Virtual Screening of DPP-4 Inhibitors Using QSAR-Based Artificial Intelligence and Molecular Docking of Hit Compounds to DPP-8 and DPP-9 Enzymes

Oky Hermansyah
Universitas Indonesia https://orcid.org/0000-0002-7701-3809

Alhadi Bustamam
Universitas Indonesia

Arry Yanuar (arry.yanuar@ui.ac.id)
Universitas Indonesia https://orcid.org/0000-0001-8895-9010

Research article

Keywords: Artificial Intelligence, DPP-4, KNIME, Machine Learning, QSAR, Virtual Screening

Posted Date: June 3rd, 2020

DOI: https://doi.org/10.21203/rs.2.22282/v2

Research article

Virtual Screening of DPP-4 Inhibitors Using QSAR-Based Artificial Intelligence and Molecular Docking of Hit Compounds to DPP-8 and DPP-9 Enzymes

Oky Hermansyah[1], Alhadi Bustamam[2], Arry Yanuar[1*]

[1] *Laboratory of Biomedical Computation and Drug Design, Faculty of Pharmacy, Universitas Indonesia, Depok, 16424, Indonesia*

[2] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, 16424, Indonesia*

Corresponding author: arry.yanuar@ui.ac.id

Abstract

Background: Dipeptidyl Peptidase-4 (DPP-4) inhibitors are becoming an essential drug in the treatment of type 2 diabetes mellitus, but some classes of these drugs have side effects such as joint pain that can become severe to pancreatitis. It is thought that these side effects appear related to their inhibition against enzymes DPP-8 and DPP-9.

Objective: This study aims to find DPP-4 inhibitor hit compounds that are selective against the DPP-8 and DPP-9 enzymes. By building a virtual screening workflow using the Quantitative Structure-Activity Relationship (QSAR) method based on artificial intelligence (AI), millions of molecules from the database can be screened for the DPP-4 enzyme target with a faster time compared to other screening methods.

Result: Five regression machine learning algorithms and four classification machine learning algorithms were used to build virtual screening workflows. The algorithm that qualifies for the regression QSAR model was Support Vector regression with R^2_{pred} 0.78, while the classification QSAR model was Random Forest with 92.21% accuracy. The virtual screening results of more than 10 million molecules from the database, obtained 2,716 hit compounds with pIC50 above 7.5. Molecular docking results of several potential hit compounds to the DPP-4, DPP-8 and DPP-9 enzymes, obtained CH0002 hit compound that has a high inhibitory potential against the DPP-4 enzyme and low inhibition of the DPP-8 and DPP-9 enzymes.

Conclusion: This research was able to produce DPP-4 inhibitor hit compounds that are potential to DPP-4 and selective to DPP-8 and DPP-9 enzymes so that they can be further developed in the DPP-4 inhibitors discovery. The resulting virtual screening workflow can be applied to the discovery of hit compounds on other targets.

Keywords:

Artificial Intelligence; DPP-4; KNIME; Machine Learning; QSAR; Virtual Screening

Background

Nowadays, Dipeptidyl Peptidase-4 (DPP-4) (EC 3.4.14.5) inhibitors become an important oral antidiabetic drug for the treatment of type 2 diabetes. Sitagliptin was introduced in 2006 as the first DPP-4 inhibitor agent. Furthermore, this class of drugs is increasingly shifting the role of sulfonylurea in type-2 diabetes treatment in some national and international guidelines.

This class of drugs works differently from most other antidiabetic drugs. The DPP-4 enzyme inhibition, besides stimulating the pancreas to produce and release insulin into the blood, also reduces or normalizes body weight and does not cause hypoglycemia. A different mechanism of action and effects with other antidiabetic groups makes this class of drugs exciting to develop [1–5].

Some of the DPP-4 inhibitors have been circulating in the market. These drugs are well-tolerated, but some of them have side effects such as joint pain and even pancreatitis. These side effects appear to be related to inhibition of enzymes with high sequence homologies, such as DPP-8 and DPP-9 [6, 7]. Therefore, it necessary to discover and develop new DPP-4 inhibitors that are selective against DPP-8 and DPP-9 enzymes.

The discovery of new DPP-4 inhibitors can be made in several ways, such as high throughput screening (HTS), which is generally carried out by the large pharmaceutical industry. Another alternative that can be done is computer-aided drug design (CADD) by conducting virtual screening [8–11] on a vast database containing millions of compounds such as ChEMBL [12], PubChem [13]. One of the virtual screening methods that can be developed using a quantitative structure-activity relationship (QSAR) that can predict the activity of a molecule against the DPP-4 enzyme by establishing a relationship between the molecular structure represented by descriptors or fingerprints with the activity of several research results on the DPP-4 enzyme. The QSAR methods that can be applied were regression QSAR or classification QSAR.

In this study, virtual screening of the database will be done by building a virtual screening workflow [14–18], which uses the QSAR classification models to predict active compounds from the database and then uses the QSAR regression method to see the value of its activity [16]. The progress of QSAR methods can benefit from modern artificial intelligence (AI) approaches to develop computation models [19–22], this study focuses on the development of a QSAR method based on AI with a supervised learning algorithm. With this method, the prediction of thousands or even millions of molecules activity from a database can be made faster than other virtual screening methods [23].

Various AI-based DPP-4 QSAR inhibitor studies have been carried out previously [24–26]. Those studies generally only used certain compound derivatives, and with a small dataset, so the prediction ability was limited only to that derivatives. The principle of QSAR is that similar structures have similar activities [27], so when using a small dataset and only limited to a derivative compound, the predictive ability is limited and will be biased, especially for virtual screening in large databases. Therefore, this study tried to develop a method for predicting the activity of DPP-4 inhibitor compounds, which is more extensive, not only limited to one derivative but also to various types of derivatives and with a dataset that is more extensive than experimental data that has ever existed before. DPP-4 inhibitor screening with a machine learning approach using thousands of DPP-4 inhibitor datasets has been conducted [28], but this method needs to be further developed with higher accuracy above 90%.

The QSAR model that will be developed to build virtual screening workflows was a model that meets QSAR statistical parameters standard [29, 30] for regression models and high accuracy for classification models. The methods used are predictive and reliable for predicting the activity of the new compound. For this purpose, five machine learning regression algorithms were built (i.e., XGboost Tree Ensemble, Random Forest, Support Vector Regression, Deep learning, and Multiple Linear Regression) and four machine learning classification algorithms (i.e., XGboost Tree Ensemble, Random Forest, Support Vector Machine, and Deep learning) [31, 32].

To produce a potential hit compound that has a high inhibitory activity against DPP-4 and low inhibitory activity against DPP-8 and DPP-9, the hit compound from the results of virtual screening is carried out molecular docking to DPP-4, DPP-8, and DPP-9. Related ligands used were compounds that have been available in the market, such as Trelagliptin, Omarigliptin, and Carmegliptin [33–36]. Potential hit compounds have inhibitory activity (Ki), which is not much different or even higher than the comparative ligand in DPP-4 enzymes and low inhibitory activity in other DPP enzymes [6].

Methods

Dataset

The dataset was downloaded from the ChEMBL website with a DPP-4 target, with an IC_{50} activity filter, and the target organism was Homo Sapiens (*https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL284/*). The Dataset containing 4,661 compounds were curated using the nM (nanomolar) activity unit by removing the empty activity value, salt, small fragments. The molecular structures were then normalized, followed by final checking for compound duplication [14, 37]. The remaining 3,933 compounds were then used for the regression modeling dataset.

For the QSAR classification models 4,355 compounds were used from the ChEMBL database with a scientific literature filter. The missing value and duplication of the dataset were curated, the salt and small fragments were removed, the molecular structure was normalized, and the duplication was removed [14, 37], leaving 3,740 compounds. Furthermore, the data are classified into active and inactive compounds that used for modeling, with pIC_{50} activity above 7.5 was active compounds, compounds with pIC_{50} below 6 were inactive compounds, and compounds with pIC_{50} between 7.5 and 6 were grey compounds that removed [28]. The remaining 2307 compounds were used for model development.

Calculation of descriptors and fingerprint

The descriptor and fingerprint calculations were performed with four nodes in KNIME [38, 39], i.e., RDKit Descriptor Calculation and Fingerprint from RDKit, and Fingerprint and Molecular Properties of CDK [40]. The total descriptors and fingerprint of each molecular structure were 17,784 features for the classification model and 19,875 features for the regression model.

Standardization of activities and partitions

The activity of each molecular structure (IC_{50}) is converted into logarithmic values in molar units [41] $pIC_{50} = -\log (IC_{50}*10^{-9})$.

The dataset is randomly partitioned, 80%: 20% and the 80% dataset is partitioned again to 90%: 10% for training set and validation set, 20% of the dataset is used for test sets. [42–44]. The training set was used for the construction of models and validation and test series were used as the internal and external evaluation of the constructed models, respectively [45–47].

Feature selection

The best features were selected from several feature selection methods. A dimension reduction with PCA, height correlation, and random forest [48], and overall features of the RDKit descriptor, ECFP2 fingerprint, MACCS, Pubchem, or a combination of RDKit descriptor features with the MACCS fingerprint were applied. These features were tested on several machine learning models, features with the highest accuracy used as features for modeling.

QSAR modeling

Five machine learning algorithms were used to build the QSAR regression models, i.e., Deep Learning, XGBoost Tree Ensamble, Multiple Linear Regression, Random Forest, and Support Vector Regression. For the QSAR classification models, four machine learning algorithms were used, i.e., Deep Learning, XGBoost Tree Ensamble, Random Forest, and Support Vector Machine.

The partitioned dataset was used to build machine learning models from training, validation to testing (Figure 1). The hyperparameter value of each model was determined through optimization with a random search (a combination of 100 experiments). The hyperparameter that produces the best performance model will be used for internal and external validation.

Evaluation of the QSAR regression model

To see the Goodness-of-fit model performance, a statistical analysis of the regression coefficient (R), determination coefficient (R^2), and Mean Squared Error (MSE) for each machine learning algorithm were performed. For the hyperparameter model selection, based on the lowest Root Mean Square Error (RMSE).

$$R = 1 - \sum_i \frac{(y_i - \hat{y}_i)}{(y_i - \bar{y}_i)} \tag{1}$$

$$R^2 = 1 - \sum_i \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2} \tag{2}$$

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

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

y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}_i$ is the actual average value. The minimum value recommended for the QSAR regression models to produce a reliable prediction and predictive value of the correlation coefficient (R) for in vivo data ≥ 0,8 and the coefficient of determination (R^2) ≥ 0,6 [49].

Internal validation (*$R^2_{(cv)}$*) was done by 8% dataset. Over-fitting of the model is usually suspected when the R^2 value of the training model is significantly higher than 25% than the R$^2_{cv}$ value [49].

Golbraikh and Tropsha, (2002) proposed a set of parameters to determine the external predictability of the QSAR Regression models, i.e. the model is considered satisfactory if all the following conditions are met:

1. $R^2_{(cv)} > 0.5$
2. $R^2_{(ext)} > 0.6$
3. $\frac{R - R_0^2}{R^2} < 0.1$ and $0.85 \leq k \leq 1.15$ or
4. $\frac{R^2 - R'^2_0}{r^2} < 0.1$ and $0.85 \leq k' \leq 1.15$ or
5. $|R_0^2 - R'^2_0| < 0.3$

$R^2_{(cv)}$ and $R^2_{(ext)}$ are the coefficient of determination from internal validation results, and the test data, respectively. R_0^2 and R'^2_0 are the coefficient of determination between actual and predicted activities at zero intercepts, and between predicted and actual, respectively. Furthermore, k and k' is the slope of the regression line through the origin point [29].
Roy, Kar, and Das (2015) proposed $r_m^2{}_{(test)}$ for external validation. $r_m^2{}_{(test)}$ value was calculated using the square of the correlation coefficient between actual and predicted activities from the test dataset. For acceptable predictions, $\overline{r_m^2}$ $_{(test)}$, the value should be lower than 0.2, provided that the value of $\Delta r_m^2{}_{(test)}$ is more than 0.5.

Evaluation of the classification QSAR model

Internal validation and external validation sets are used to test the classification model performance. All models are evaluated with:

Sensitivity = Recall = TP / (TP + FN) .... (5)

Specificity = TN /(TN + FP) .... (6)

Accuracy = (TP + TN) / (TP + FP +TN + FN) .... (7)

F-Measure = 2(Recall x Precision) / (Recall + Precision) .... (8)

Precision (FP Rate) = TP / (TP +FP) .... (9)

Receiver Operating Characteristic (ROC) plot was used to display the graphical behavior of the model. Visualization between the X-axis (1- Specificity) and the Y-axis (Sensitivity) will show the perfect model with the area under the curve (AUC) 1. The AUC of 0.5, will then the classification model has no discriminatory power at all [30, 50].

Testing the QSAR model workflow on other targets

To see the ability of QSAR method workflow automation implementation, several targets from the ChEMBL database (*https://www.ebi.ac.uk/chembl/*) were used as a modeling dataset to see workflow prediction capabilities, i.e., the opioid sigma receptor (CHEMBL287) and the adrenergic Beta-1 receptor (CHEMBL213). Activity data for each target was downloaded from the ChEMBL website based on its ID and analyzed with KNIME. Five machine learning regression algorithm was used to see its ability on other models, i.e., deep learning, XGBoost tree ensemble, multiple linear regression, random forest, and support vector machine.

Virtual screening from database

To find and develop the new DPP-4 inhibitor hit compound, virtual screening was performed from various databases, including 1,870,461 molecules from the ChEMBL [51, 52] (*http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_25/*) in SDF format. 1,424,986 molecules from PubChem [53] database in SDF format (*https://ftp.ncbi.nih.gov/pubchem/Compound/CURRENT-Full/SDF/*) update 18 October 2019), 7,869,542 molecules from Molport database in smiles format (*ftp://molport.com/*) update 18 October 2019. These molecules were tested by the QSAR classification models to determine their activity. The active molecules are further processed for similarity tests with ChEMBL DPP-4 inhibitor data to find out whether this molecule is a new compound or is the same compound as the existing DPP-4 inhibitors. The similarity value used is the Tanimoto coefficient with a similarity limit of 0.85 using a 166-bits MACCS fingerprint [27, 54].

The activity value of the active molecules that meet the criteria was determined with the QSAR regression models. The obtained hit compounds were then tested for similarity with DPP-4 target decoys from DUD-E [55] (*http://dude.docking.org/targets/dpp4*) to see whether that hit compounds were similar to pan assay interference compounds (PAINS) or not. The final stage of the hit compound is the value of the Lipinski's rule of 5 [56] to see whether this hit compound can be developed as a DPP-4 inhibitor (drug-likeness).

Molecular docking of hit compound into DPP-4, DPP-8, and DPP-9 enzymes

The most active of the hit compounds were collected from the virtual screening results. The molecular docking was carried out in the DPP-4 enzyme, while the selectivity is compared to the DPP-8 and DPP-9 enzymes [57]. The crystal structure of the enzyme used was downloaded from the protein data bank (PDB) [58] with criteria derived from the *Homo sapiens* organism. The crystal structure of the DPP-4 enzyme used is bound to various ligands that have been available in the market such as Trelagliptin (PDB id 5KBY) [59], Alogliptin (PDB id 2ONC) [60], Omarigliptin (PDB id 4PNZ) [61], and Carmegliptin (PDB id 3KWF) [35]. For the DPP-8 enzyme, we used macromolecules PDB id 6HP8 [62] and DPP-9 with PDB id 6EOR [63].

The potential hit compound has an inhibitory activity against the DPP-4 enzyme that is higher or equal to the original ligand. In contrast, for selectivity to DPP-8 and DPP-9, the potential hit compound has a low inhibitory activity > 10,000 nM [6]. Molecular docking is done by Autodock version 1.5.6, for visualization using Discovery Studio R2 Client 2017, and PyMol version 2.3.4. Before docking the molecular compound the results of virtual screening were performed by redocking the original ligand and validating its root mean square deviation (RMSD) with PyMOL. In general, a good quality of redocking is RMSD below 2 Å. All redocking value is fall under this value.

Results and Discussion

The diversity of datasets is a critical factor in the QSAR method. Several QSAR methods have been developed to study DPP-4 inhibitors. Some previous research mainly focuses on local (conventional) QSAR modeling studies, both with QSAR-2D, QSAR-3D, and pharmacophore. However, this method only provides reliable predictions is limited chemical space. So, a more extensive and more diverse set of data from the ChEMBL database is used to build the global QSAR method [64].

Chemical space analysis

For the regression model (Figure 2 (A)) the training sets used were 2,831 molecules scattered with the highest test molecular activity value 0.064 nM (pIC_{50} = 10.19) and the lowest molecular activity 9,100 µM (pIC_{50} = 2.04) while 787 molecules as the test set (external validation) with the highest test molecular activity 0.012 nM (pIC_{50} = 10.92) and the lowest molecular activity 1,000 µM (pIC_{50} = 3).

In the classification models (Figure 2 (B)), the training set used was 1,845 molecules consisting of 879 active molecules and 976 inactive molecules. In contrast, the test set (external validation) was 462 molecules with 237 active molecules and 135 inactive molecules. In general, the performance and predictive ability of the model are influenced by chemical compounds diversity in training data and test data. A simple chemical space analysis was done to examine the chemical diversity of the research dataset. Figure 2 is a plot showing the distribution of DPP-4 inhibitor compounds in training data and external validation data based on molecular weight (MW) and partition coefficient (XLogP). In this plot, a similar distribution of 784 external validation data to 2,841 training data was observed. Also in the chemical space with MW ranging from 128.094 to 1,173.69 and XLogP from -4.107 to 18.493 illustrates that the modeling data has a significant heterogeneity of chemical spaces so that the prediction ability becomes wider (global) to predict the new compounds pIC_{50} value from various compounds derivative against the DPP-4 enzyme [65].

Feature selection

The feature selection method developed for seven types of features (Figure 3), produces the best features in reducing feature selection method with Random Forest. The combination of this feature with SVR consistently produces the highest R^2 and lowest error at the internal and external validation stage compared to EFCP2 features. The Random Forest feature selection technique produces a deep reduction level, from 17,569 features reduced to 98.82% features. Dimension reduction with random forest, developed from the methods of Silipo, Adae, Hart, & Berthold (2014), although the reduction rate is high, the accuracy is still maintained. By using the Random Forest Regression, the feature was reduced by the parameters nModels 100 and depth level 10 and k-number of iteration 100. The feature selection results by reduction method with Random Forest come from a combination of descriptors and fingerprints from RDKit and fingerprints from CDK. Two hundred eight features for the QSAR regression model and 200 features for the QSAR classification model.

Optimization of machine learning algorithms

The model with the lowest error and highest accuracy were obtained from algorithm parameter optimization through the lowest random error search method in a specific parameter range with 100 parameter values and 100 repetitions. The deep learning model was built using two dense layers with a learning rate of 0.01, and the number of neurons is 100. The weight initiation strategy used was Sigmoid_Uniform, activation function: HardSigmoid, updater: RMSPROP. The parameter optimization results show the number of batches and epochs with the lowest error RMSE 0.8335 with the parameter values of Batches 65 and Epoch 505. For XGBoost parameters that are optimized boosting rounds, maximum depth, and eta, the lowest error value was RMSE 0.7684 with boosting rounds parameter values 1000 Maximum depth 13 and eta 0.2994. The optimum value of Multiple Linear Regression was obtained from the optimization of the offset value. The lowest error value RMSE was 0.9605, with the Offset value parameter 0.23707. In Random Forest, the optimum parameters set are a number of methods and tree depth, and it obtained the lowest error value was RMSE 0.7765 with a number of models 109 and tree depth 15. In Support Vector Regression, the lowest error parameter obtained using a local radial base function (RBF), the lowest error value obtained is RMSE 0.7573, with cost parameters 79 and degree 5. Other parameters that are not specified are in default settings in KNIME.

Optimization of classification methods is the same as the regression models, but in the deep learning optimal parameters on the dense layer obtained by the Relu Weight Initiation Strategy and LeakyRelu Activation Function, Adagrad Updater, batch Size 15 and Epoch 843. The highest DL accuracy was 0.9431. XGBoost performed the highest accuracy with 0.9404 at nRounds 500 and maxDepth 12. In Random Forest, the highest accuracy was 0.9404 in nMethod 216 and treeDepth 21. In SVM, the highest accuracy was 0.6938 with sigma 0.9662 and penalty 25.

The best model of optimization results with the lowest error was the SVR. This model widely applied in various QSAR methods analyses, including the DPP-4 inhibitor QSAR model conducted by Gu et al. (2013) and Yang et al. (2013). This method also widely used in managing high-dimensional variables, especially with small datasets, by mapping and transforming non-linear data kernels into high-dimensional features. Random Forest also widely applied to the QSAR method. While XGBoost still has little application in QSAR modeling because this method is a relatively new developed method compared to other machine learning methods, however, XGBoost has a faster analysis capability, with satisfactory predictions.

The optimization results on various classification methods such as DL, XGBoost, and Random Forest showed satisfactory results with an accuracy of up to 94%. However, in SVM, the optimum accuracy was only 69%. Low SVM accuracy showed that hyperplane was not able to separate active and inactive compounds in the classification model dataset properly. However, ensemble-based models such as Random Forest and XGboost were able to predict active and inactive compounds well. The results give a contrasts with the Regression QSAR model in which SVM was the best prediction method. In the regression method, the transformation of the variable to higher dimensions results in a better prediction of pIC_{50} compared to the trees method.

In the regression model, goodness-of-fit of MLR and deep learning were very low when compared to the SVR and ensemble methods. That is because MLR itself is intended for linear data, so when handling nonlinear data the performance becomes low. At the same time, in deep learning, this is likely due to the lack of training data so that the model performance was not optimal. Generally, a small dataset on deep learning shows performance under conventional machine learning methods, but in a large dataset, the ability of deep learning model remains consistent. These results showed contrary to the machine learning method, which optimal on datasets [66].

Internal validation

From the internal validation results of the regression QSAR model, the comparison between the model training results error and the internal validation is no more than 25% for all models. It means that the resulting QSAR model did not experience overfitting [49]. The internal validation model with the lowest error (MSE) (Figure 4) was the SVR model, then XGBoost Tree Ensemble and Random Forest.

From the results of internal validation on the QSAR classification models (Table 1), three algorithms, i.e., Deep Learning, XGBoost, and Random Forest, showed excellent performance with accuracy above 90%. Of all the QSAR classification models, the model with the best performance was Random Forest.

External validation

The ideal QSAR regression model with high predictability has intercept values 0 and slope 1 in the equation of the line between actual and predictions, and the regression coefficient is equal to 1. Seen from the k value (slope) between actual and prediction and k' between prediction and actual without an interception, the regression line approaches 1 [29]. All the resulting QSAR regression models had met the requirements for the parameters k and k'. Other than that, the predictive model must have proximity between R^2, R_0^2, and R'^2_0 values with the condition < 0.1. The method developed in the test with external validation data, i.e., DL, MLR, and RF method, has a difference with R_0^2 value more than 0.1 so that these models did not meet the requirements to become predictive QSAR models.

To verify the closeness between the observed and predicted data, the minimum value of the parameter $\overline{R_m^2}$ and ΔR_m^2 must be met [30]. From the calculation results (Table 2), DL with MLR does not meet the $\overline{R_m^2}$ requirement because the value was smaller than 0.5, while other models XGBoost, RF, and SVR were eligible. Calculation results for the parameter ΔR_m^2 showed only the SVR model that eligible for the QSAR regression model.

The SVR model fulfills all the requirements for the QSAR regression model on the dataset that used, both goodness of fit, which is proven by R^2 training values above 0.5. Robustness is proven by Q^2 values above 0.5 and predictive methods, which are proven by the fulfillment of all statistical parameters of the QSAR method. So the SVR model can be used to activity predictions model in virtual screening workflows while the other methods fail to meet the requirements, especially in parameter values (R^2-R_0^2)/R^2 and ΔR_m^2. The SVR model showed that there is a significant difference between the curves formed by the predictive value and the actual point of origin (zero intercepts). The perfect method will produce a regression line that lies in the regression coefficient one and intercept 0. The further the slope between the plot formed from predictions results of the actual at zero intercepts, the predictive ability of the method will decrease because it will further increase the residual value between the actual data and the predicted results.

External validation results of the classification method in table 3 and the ROC curve (Figure 5) showed three algorithms, Deep Learning, XGBoost Tree Ensemble, and Random Forest were the classification models with the excellent predictive ability (higher than previous studies with an accuracy above 80% [28]). Deep learning and XGBoost algorithms showed better performance on external validation and ROC curves. However, to build a virtual screening workflow, the Random Forest algorithm was used. This algorithm returned ROC curves with the highest accuracy compared to other algorithms, using internal and external validation. The value is not much different from deep learning and XGBoost algorithms.

Testing QSAR regression workflow on other targets

The results of the QSAR model's prediction performance test against several targets (Table 4), produced predictions with a high enough coefficient of determination reaching 0.7 in the SVR, XGBoost, and Random Forest models. This prediction results showed the ability to predict our workflows against other targets automatically from raw datasets that are directly downloaded from the ChEMBL database.

Virtual screening results

Virtual screening results of the ChEMBL, PubChem, and Molport databases obtained several hit compounds (Figure 6). This hit compound was tested for similarity with the DPP-4 ChEMBL database, with similarity values below 0.85. Furthermore, it was also tested on DPP-4 DUD-E decoys. Finally, the compound must meet Lipinski's rule of five to be absorbed into the body. As a result, about 2,716 compounds can pass through this phase.

Molecular docking results

Molecules that show potential with high inhibitory activity (Ki <100 nM) in the crystal structure of the DPP-4 enzyme include hit compounds with id CH0001, CH0002, and CH0003. These hit compounds come from the ChEMBL database; some of these compounds even have higher inhibitory activities than the original ligand. For example, CH0001 and CH0002 hit compounds showed higher inhibitory activity compared to trelagliptin ligands in DPP-4 (5KBY) crystal structure. The CH0003 hit compound has a higher inhibitory activity than the alogliptin ligand in DPP-4 (2ONC) crystal structure.

Molecular docking results on the DPP-8 enzyme, CH0003 hit compound showed high inhibitory activity (Ki <100 nM) while CH0001, CH002 hit compounds showed moderate inhibitory activity. Some ligands such as omarigliptin and trelagliptin showed moderate inhibitory activity against the DPP-8 enzyme. While the molecular docking results on the DPP-9 enzyme crystal structure, CH0003 hit compounds showed high inhibitory activity, while the CH0002 hit compound showed moderate inhibitory activity. The tetheragliptin, alogliptin, omarigliptin, and carmegliptin exhibits moderate inhibitory activity.

From molecular docking results of DPP-4, DPP-8, and DPP-9 enzymes (Table 5), resulting in the potential and selective hit compound was CH0002. The CH0003 hit compound, on the molecular docking visualization, showed unfavorable donor-donor interactions. This interaction occurs because the explicit hydrogen of the hit compound is in the range of hydrogen bonding with hydrogen atoms from amino acids. For example (Figure 7), the interaction of H atoms in the hydroxy group CH0003 hit compounds with the amino acid Arginine 125 DPP-4 (2ONC).

The CH0002 hit compound based on visualization results did not fully meet the eight potential and selective ligand criteria proposed by Ojeda-montes et al. [35] because there was an interaction with the amino acid ligand Trp629 on DPP-4 (5KBY), which could reduce its activity. However, the bonding of the CH0002 hit compound to the DPP-4 enzyme (4PNZ) has a bond with the active site S3 with the amino acids Phe357 and Arg358 which are essential for selectivity (Figure 8) so that this hit compound deserves to be the best hit compound from this virtual screening results.

Conclusion

It can be concluded that the QSAR method based on artificial intelligence developed both the regression method and the classification method, produces a virtual screening workflow that meets the qsar statistical parameter standards. In the regression model, this is evidenced by the fulfillment of all statistical parameters proposed by Golbraikh & Tropsha, (2002) and Roy & Kur, (2016). The algorithm that satisfies the QSAR statistical parameters was Support Vector Regression (SVR). In the classification model, the accuracy obtained was > 90%, and the highest ROC curve obtained was 0.96, far above the standard > 0.5. The best algorithm in the classification method is Random Forest.

The best machine learning algorithm of classification and regression models was used to build virtual screening workflow. Random Forest was used to predicting active compounds and SVR was used to predict its activity. From the results of MolPort, ChEMBL and Pubchem database screening with a total of more than 10 million compounds, hit compounds with pIC50 activity above 7.5 were 2,716 compounds.

Molecular docking results several hit compounds carried out to DPP-4, DPP-8 and DPP-9 enzymes, hit compounds that produce high inhibitory activity against DPP-4 enzymes and low inhibition of DPP-8 and DPP-9 enzymes ware hit compounds CH0002. this molecule can be further developed as a DPP-4 inhibitor.

Abbreviations

QSAR	= Quantitative Structure–Activity Relationship;
DL	= Deep learning;
XGBoost	= XGBboost tree ensemble;
MLR	= Multiple Linear Regression
RF	= Random Forest;
SVR	= Support Vector Regression.
PCA	= Principle Component Analysis
AUC	= Area Under Curve

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of data and material

Data are available at Hermansyah, Oky; Bustamam, Alhadi; Yanuar, Arry (2019), "Dataset for QSAR Modeling of DPP-4 Inhibitors", Mendeley Data, v2. http://dx.doi.org/10.17632/4sw5hr2yz7.2

Competing interests

We declare that we have no significant competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

Funding

This work was supported by the fund through PITTA (International Indexed Publication for UI Student Final Project) 2019 grant (NKB-0472/UN2.R3.1/HKP.05.00/2019) awarded to A.Y. from the Directorate of Research and Community Service at the University of Indonesia for publication manuscript.

Authors' contributions

OH designed the study, implemented all KNIME workflow development and prepared the manuscript. AB gave helpful information for design of the study. AY supervised and provided input on all aspects of the study. OH, AB and AY discussed the results. OH and AY analyzed data and wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We thank to Prof. Heru Suhartanto (Faculty of Computer Sciences, Universitas Indonesia) for the advice and discussion of a good machine learning model.

References

1. Gallwitz B. Clinical Use of DPP-4 Inhibitors. Frontiers in endocrinology. 2019;10:389. doi:10.3389/fendo.2019.00389.
2. Sesti G, Avogaro A, Belcastro S, Bonora BM, Croci M, Daniele G, et al. Ten years of experience with DPP-4 inhibitors for the treatment of type 2 diabetes mellitus. Acta Diabetologica. 2019;56:605–17. doi:10.1007/s00592-018-1271-3.
3. Alam* F, Islam MA, Gan* MAK and SH. Updates on Managing Type 2 Diabetes Mellitus with Natural Products: Towards Antidiabetic Drug Development. Current Medicinal Chemistry. 2018;25:5395–431. doi:http://dx.doi.org/10.2174/0929867323666160813222436.
4. Chylewska* A, Biedulska M, Makowski* PS and M. Metallopharmaceuticals in Therapy - A New Horizon for Scientific Research. Current Medicinal Chemistry. 2018;25:1729–91.

doi:http://dx.doi.org/10.2174/0929867325666171206102501.

5. Popovic-Djordjevic* JB, Stanojkovic IIJ and TP. Antidiabetics: Structural Diversity of Molecules with a Common Aim. Current Medicinal Chemistry. 2018;25:2140–65. doi:http://dx.doi.org/10.2174/0929867325666171205145309.

6. Huan Y, Jiang Q, Liu J, Shen Z. Establishment of a dipeptidyl peptidases (DPP) 8/9 expressing cell model for evaluating the selectivity of DPP4 inhibitors. Journal of Pharmacological and Toxicological Methods. 2015;71:8–12. doi:https://doi.org/10.1016/j.vascn.2014.11.002.

7. Patel BD, Ghate MD. Recent approaches to medicinal chemistry and therapeutic potential of dipeptidyl peptidase-4 (DPP-4) inhibitors. European Journal of Medicinal Chemistry. 2014;74:574–605. doi:https://doi.org/10.1016/j.ejmech.2013.12.038.

8. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. British journal of pharmacology. 2011;162:1239–49. doi:10.1111/j.1476-5381.2010.01127.x.

9. Pei L, Shen X, Yan Y, Tan* C, Qu K, Zou J, et al. Virtual Screening of the Multi-pathway and Multi-gene Regulatory Molecular Mechanism of Dachengqi Decoction in the Treatment of Stroke Based on Network Pharmacology. Combinatorial Chemistry & High Throughput Screening. 2020;23:1–13. doi:http://dx.doi.org/10.2174/1386207323666200311113747.

10. Wang Z-F, Hu Y-Q, Zhang Q-GW and R. Virtual Screening of Potential Anti-fatigue Mechanism of Polygonati Rhizoma Based on Network Pharmacology. Combinatorial Chemistry & High Throughput Screening. 2019;22:612–24. doi:http://dx.doi.org/10.2174/1386207322666191106110615.

11. Shamsara* J. A Random Forest Model to Predict the Activity of a Large Set of Soluble Epoxide Hydrolase Inhibitors Solely Based on a Set of Simple Fragmental Descriptors. Combinatorial Chemistry & High Throughput Screening. 2019;22:555–69. doi:http://dx.doi.org/10.2174/1386207322666191016110232.

12. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research. 2012;40 Database issue:D1100–7. doi:10.1093/nar/gkr777.

13. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic acids research. 2016;44:D1202–13. doi:10.1093/nar/gkv951.

14. Kausar S, Falcao AO. An automated framework for QSAR model building. Journal of cheminformatics. 2018;10:1. doi:10.1186/s13321-017-0256-5.

15. P. Mazanetz M, J. Marmon R, B. T. Reisser C, Morao I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. Current Topics in Medicinal Chemistry. 2012;12:1965–79. doi:10.2174/156802612804910331.

16. Arunachalam MR and M. Quantitative Structure-Activity Relationship (QSAR) Studies for the Inhibition of MAOs. Combinatorial Chemistry & High Throughput Screening. 2020;23:1–11. doi:http://dx.doi.org/10.2174/1386207323666200324173231.

17. Wójcikowski M, Siedlecki P, Ballester PJ. Building Machine-Learning Scoring Functions for Structure-Based Prediction of Intermolecular Binding Affinity BT - Docking Screens for Drug Discovery. In: de Azevedo Jr. WF, editor. New York, NY: Springer New York; 2019. p. 1–12. doi:10.1007/978-1-4939-9752-7_1.

18. Santos LHS, Ferreira RS, Caffarena ER. Integrating Molecular Docking and Molecular Dynamics Simulations BT - Docking Screens for Drug Discovery. In: de Azevedo Jr. WF, editor. New York, NY: Springer New York; 2019. p. 13–34. doi:10.1007/978-1-4939-9752-7_2.

19. Kumar* R, Sharma A, Tiwari MHS and RK. Prediction of Drug-Plasma Protein Binding Using Artificial Intelligence Based Algorithms. Combinatorial Chemistry & High Throughput Screening. 2018;21:57–64. doi:http://dx.doi.org/10.2174/1386207321666171218121557.

20. Bitencourt-Ferreira G, Jr.* AD da S and WF de A. Application of Machine Learning Techniques to Predict Binding Affinity for Drug Targets. A Study of Cyclin-Dependent Kinase 2. Current Medicinal Chemistry. 2019;26:1–11. doi:http://dx.doi.org/10.2174/2213275912666191102162959.

21. da Silva AD, Bitencourt-Ferreira G, de Azevedo Jr WF. Taba: A Tool to Analyze the Binding Affinity. Journal of Computational Chemistry. 2020;41:69–73. doi:10.1002/jcc.26048.

22. Bitencourt-Ferreira G, de Azevedo WF. Machine Learning to Predict Binding Affinity BT - Docking Screens for Drug Discovery. In: de Azevedo Jr. WF, editor. New York, NY: Springer New York; 2019. p. 251–73. doi:10.1007/978-1-4939-9752-7_16.

23. Neves BJ, Braga RC, Melo-Filho CC, Moreira-Filho JT, Muratov EN, Andrade CH. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. Frontiers in pharmacology. 2018;9:1275. doi:10.3389/fphar.2018.01275.

24. Gu T, Yang X, Li M, Wu M, Su Q, Lu W, et al. Predicting the DPP-IV inhibitory activity pIC_{50} based on their physicochemical properties. BioMed research international. 2013;2013:798743. doi:10.1155/2013/798743.

25. Sokolović D, Ranković J, Stanković V, Stefanović R, Karaleić S, Mekić B, et al. QSAR study of dipeptidyl peptidase-4 inhibitors based on the Monte Carlo method. Medicinal Chemistry Research. 2017;26:796–804. doi:10.1007/s00044-017-1792-2.

26. Al-Fakih AM, Algamal ZY, Lee MH, Aziz M, Ali HTM. A QSAR model for predicting antidiabetic activity of dipeptidyl peptidase-IV inhibitors by enhanced binary gravitational search algorithm. SAR and QSAR in Environmental Research. 2019;30:403–16. doi:10.1080/1062936X.2019.1607899.

27. Martin YC, Kofron JL, Traphagen LM. Do Structurally Similar Molecules Have Similar Biological Activity? Journal of Medicinal Chemistry. 2002;45:4350–8. doi:10.1021/jm020155c.

28. Cai J, Li C, Liu Z, Du J, Ye J, Gu Q, et al. Predicting DPP-IV inhibitors with machine learning approaches. Journal of Computer-Aided Molecular Design. 2017;31:393–402. doi:10.1007/s10822-017-0009-6.

29. Golbraikh A, Tropsha A. Beware of q2! Journal of Molecular Graphics and Modelling. 2002;20:269–76. doi:https://doi.org/10.1016/S1093-3263(01)00123-1.

30. Roy K, Kar S, Das RN. Statistical Methods in QSAR/QSPR BT - A Primer on QSAR/QSPR Modeling: Fundamental Concepts. In: Roy K, Kar S, Das RN, editors. Cham: Springer International Publishing; 2015. p. 37–59. doi:10.1007/978-3-319-17281-1_2.

31. Babajide Mustapha I, Saeed F. Bioactive Molecule Prediction Using Extreme Gradient Boosting. Molecules (Basel, Switzerland). 2016;21:983. doi:10.3390/molecules21080983.

32. Roy K, Kar S, Das RN. QSAR/QSPR Methods BT - A Primer on QSAR/QSPR Modeling: Fundamental Concepts. In: Roy K, Kar S, Das RN, editors. Cham: Springer International Publishing; 2015. p. 61–103. doi:10.1007/978-3-319-17281-1_3.

33. McKeage K. Trelagliptin: First Global Approval. Drugs. 2015;75:1161–4. doi:10.1007/s40265-015-0431-9.

34. Burness CB. Omarigliptin: First Global Approval. Drugs. 2015;75:1947–52. doi:10.1007/s40265-015-0493-8.

35. Mattei P, Boehringer M, Di Giorgio P, Fischer H, Hennig M, Huwyler J, et al. Discovery of carmegliptin: A potent and long-acting dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes. Bioorganic & Medicinal Chemistry Letters. 2010;20:1109–13. doi:10.1016/j.bmcl.2009.12.024.

36. Makrilakis K. The Role of DPP-4 Inhibitors in the Treatment Algorithm of Type 2 Diabetes Mellitus: When to Select, What to Expect. International Journal of

Environmental Research and Public Health. 2019;16:2720. doi:10.3390/ijerph16152720.

37. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? Journal of medicinal chemistry. 2014;57:4977–5010. doi:10.1021/jm4004285.

38. Berthold MR, Cebron N, Dill F, Gabriel TR, Kotter T, Meinl T, et al. Data Analysis, Machine Learning and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. doi:10.1007/978-3-540-78246-9.

39. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME - the Konstanz information miner. ACM SIGKDD Explorations Newsletter. 2009;11:26. doi:10.1145/1656274.1656280.

40. Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C. KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics. 2013;14:257. doi:10.1186/1471-2105-14-257.

41. Selvaraj C, Tripathi S, Reddy K, Singh SK. Tool development for Prediction of pIC50 values from the IC50 values-A pIC50 value calculator. 2011. https://www.semanticscholar.org/paper/Tool-development-for-Prediction-of-pIC50-values-the-Selvaraj-Tripathi/1e679814fffa1a71fadbed952305db1528cdec21.

42. Ripley BD. Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press; 1996. doi: 10.1017/CBO9780511812651.

43. Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach. MIT Press; 2001.

44. Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M, Hu J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Computational Materials Science. 2020;171:109203. doi:https://doi.org/10.1016/j.commatsci.2019.109203.

45. Mozafari Z, Arab Chamjangali M, Arashi M. Combination of least absolute shrinkage and selection operator with Bayesian Regularization artificial neural network (LASSO-BR-ANN) for QSAR studies using functional group and molecular docking mixed descriptors. Chemometrics and Intelligent Laboratory Systems. 2020;200:103998. doi:https://doi.org/10.1016/j.chemolab.2020.103998.

46. Liu W, Lu H, Cao C, Jiao Y, Chen G. An Improved Quantitative Structure Property Relationship Model for Predicting Thermal Conductivity of Liquid Aliphatic Alcohols. Journal of Chemical & Engineering Data. 2018;63:4735–40. doi:10.1021/acs.jced.8b00764.

47. Myint K-Z, Wang L, Tong Q, Xie X-Q. Molecular Fingerprint-Based Artificial Neural

Networks QSAR for Ligand Biological Activity Predictions. Molecular Pharmaceutics. 2012;9:2912–23. doi:10.1021/mp300237z.

48. Silipo R, Adae I, Hart A, Berthold M. Seven Techniques for Dimensionality Reduction: Missing Values, Low Variance Filter, High Correlation Filter, PCA, Random Forests, Backward Feature Elimination, and Forward Feature Construction. Knime. 2014;:1–21. https://www.semanticscholar.org/paper/Seven-Techniques-for-Dimensionality-Reduction-%2C-Low-Silipo/594c5de1f7c466756b405174a991eab34c1e3e66.

49. Veerasamy R, Rajak H, Jain A, Sivadasan S, Christapher PV, Agrawal RK. Validation of QSAR Models - Strategies and Importance. 2011. https://www.semanticscholar.org/paper/Validation-of-QSAR-Models-Strategies-and-Importance-Veerasamy-Rajak/4eb25ff5a87f2fd6789c5b9954eddddfd1c59dab.

50. Gramatica P. On the Development and Validation of QSAR Models BT - Computational Toxicology: Volume II. In: Reisfeld B, Mayeno AN, editors. Totowa, NJ: Humana Press; 2013. p. 499–526. doi:10.1007/978-1-62703-059-5_21.

51. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic acids research. 2017;45:D945–54. doi:10.1093/nar/gkw1074.

52. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic acids research. 2015;43:W612–20. doi:10.1093/nar/gkv352.

53. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Research. 2018;47:D1102–9. doi:10.1093/nar/gky1033.

54. Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discovery Today. 2016;21:1291–302. doi:https://doi.org/10.1016/j.drudis.2016.06.013.

55. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. Journal of Medicinal Chemistry. 2012;55:6582–94. doi:10.1021/jm300687e.

56. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3. Advanced Drug Delivery Reviews. 2001;46:3–26. doi:https://doi.org/10.1016/S0169-409X(00)00129-0.

57. Kang NS, Ahn JH, Kim SS, Chae CH, Yoo S-E. Docking-based 3D-QSAR study for selectivity of DPP4, DPP8, and DPP9 inhibitors. Bioorganic & Medicinal Chemistry Letters. 2007;17:3716–21. doi:https://doi.org/10.1016/j.bmcl.2007.04.031.

58. Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000;28:235–42. doi:10.1093/nar/28.1.235.

59. Grimshaw CE, Jennings A, Kamran R, Ueno H, Nishigaki N, Kosaka T, et al. Trelagliptin (SYR-472, Zafatek), Novel Once-Weekly Treatment for Type 2 Diabetes, Inhibits Dipeptidyl Peptidase-4 (DPP-4) via a Non-Covalent Mechanism. PLOS ONE. 2016;11:e0157509. doi:10.1371/journal.pone.0157509.

60. Feng J, Zhang Z, Wallace MB, Stafford JA, Kaldor SW, Kassel DB, et al. Discovery of Alogliptin: A Potent, Selective, Bioavailable, and Efficacious Inhibitor of Dipeptidyl Peptidase IV †. Journal of Medicinal Chemistry. 2007;50:2297–300. doi:10.1021/jm070104l.

61. Biftu T, Sinha-Roy R, Chen P, Qian X, Feng D, Kuethe JT, et al. Omarigliptin (MK-3102): A Novel Long-Acting DPP-4 Inhibitor for Once-Weekly Treatment of Type 2 Diabetes. Journal of Medicinal Chemistry. 2014;57:3205–12. doi:10.1021/jm401992e.

62. Ross BH. Improvement of Protein Crystal Diffraction Using Post-Crystallization Methods: Infrared Laser Radiation Controls Crystal Order. Thesis. 2019. doi:10.2210/PDB6HP8/PDB.

63. Ross B, Krapp S, Augustin M, Kierfersauer R, Arciniega M, Geiss-Friedlander R, et al. Structures and mechanism of dipeptidyl peptidases 8 and 9, important players in cellular homeostasis and cancer. Proceedings of the National Academy of Sciences. 2018;115:E1437–45. doi:10.1073/pnas.1717565115.

64. Shi J, Zhao G, Wei Y. Computational QSAR model combined molecular descriptors and fingerprints to predict HDAC1 inhibitors. Med Sci (Paris). 2018;34:52–8. https://doi.org/10.1051/medsci/201834f110.

65. Kong Y, Yan A. QSAR models for predicting the bioactivity of Polo-like Kinase 1 inhibitors. Chemometrics and Intelligent Laboratory Systems. 2017;167:214–25. doi:10.1016/j.chemolab.2017.06.011.

66. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chemical Reviews. 2019. doi:10.1021/acs.chemrev.8b00728.

Figure 1:

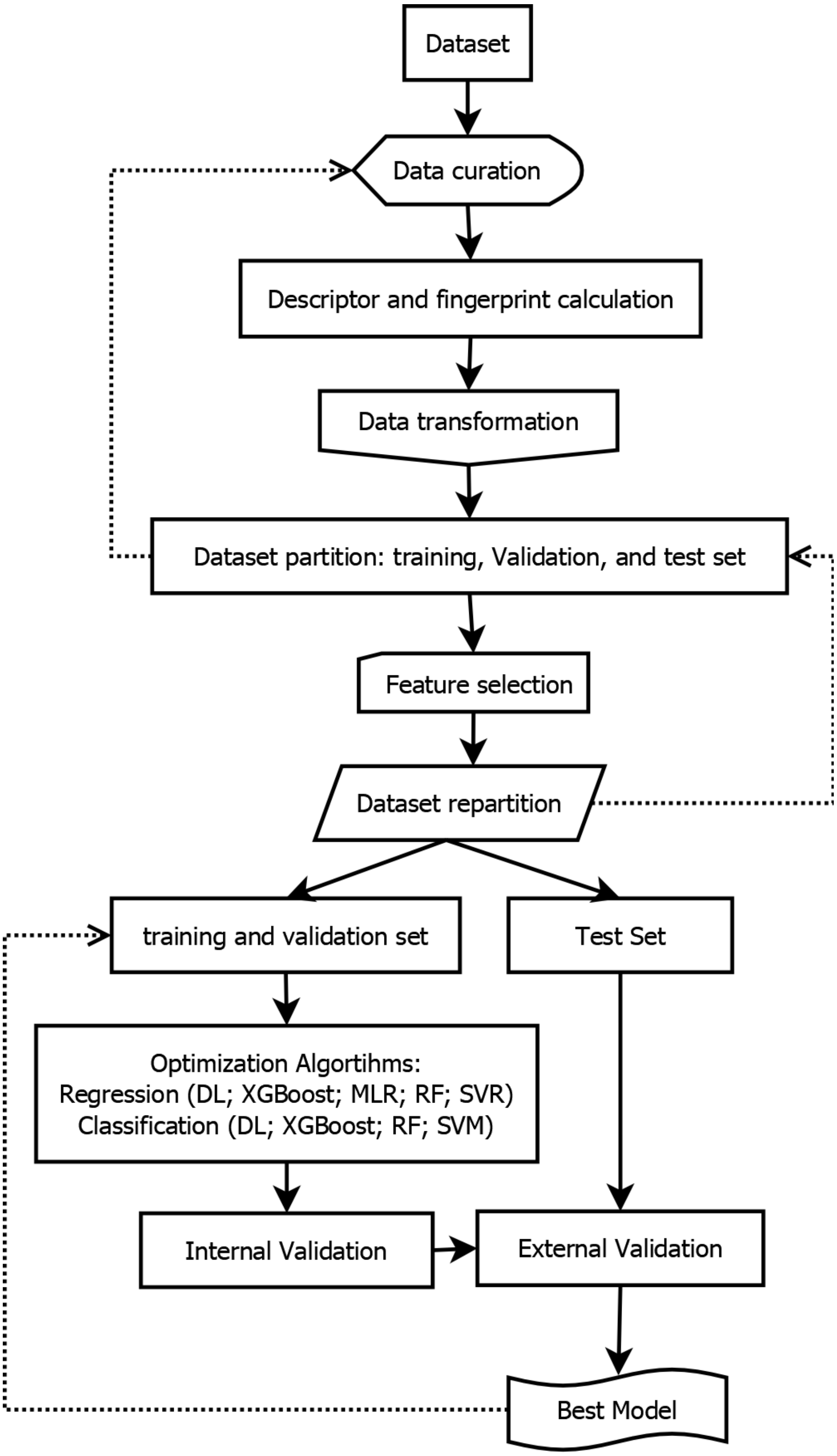
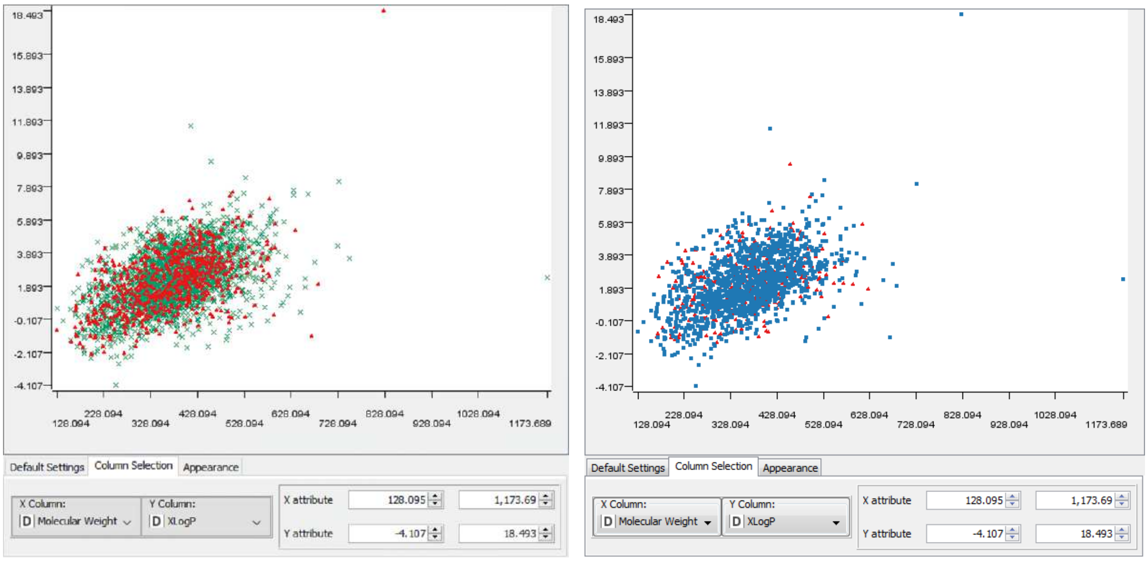

Dataset
Data curation
Descriptor and fingerprint calculation
Data transformation
Dataset partition: training, Validation, and test set
Feature selection
Dataset repartition
training and validation set
Test Set
Optimization Algortihms:
Regression (DL; XGBoost; MLR; RF; SVR)
Classification (DL; XGBoost; RF; SVM)
Internal Validation
External Validation
Best Model

Figure 1. The workflow of QSAR modeling DPP 4 inhibitors

Figure 2:



X Column: D Molecular Weight

Y Column: D XLogP

X attribute 128.095 1,173.69

Y attribute -4.107 18.493

(A) (B)

Figure 2. Chemical space training set versus test set (external validation) defined by MW and ALogP (A) For the regression model (green (x) is a training set, red (Δ) is a test set), and (B) for classification model (blue (□) is a training set, red (Δ) is a test set.

Figure 3:

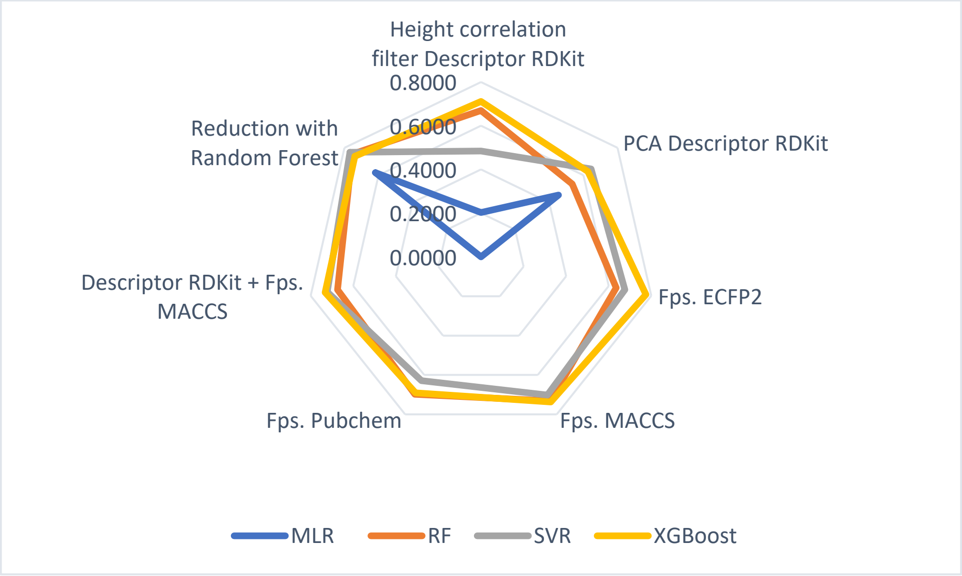


Figure 3. Feature selection.

There are seven features developed to get the best method, using four learning models.

Figure 4:

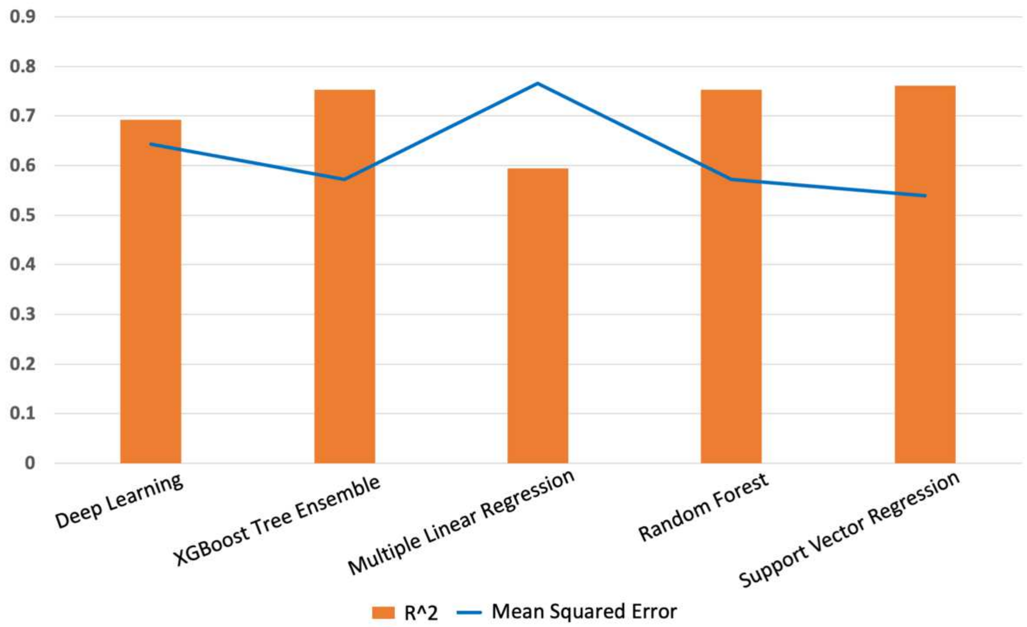


Figure 4. Internal validation results in the regression model

The SVR model produces the best performance among other models with the lowest MSE

Figure 5:

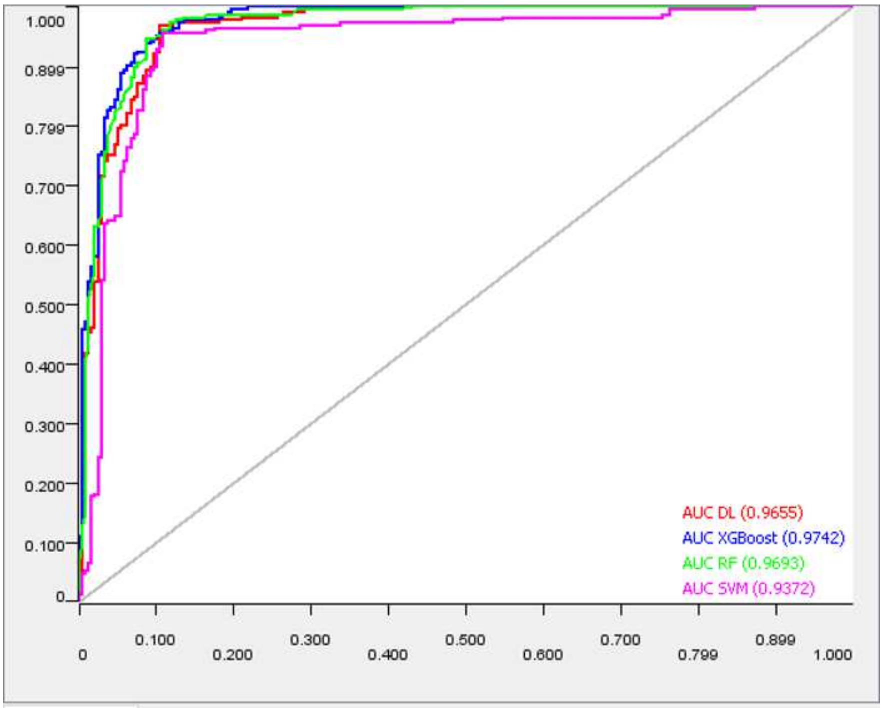
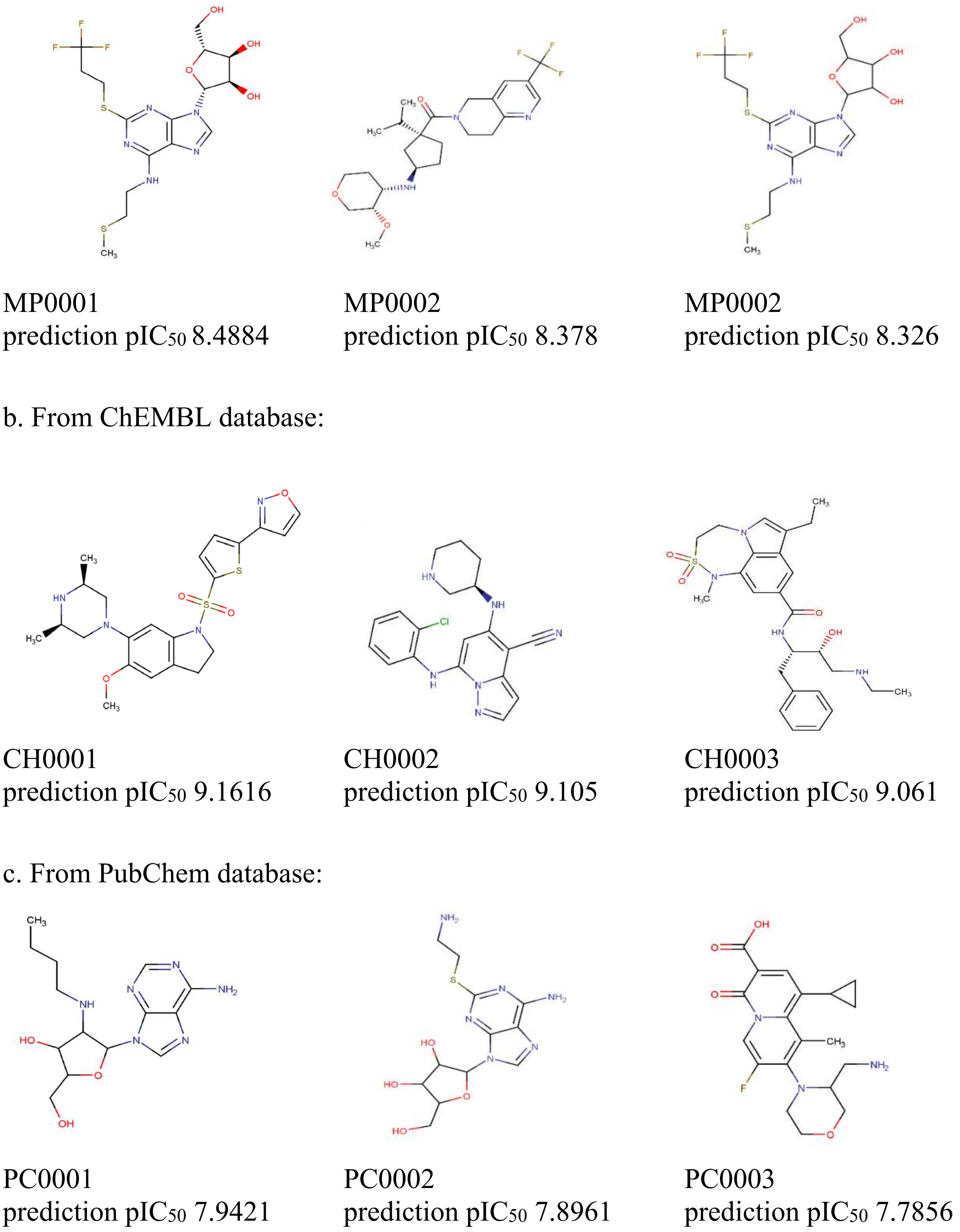


Figure 5. The ROC curve of the four classification models that developed

Figure 6:

a. From MolPort database:



MP0001
prediction pIC50 8.4884

MP0002
prediction pIC50 8.378

MP0002
prediction pIC50 8.326

b. From ChEMBL database:

CH0001
prediction pIC50 9.1616

CH0002
prediction pIC50 9.105

CH0003
prediction pIC50 9.061

c. From PubChem database:

PC0001
prediction pIC50 7.9421

PC0002
prediction pIC50 7.8961

PC0003
prediction pIC50 7.7856

Figure 6. Molecular structure of several hit compounds that was resulting from virtual screening

Figure 7:

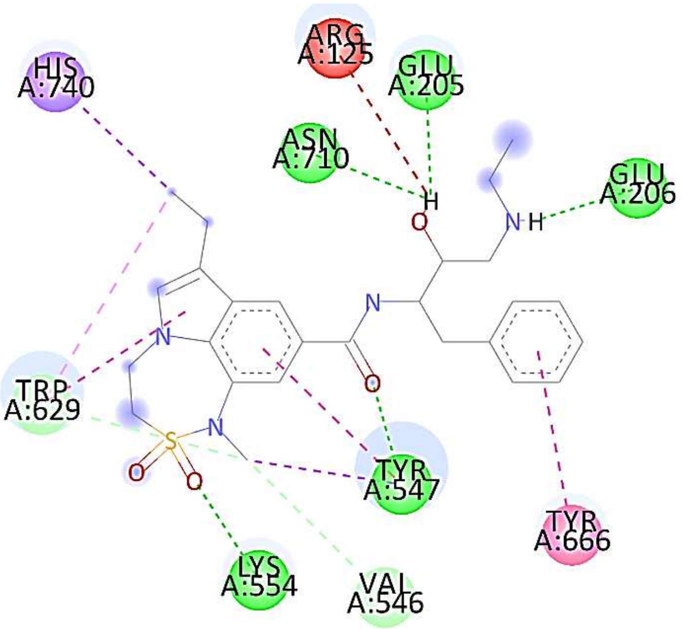


Figure 7. Unfavorable donor-donors interactions (in red) of hit compound CH0003 with DPP-4 (2ONC) crystal structure

Figure 8:

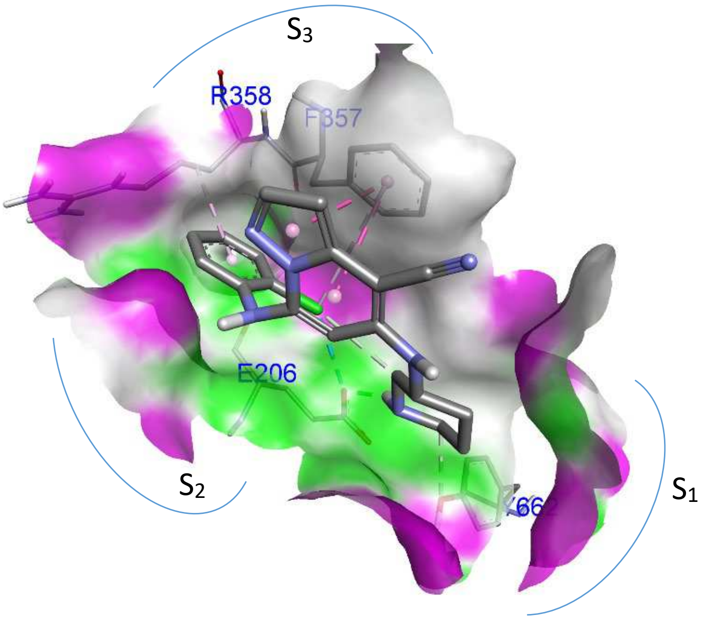


Figure 8. Visualization of the CH0002 hit molecule on DPP-4 (4PNZ).

Table 1:

Table 1. Internal validation results on the classification model

Models	TP	FP	TN	FN	Sensitivity	Specificity	F-Measure	Precision	Accuracy
Deep Learning	891	93	786	75	0.9224	0.8942	0.9138	0.9055	0.9079
Random Forest	925	105	774	41	0.9576	0.8805	0.9269	0.8981	0.9209
SVM	952	457	422	14	0.9855	0.4801	0.8017	0.6757	0.7447
XGBoost	907	93	786	59	0.9389	0.8942	0.9227	0.9070	0.9176

*TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative

Table 2:

Table 2. External validation statistical parameters of various models

Metric	DL	XGBoost	MLR	RF	SVR	Standard
Q^2	0.6920	0.7530	0.5939	0.7532	0.7607	> 0.5 [a]
$R^2_{(pred)}$	0.5910	0.7617	0.6013	0.7668	0.7761	> 0.6 [a]
MSE (validation)	0.7686	0.6163	1.0134	0.6157	0.5971	-
MSE (ext)	1.0370	0.6043	1.0109	0.5914	0.5679	-
R_0^2	0.2564	0.6914	0.4597	0.6319	0.7281	-
R'^2_0	0.5911	0.7618	0.6014	0.7668	0.7762	-
$(R^2 - R_0^2) / R^2$	0.5672	0.0924	0.2422	0.1847	0.0624	< 0.1 [a]
$(R^2 - R'^2_0) / R^2$	0.0023	0.0000	0.0086	0.0107	0.0006	< 0.1 [a]
$\lvert R_0^2 - R'^2_0 \rvert$	0.3347	0.0704	0.1417	0.1349	0.0481	< 0.3 [a]
k	0.9979	1.0024	0.9976	1.0005	0.9975	0.85 ≤ k ≤ 1.15 [a]
k'	0.9797	0.9845	0.9805	0.9867	0.9902	0.85 ≤ k' ≤ 1.15 [a]
R_m^2	0.2760	0.5181	0.3665	0.4272	0.5668	-
R'^2_m	0.5686	0.7618	0.5586	0.6874	0.7563	-
$\overline{R_m^2}$	0.4223	0.6400	0.4625	0.5573	0.6616	> 0.5 [b]
ΔR_m^2	0.2925	0.2437	0.1920	0.2602	0.1895	< 0.2 [b]
Model Predictive	Fail	Fail	Fail	Fail	Yes	

*a = standard Golbraikh & Tropsha (2002)
b = standard Roy, Kar & Das (2015)

Table 3:

Table 3. External validation results on the classification model

Models	TP	FP	TN	FN	Sensitivity	Specificity	F-measure	Precision	Accuracy
Deep Learning	215	25	212	10	0.9556	0.8945	0.9247	0.8958	0.9242
Random Forest	219	29	208	6	0.9733	0.8776	0.9260	0.8831	0.9242
SVM	221	137	100	4	0.9822	0.4219	0.7581	0.6173	0.6948
XGBoost	215	26	211	10	0.9556	0.8903	0.9227	0.8921	0.9221

Table 4:

Table 4. Performance of QSAR method workflows that are automated on various targets

Target	Models	Q^2	MSE	$R^2_{(ext)}$	Dataset	Curation	Training	Validation	Test
Beta-1 adrenergic receptor (CHEMBL213)	Deep Learning	0.6601	0.4265	0.9134	1508	620	446	496	50
	MLR	0.1570	1.0578	-0.2724					
	Random Forest	0.7349	0.3326	0.6462					
	SVR	0.7312	0.3373	0.6515					
	XGBoost	0.7099	0.3641	0.6676					
Sigma Opioid receptor (CHEMBL233)	Deep Learning	0.6730	0.6113	0.0736	2280	1157	832	925	232
	MLR	0.4672	0.9959	0.5693					
	Random Forest	0.7725	0.4253	0.7318					
	SVR	0.7543	0.4593	0.7311					
	XGBoost	0.7453	0.4762	0.7422					

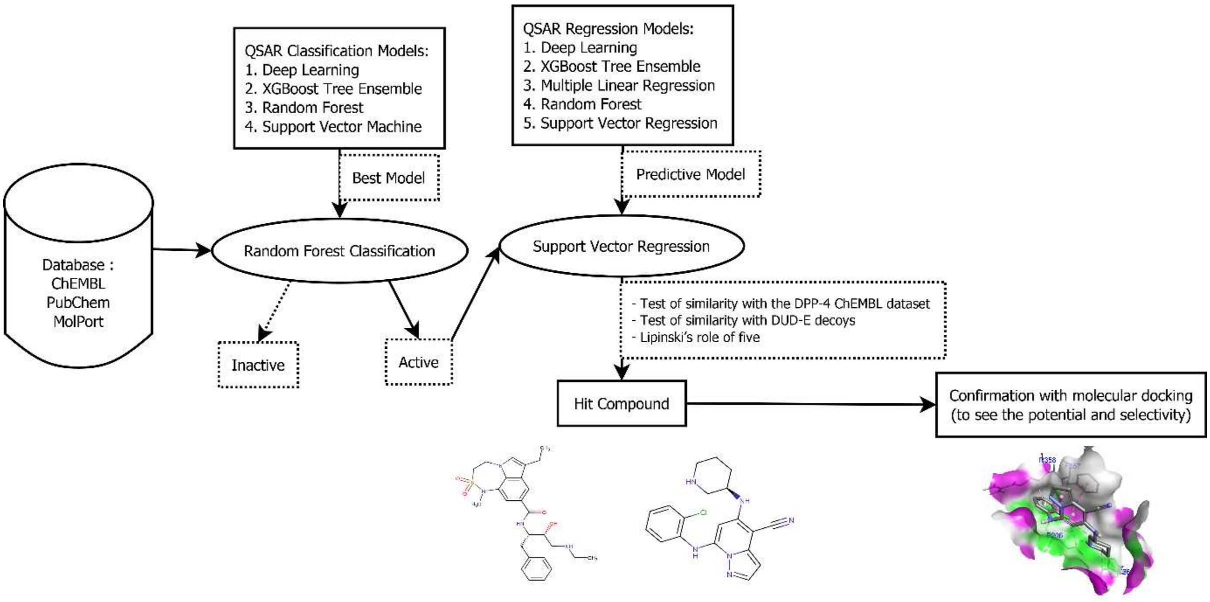
Table 5:

Table 5. Molecular docking results of hit compounds to DPP-4, DPP-8, and DPP-9 enzymes.

No.		Macromolecule (PDB ID)	Ligand	Binding Energy	*Ki*	Unit	Molecule
A	1	5KBY	6RL1510	-9.66	82.49	nM	Trelagliptin
	2		CH0001	-9.42	125.21	nM	
	3		CH0002	-9.67	81.13	nM	
	4		CH0003	-11.27	5.51	nM	
	5		MP0001	-5.67	69420	nM	
	6		MP0002	-8.54	551.45	nM	
	7		MP0005	-4.55	459580	nM	
	8		PC0001	-6.68	12780	nM	
	9		PC0002	-7.13	5910	nM	
	10		PC0003	-8.43	658	nM	
B	1	2ONC	SY1800	-10.43	22.83	nM	Native Ligand of 2ONC
	2		CH0001	-9.51	106.54	nM	
	3		CH0002	-9.78	67.62	nM	
	4		CH0003	-11.46	4	nM	
	5		MP0001	-5.27	136760	nM	
	6		MP0002	-8.5	586.04	nM	
	7		MP0005	-4.7	358230	nM	
	8		PC0001	-6.43	19500	nM	
	9		PC0002	-6.77	10940	nM	
	10		PC0003	-7.97	1440	nM	
C	1	4PNZ	2VH802	-10.37	25.12	nM	Omarigliptin
	2		CH0001	-9.78	67.96	nM	
	3		CH0002	-8.22	940.40	nM	
	4		CH0003	-8.91	293.62	nM	
	5		MP0001	-4.89	261080	nM	
	6		MP0002	-7	7420	nM	
	7		MP0005	-3.69	1980000	nM	
	8		PC0001	-6.86	9350	nM	
	9		PC0002	-6.65	13280	nM	
	10		PC0003	-7.66	2410	nM	
D	1	3KWF	B1Q1	-9.75	71.41	nM	Carmegliptin
	2		CH0001	-9.05	234.16	nM	
	3		CH0002	-9.21	177.65	nM	
	4		CH0003	-10.11	39.1	nM	
	5		MP0001	-4.04	1090000	nM	
	6		MP0002	-7.41	3730	nM	
	7		MP0005	-4.16	897610	nM	
	8		PC0001	-6.94	8170	nM	
	9		PC0002	-6.35	22310	nM	

	10		PC0003	-7.59	2720	nM	
E	1	6HP8	GK2901	-6.69	12490	nM	Native Ligand of 6HP8 (DPP-8)
	2		CH0001	-8.88	310.16	nM	
	3		CH0002	-8.08	1190	nM	
	4		CH0003	-9.93	52.98	nM	
	5		MP0001	-6.42	19660	nM	
	6		MP0002	-5.88	48820	nM	
	7		MP0005	-4.17	879860	nM	
	8		PC0001	-6.91	8650	nM	
	9		PC0002	-6.22	27410	nM	
	10		PC0003	-7.4	3790	nM	
	11		2VH802	-5.21	152160	nM	Omarigliptin
	12		B1Q1	-6.26	25890	nM	Carmegliptin
	13		6RL1510	-7.98	1410	nM	Trelagliptin
	14		SY1800	-8.54	551	nM	Native Ligand of 2ONC
F	1	6EOR	9XH901	-10.32	27.43	nM	Native Ligand of 6EOR (DPP-9)
	2		CH0001	-10.61	16.69	nM	
	3		CH0002	-8.13	1090	nM	
	4		CH0003	-11.4	4.43	nM	
	5		MP0001	-5.26	140350	nM	
	6		MP0002	-6.41	19930	nM	
	7		MP0005	-4.58	435740	nM	
	8		PC0001	-6.77	10990	nM	
	9		PC0002	-5.81	54970	nM	
	10		PC0003	-7.52	3100	nM	
	11		2VH802	-9.17	188.59	nM	Omarigliptin
	12		B1Q1	-7.77	2000	nM	Carmegliptin
	13		6RL1510	-6.55	15850	nM	Trelagliptin
	14		SY1800	-6.82	10060	nM	Native Ligand of 2ONC

Graphical Abstract


QSAR Classification Models:
1. Deep Learning
2. XGBoost Tree Ensemble
3. Random Forest
4. Support Vector Machine
QSAR Regression Models:
1. Deep Learning
2. XGBoost Tree Ensemble
3. Multiple Linear Regression
4. Random Forest
5. Support Vector Regression
Best Model
Predictive Model
Database :
ChEMBL
PubChem
MolPort
Random Forest Classification
Support Vector Regression
Inactive
Active
- Test of similarity with the DPP-4 ChEMBL dataset
- Test of similarity with DUD-E decoys
- Lipinski's role of five
Hit Compound
Confirmation with molecular docking
(to see the potential and selectivity)

Figures

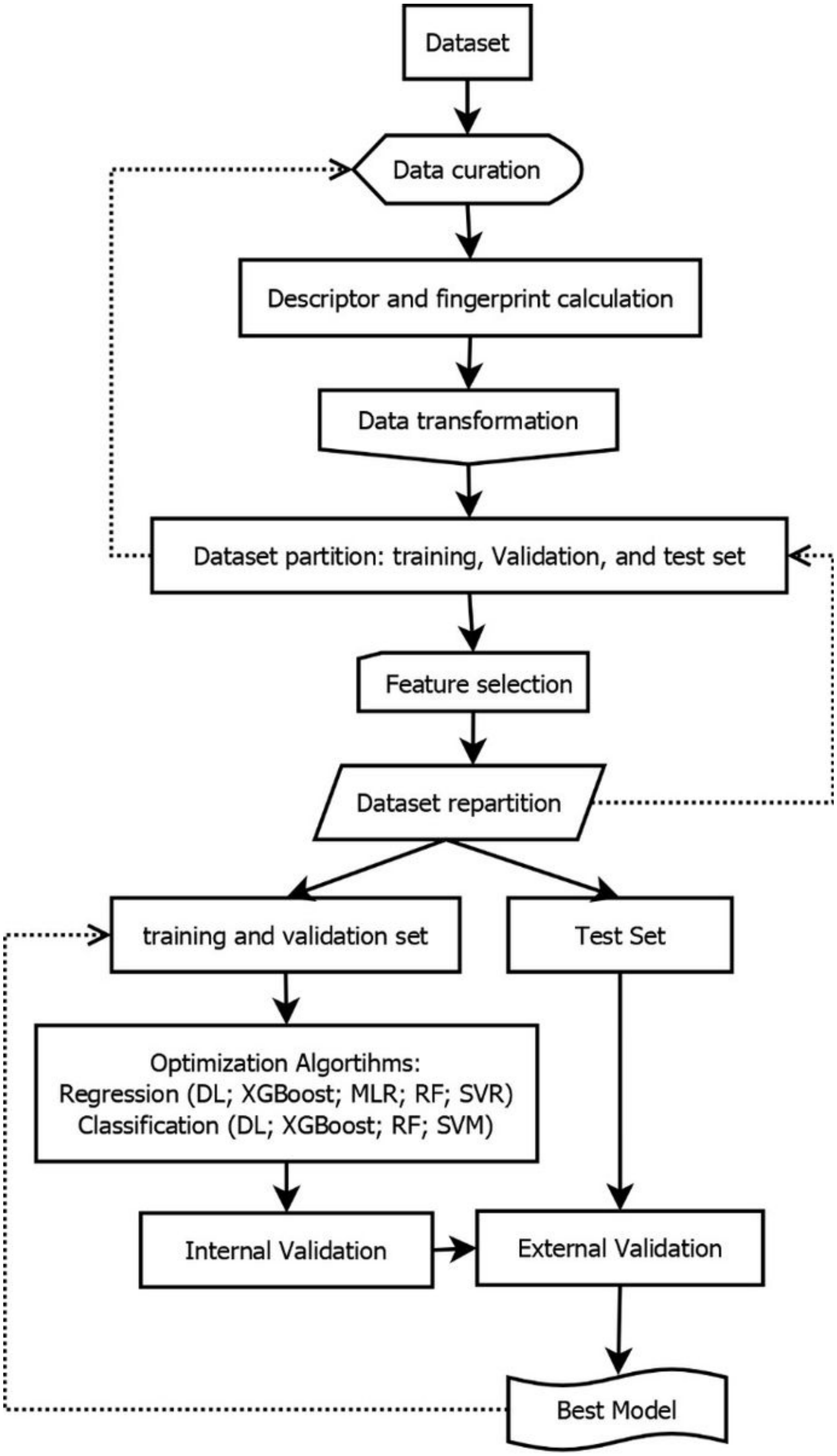
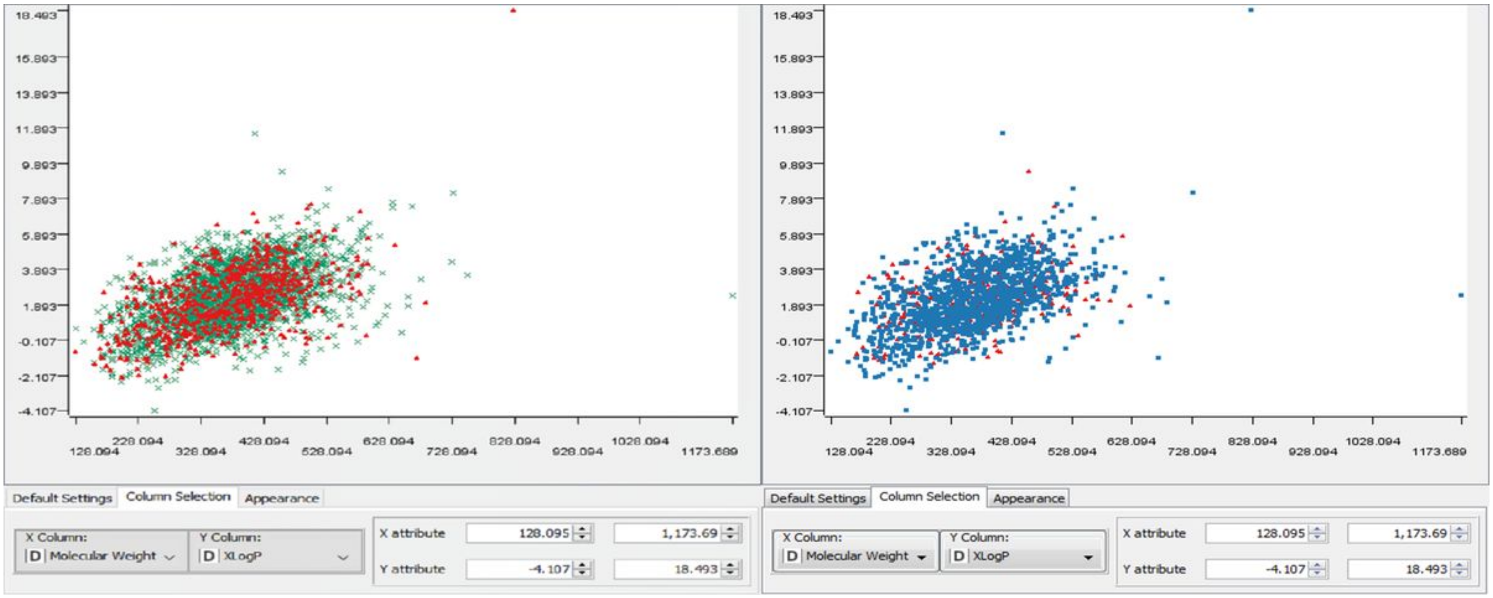

Dataset
Data curation
Descriptor and fingerprint calculation
Data transformation
Dataset partition: training, Validation, and test set
Feature selection
Dataset repartition
training and validation set
Test Set
Optimization Algortihms:
Regression (DL; XGBoost; MLR; RF; SVR)
Classification (DL; XGBoost; RF; SVM)
Internal Validation
External Validation
Best Model

Figure 1

The workflow of QSAR modeling DPP 4 inhibitors



X Column: D Molecular Weight

Y Column: D XLogP

X attribute 128.095 1,173.69

Y attribute -4.107 18.493

X Column: D Molecular Weight

Y Column: D XLogP

X attribute 128.095 1,173.69

Y attribute -4.107 18.493

Figure 2

Chemical space training set versus test set (external validation) defined by MW and ALogP (A) For the regression model (green (x) is a training set, red (Δ) is a test set), and (B) for classification model (blue () is a training set, red (Δ) is a test set.

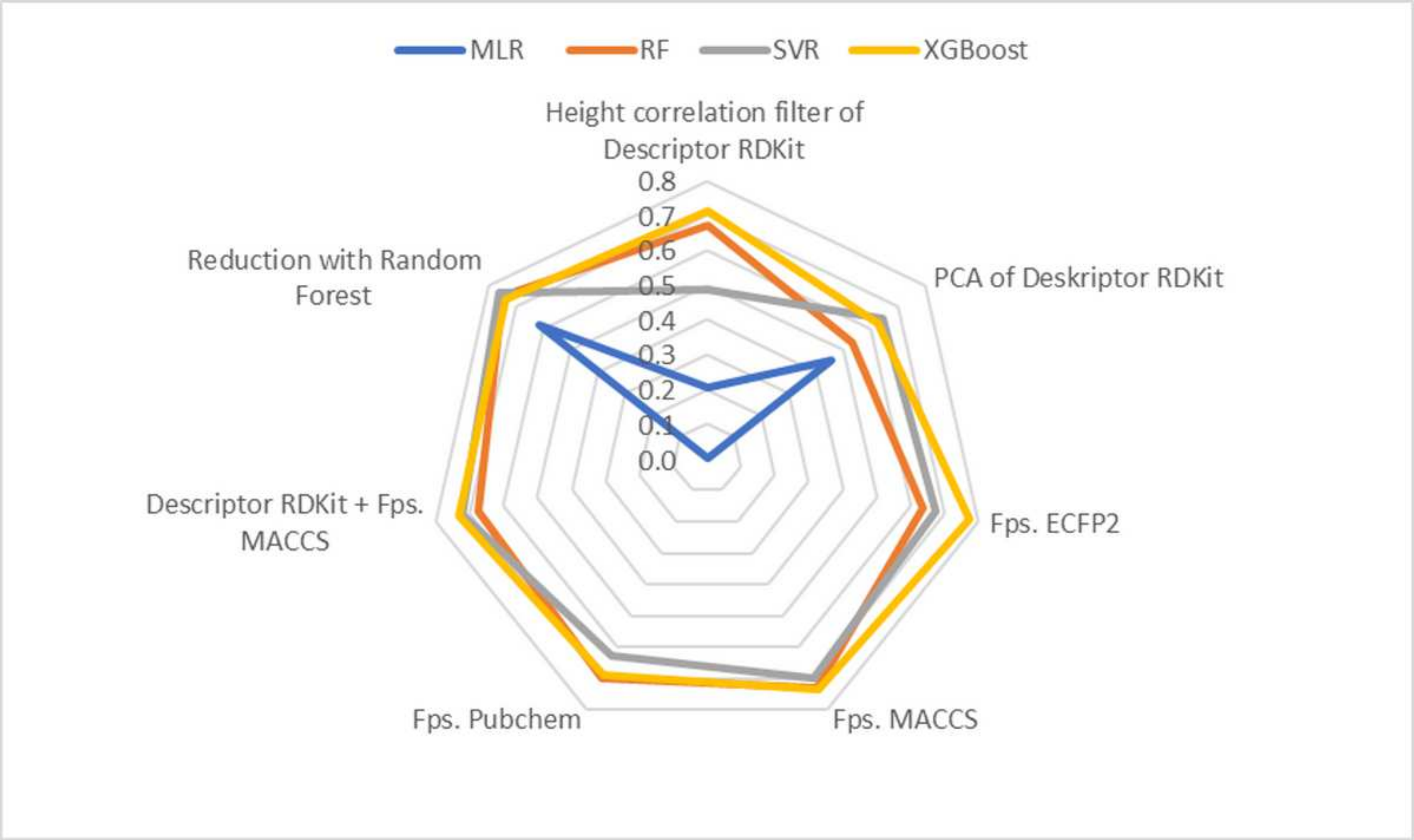


Figure 3

Feature selection. There are seven features developed to get the best method, using four learning models.

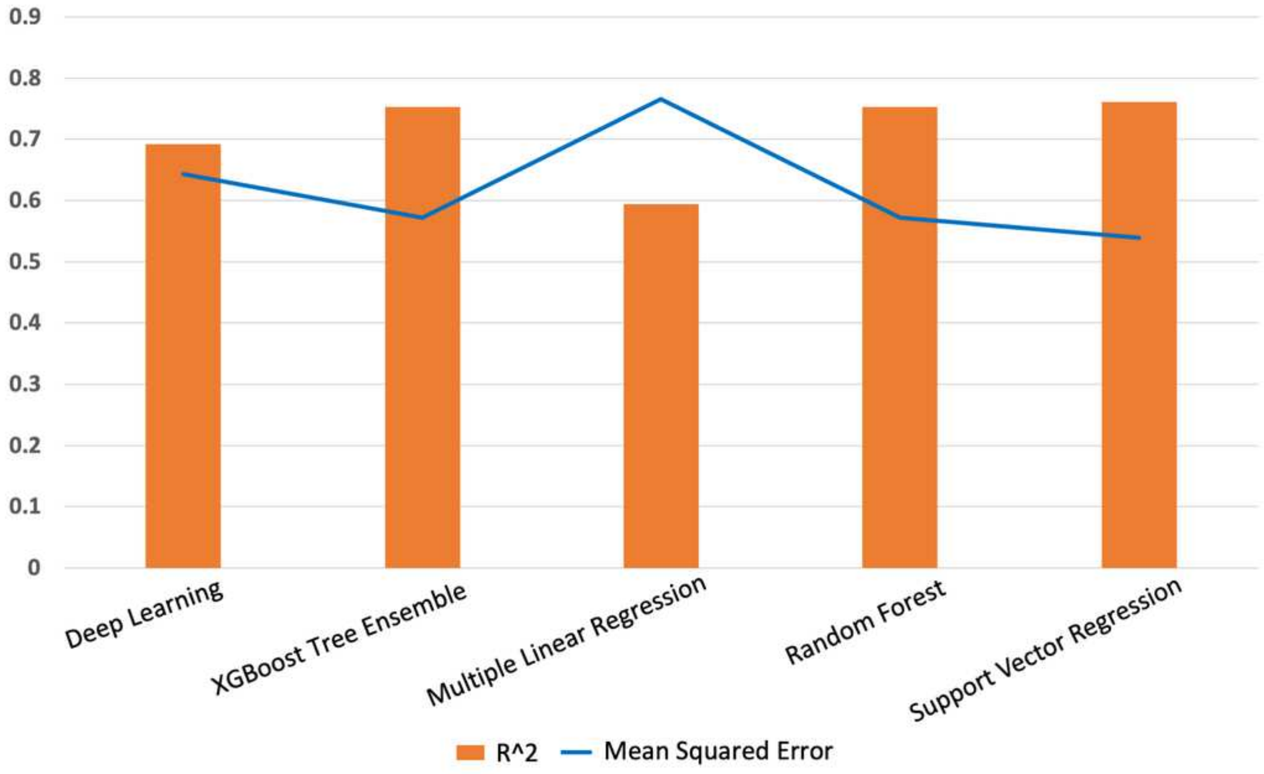


Figure 4

Internal validation results in the regression model The SVR model produces the best performance among other models with the lowest MSE

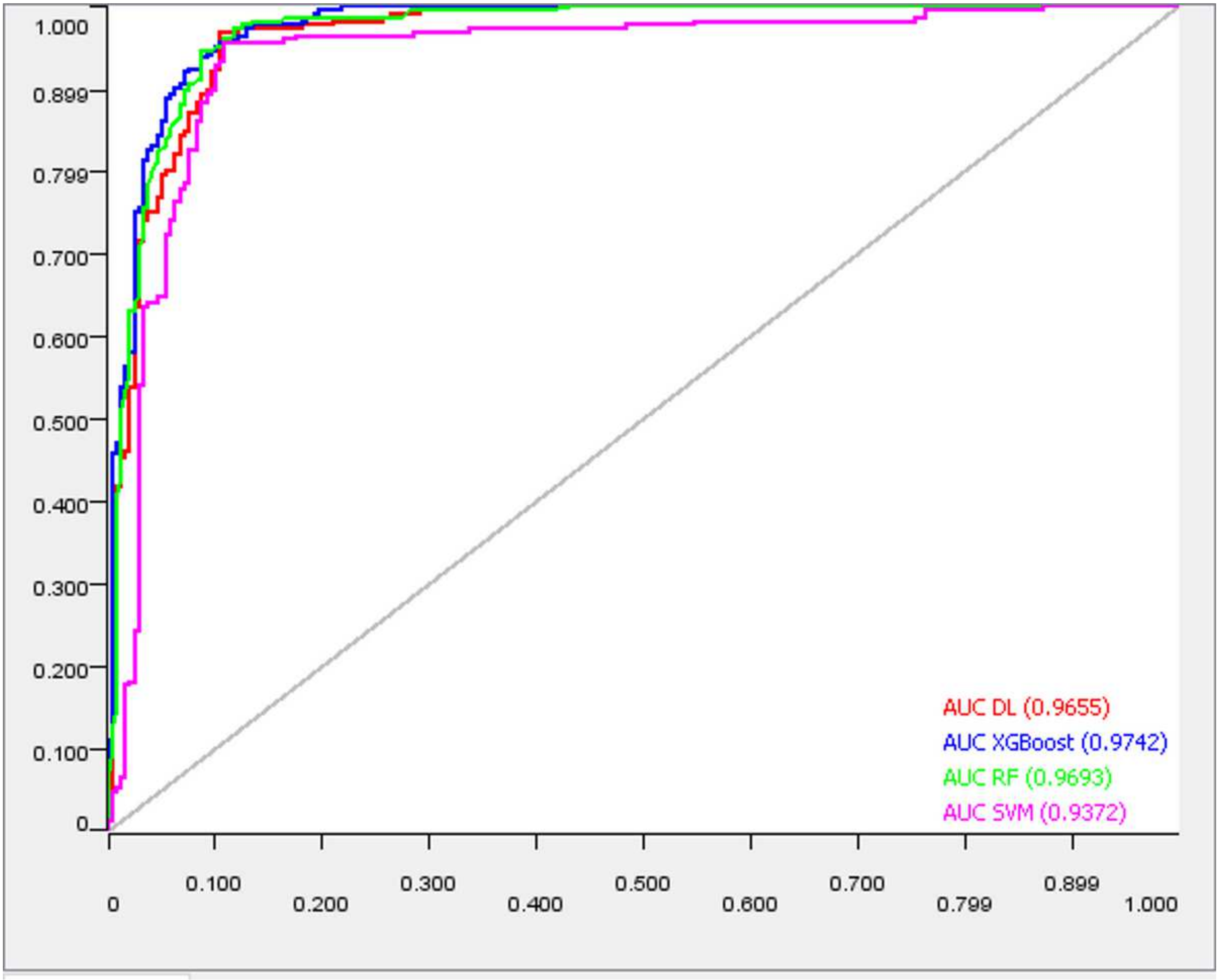
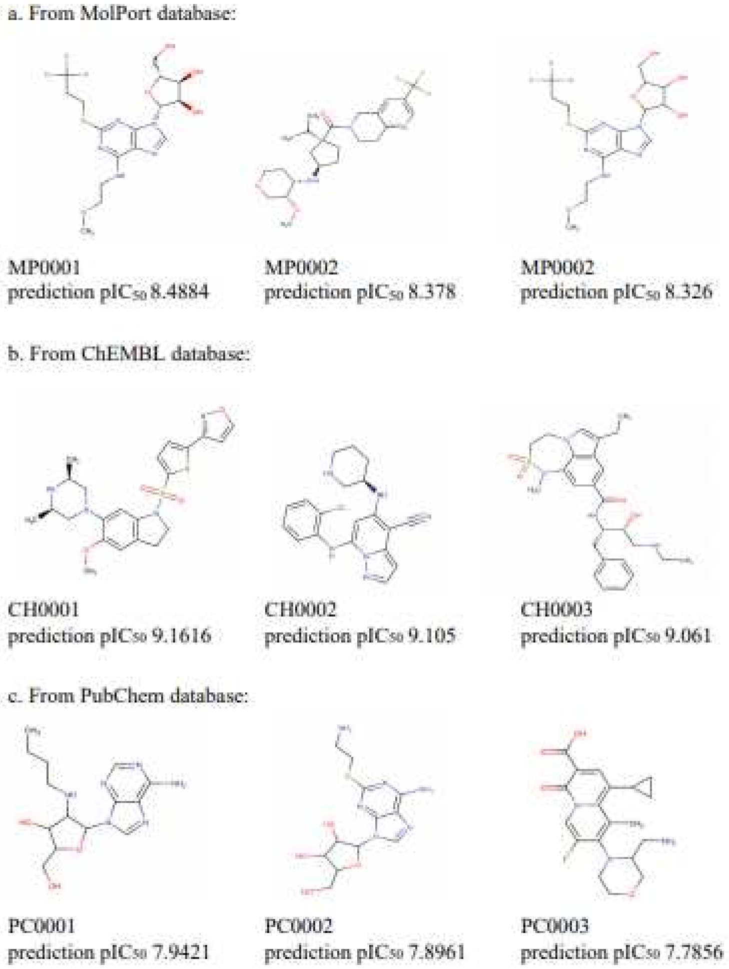


Figure 5

The ROC curve of the four classification models that developed



a. From MolPort database:

MP0001
prediction pIC_{50} 8.4884

MP0002
prediction pIC_{50} 8.378

MP0002
prediction pIC_{50} 8.326

b. From ChEMBL database:

CH0001
prediction pIC_{50} 9.1616

CH0002
prediction pIC_{50} 9.105

CH0003
prediction pIC_{50} 9.061

c. From PubChem database:

PC0001
prediction pIC_{50} 7.9421

PC0002
prediction pIC_{50} 7.8961

PC0003
prediction pIC_{50} 7.7856

Figure 6

Molecular structure of several hit compounds that was resulting from virtual screening

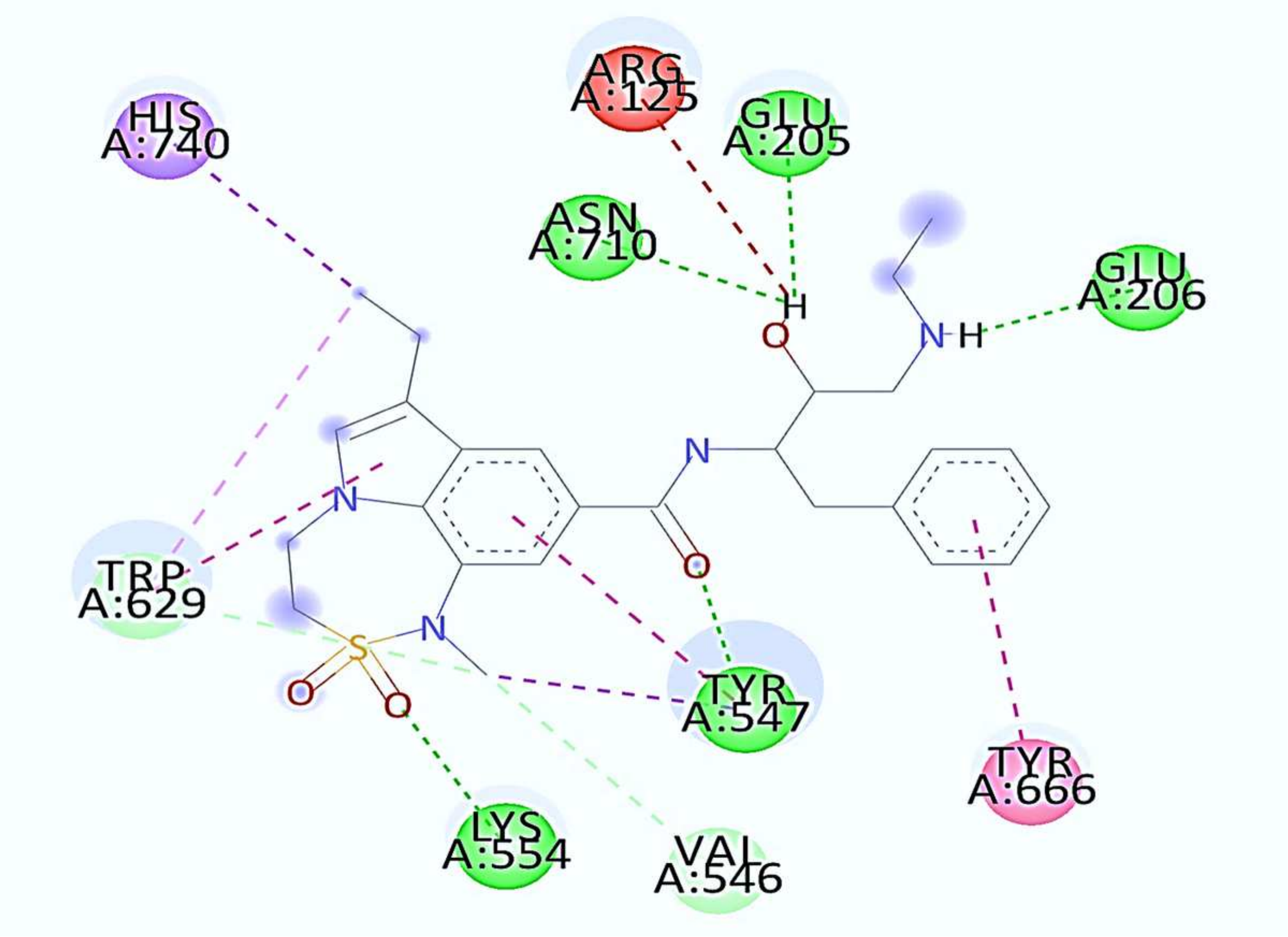


Figure 7

Unfavorable donor-donors interactions (in red) of hit compound CH0003 with DPP-4 (2ONC) crystal structure

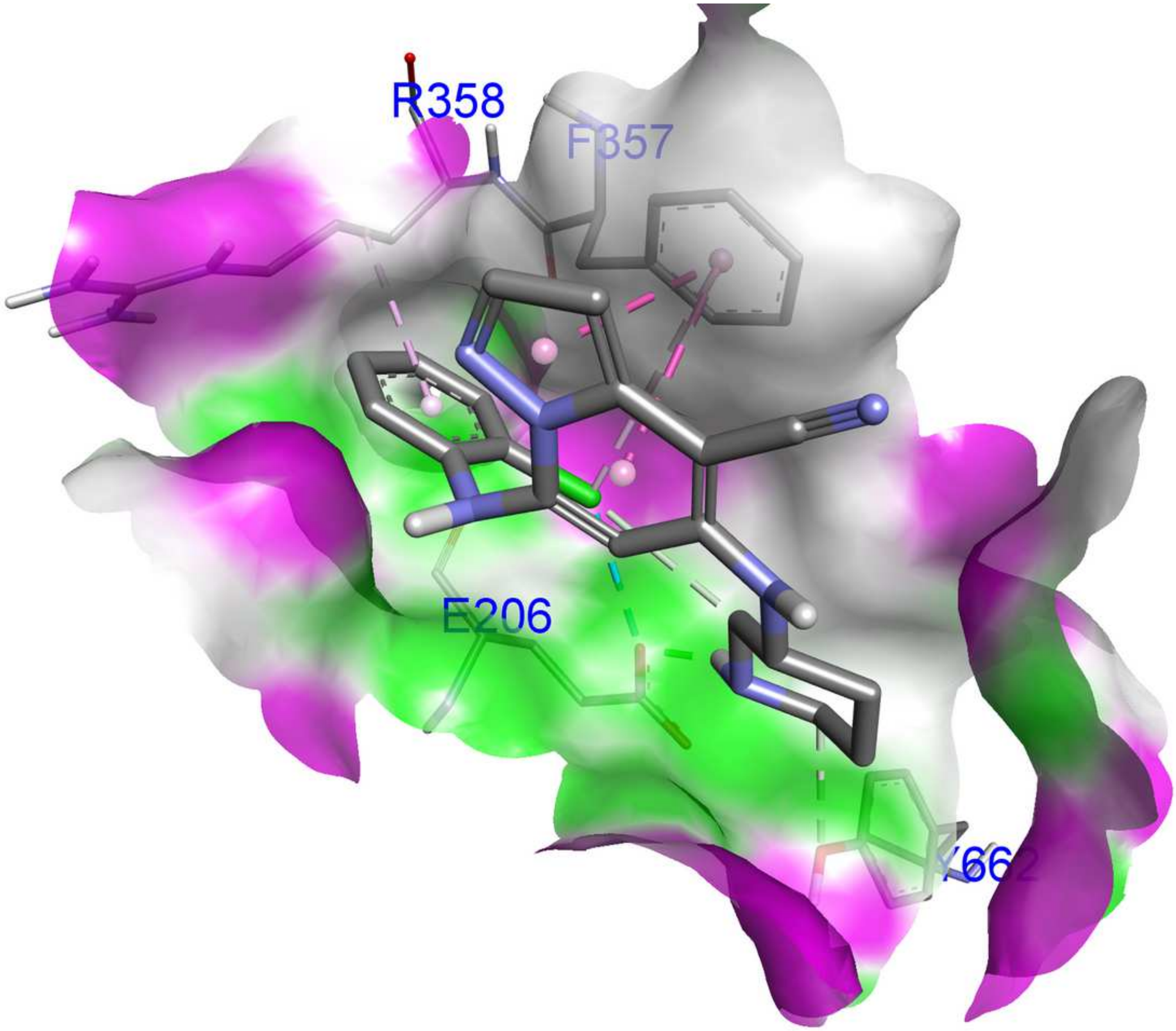


Figure 8

Visualization of the CH0002 hit molecule on DPP-4 (4PNZ).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuplementaryFile.rar
- GraphicalAbstract.png