in Structure-Activity Analysis. New York: John Wiley and Sons; 1986.
[82] Pearlman SR. Molecule Structure Description: The Electrotopological State. J. Am. Chem. Soc. 2000;122(26):6340.
[83] Ruark CD, Hack CE, Robinson PJ, Anderson PE, Gearhart JM. Quantitative structure–activity relationships for organophosphates binding to acetylcholinesterase. Arch Toxicol 2013;87:281–9.

[84] Netzeva TI, et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships: The Report and Recommendations of ECVAM Workshop 521,2. Altern Lab Anim 2005;33:155–73.

[85] Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. Altern Lab Anim 2005;33:445–59.

[86] Friesner RA, Glide,, et al. A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J Med Chem 2004;47:1739–49.

[87] Velázquez-Libera JL, Navarro-Retamal C, Caballero J. Insights into the Structural Requirements of 2(S)-Amino-6-Boronohexanoic Acid Derivatives as Arginase I Inhibitors: 3D-QSAR, Docking, and Interaction Fingerprint Studies. Int J Mol Sci 2018;19.

[88] Caballero J, Morales-Bayuelo A, Navarro-Retamal C. Mycobacterium tuberculosis serine/threonine protein kinases: structural information for the design of their specific ATP-competitive inhibitors. J Comput Aided Mol Des 2018;32:1315–36.

[89] Velázquez-Libera JL, Rossino G, Navarro-Retamal C, Collina S, Caballero J. Docking, Interaction Fingerprint, and Three-Dimensional Quantitative Structure-Activity Relationship (3D-QSAR) of Sigma1 Receptor Ligands, Analogs of the Neuroprotective Agent RC-33. Front Chem 2019;7.