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

CURRENT STATUS: UNDER REVIEW

BMC Bioinformatics BMC Series

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

Alhadi Bustamam
Universitas Indonesia

Arry Yanuar
arry.yanuar@ui.ac.id*Corresponding Author*
ORCiD: https://orcid.org/0000-0001-8895-9010

DOI:
10.21203/rs.2.22282/v1

SUBJECT AREAS
Bioinformatics

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

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

Full-text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Tables

Table 1. Internal validation results on the classification model

Models	TP	FP	TN	FN	Sensitivity	Specificity	F-Measure	Precision	Acc
Deep Learning	891	93	786	75	0.9224	0.8942	0.9138	0.9055	0.9
Random Forest	925	105	774	41	0.9576	0.8805	0.9269	0.8981	0.9
SVM	952	457	422	14	0.9855	0.4801	0.8017	0.6757	0.7
XGBoost	907	93	786	59	0.9389	0.8942	0.9227	0.9070	0.9

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

Table 2. External validation statistical parameters of various models

Metric	DL	XGBoost	MLR	RF	SVR	Standard
$R^2_{(cv)}$	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 $_{(cv)}$	0.7686	0.6163	1.0134	0.6157	0.5971	-
MSE $_{(pred)}$	1.0370	0.6043	1.0109	0.5914	0.5679	-
	0.2564	0.6914	0.4597	0.6319	0.7281	-
	0.5911	0.7618	0.6014	0.7668	0.7762	-
( ) /	0.5672	0.0924	0.2422	0.1847	0.0624	< 0.1 [a]
( ) /	0.0023	0.0000	0.0086	0.0107	0.0006	< 0.1 [a]
	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]
	0.2760	0.5181	0.3665	0.4272	0.5668	-
	0.5686	0.7618	0.5586	0.6874	0.7563	-
	0.4223	0.6400	0.4625	0.5573	0.6616	> 0.5 [b]
	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. External validation results on the classification model

Models	TP	FP	TN	FN	Sensitivity	Specificity	F-measure	Precision	Ac
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. Performance of QSAR method workflows that are automated on various targets

Target	Models	$R^2_{(cv)}$	MSE	$R^2_{(Pred)}$	Dataset	Curation	Training	Validation
Beta-1 adrenergic receptor (CHEMBL213)	Deep Learning	0.6601	0.4265	0.9134	1508	620	446	496
	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
	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				

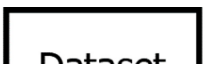
Table 5. Molecular docking results of hit compounds to DPP-4, DPP-8, and DPP-9 enzymes.

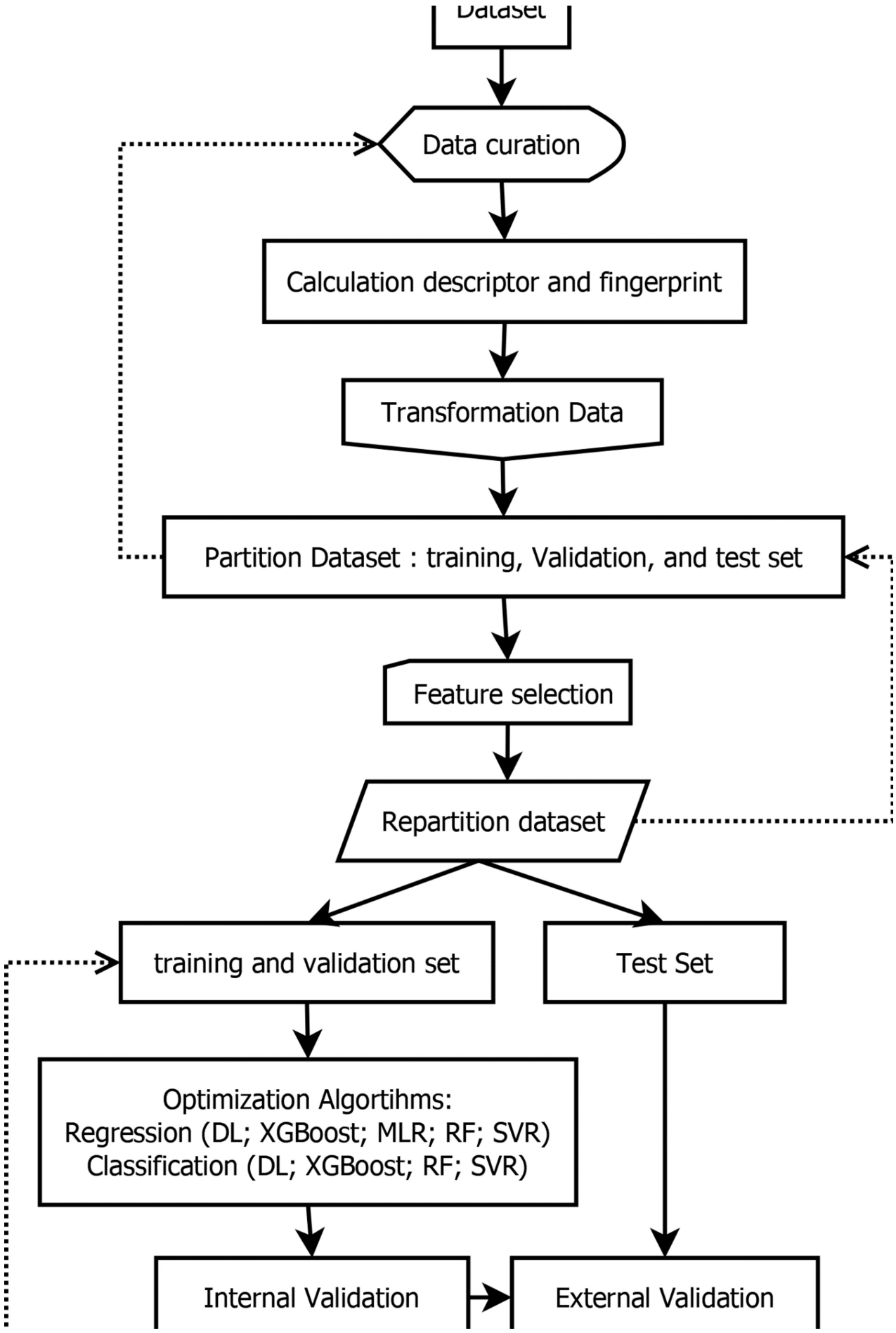
No.		Macromolecule (PDB ID)	Ligand	Binding Energy	*Ki*	Unit	Molecule
A	1	5KBY	6RL1510	-9.66	82.49	nM	Trelagliptin
	2		CH0001	-9.42	125.21	nM	
	3		CH0002	-9.67	81.13	nM	
	4		CH0003	-11.27	5.51	nM	
	5		MP0001	-5.67	69420	nM	
	6		MP0002	-8.54	551.45	nM	
	7		MP0005	-4.55	459580	nM	

	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	Alogliptin
	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	Omargliptin
	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	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	Omargliptin
	12		B1Q1	-6.26	25890	nM	Carmegliptin
	13		6RL1510	-7.98	1410	nM	Trelagliptin
	14		SY1800	-8.54	551	nM	Alogliptin
F	1	6EOR	9XH901	-11.24	5.8	nM	DPP-9
	2		CH0001	-8.84	332.85	nM	
	3		CH0002	-8.07	1220	nM	
	4		CH0003	-10.44	22.27	nM	
	5		MP0001	-5.28	135300	nM	
	6		MP0002	-7.11	6130	nM	
	7		MP0005	-4.37	630360	nM	
	8		PC0001	-7.23	5000	nM	
	9		PC0002	-6.82	10060	nM	
	10		PC0003	-7.76	2050	nM	
	11		2VH802	-8.32	792	nM	Omargliptin
	12		B1Q1	-7.86	1730	nM	Carmegliptin
	13		6RL1510	-7.42	3660	nM	Trelagliptin
	14		SY1800	-7.34	4150	nM	Alogliptin

Figures




Dataset
Data curation
Calculation descriptor and fingerprint
Transformation Data
Partition Dataset : training, Validation, and test set
Feature selection
Repartition dataset
training and validation set
Test Set
Optimization Algortihms:
Regression (DL; XGBoost; MLR; RF; SVR)
Classification (DL; XGBoost; RF; SVR)
Internal Validation
External Validation

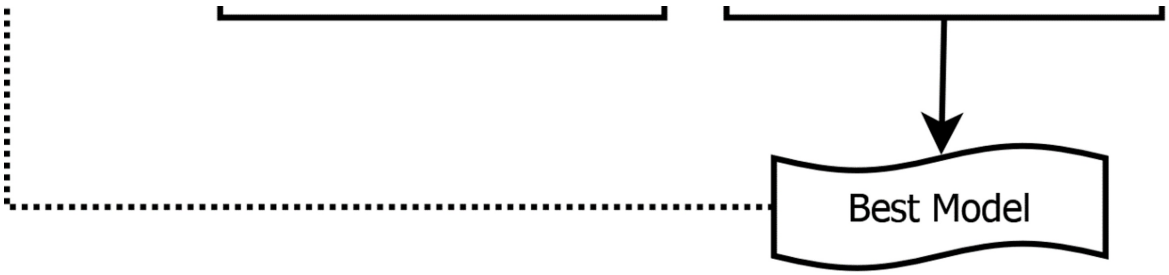


Figure 1

the workflow of QSAR modeling DPP 4 inhibitors

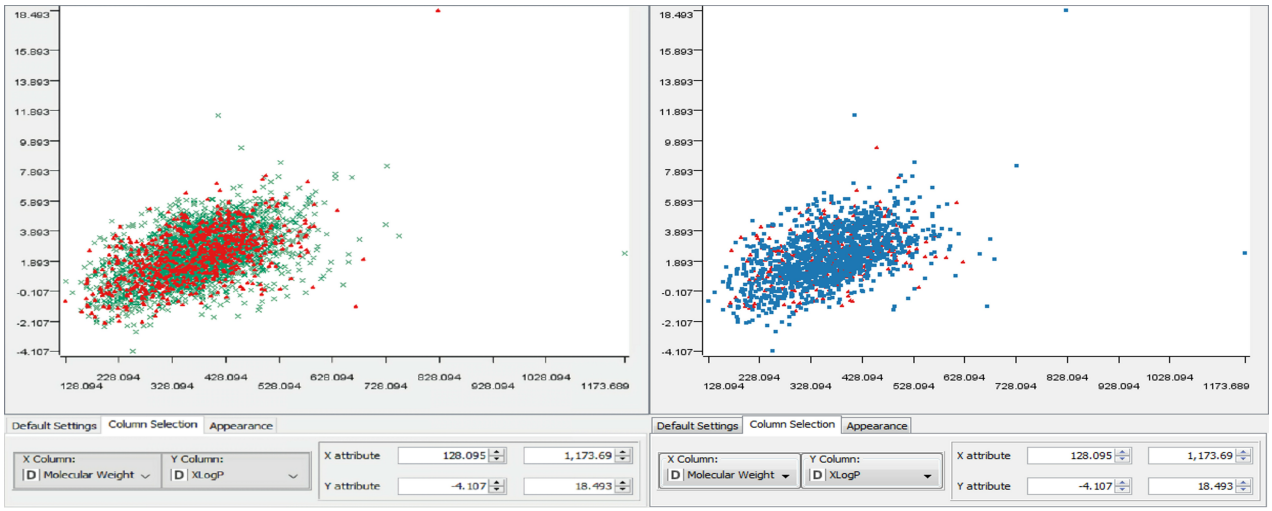


Figure 2

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.

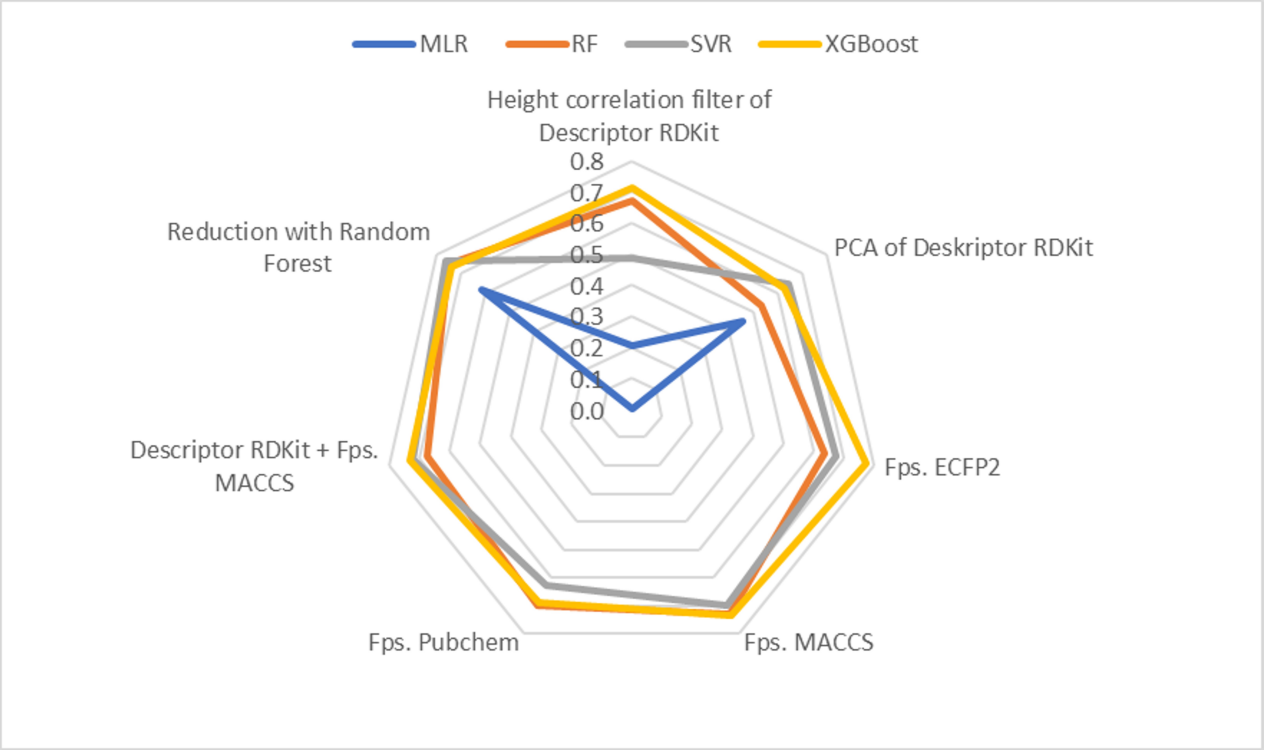


Figure 3

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

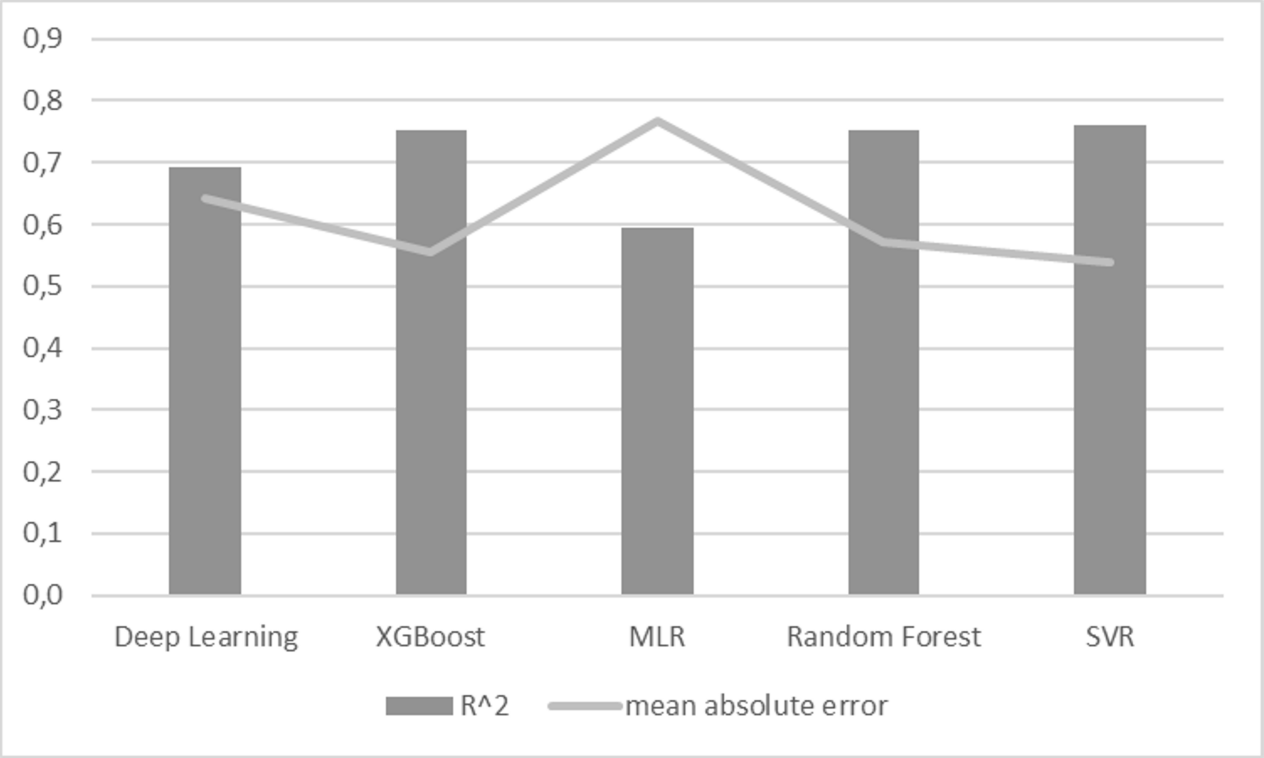
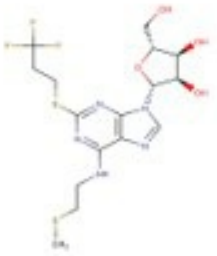


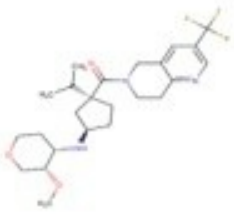
Figure 4

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

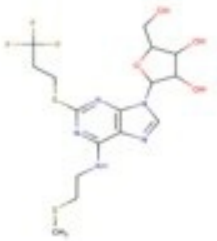
a. From MolPort database:



MP0001
prediction pIC_{50} 8,4884

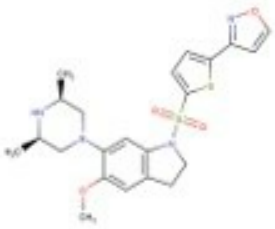


MP0002
prediction pIC_{50} 8,378

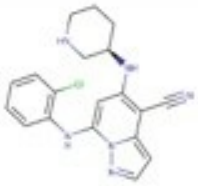


MP0002
prediction pIC_{50} 8,326

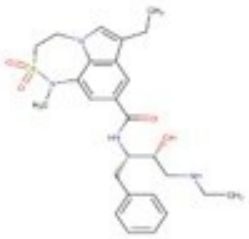
b. From ChEMBL database:



CH0001
prediction pIC_{50} 9,1616

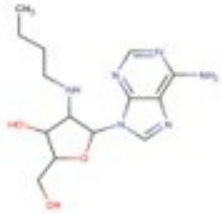


CH0002
prediction pIC_{50} 9,105

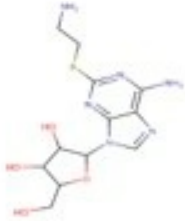


CH0003
prediction pIC_{50} 9,061

c. From PubChem database:



PC0001
prediction pIC_{50} 7,9421



PC0002
prediction pIC_{50} 7,8961

PC0003
prediction pIC_{50} 7,7856

Figure 5

Figure 5. Molecular structure of several hit compounds that was resulting from virtual screening

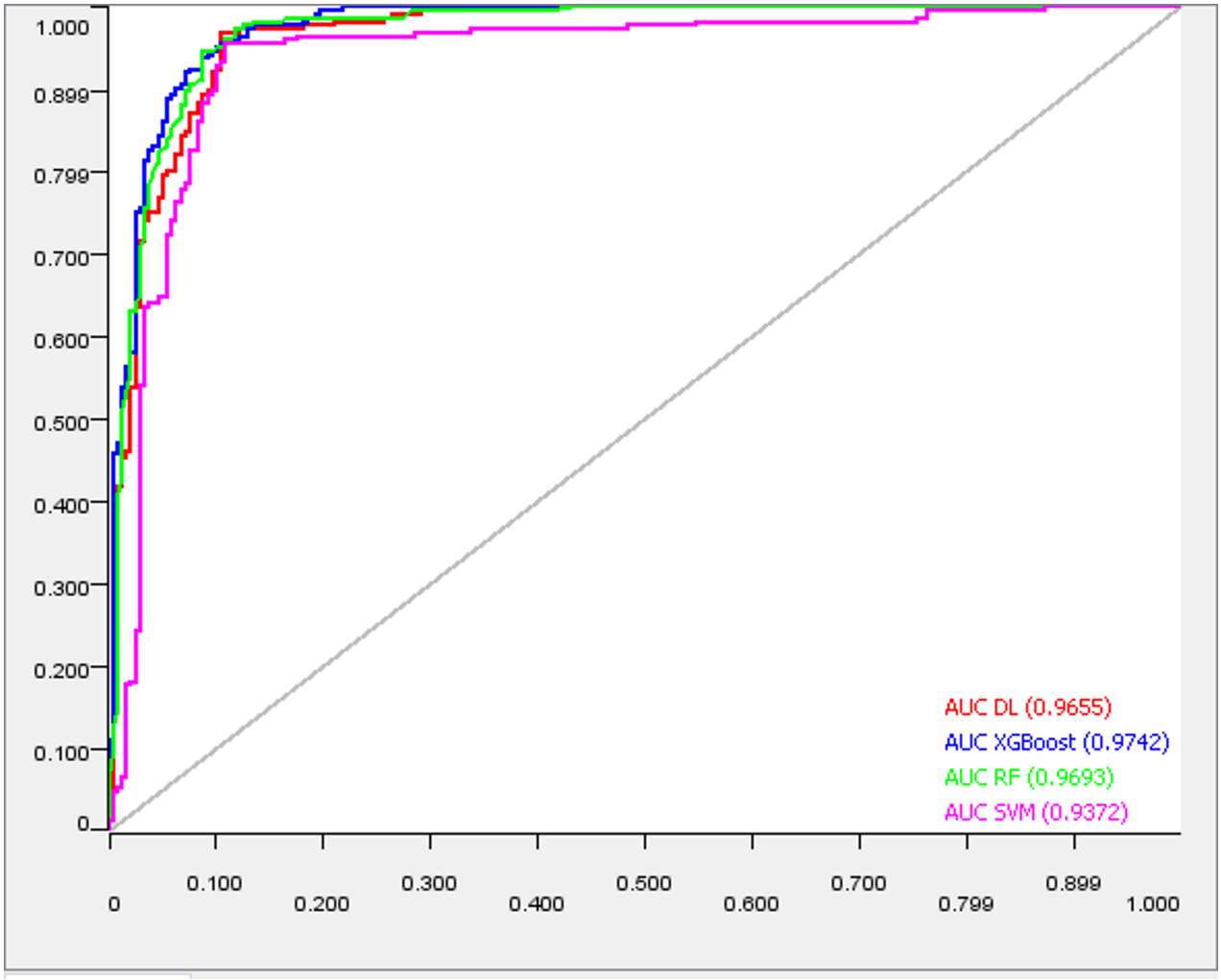


Figure 6

The ROC curve of the four classification models that developed

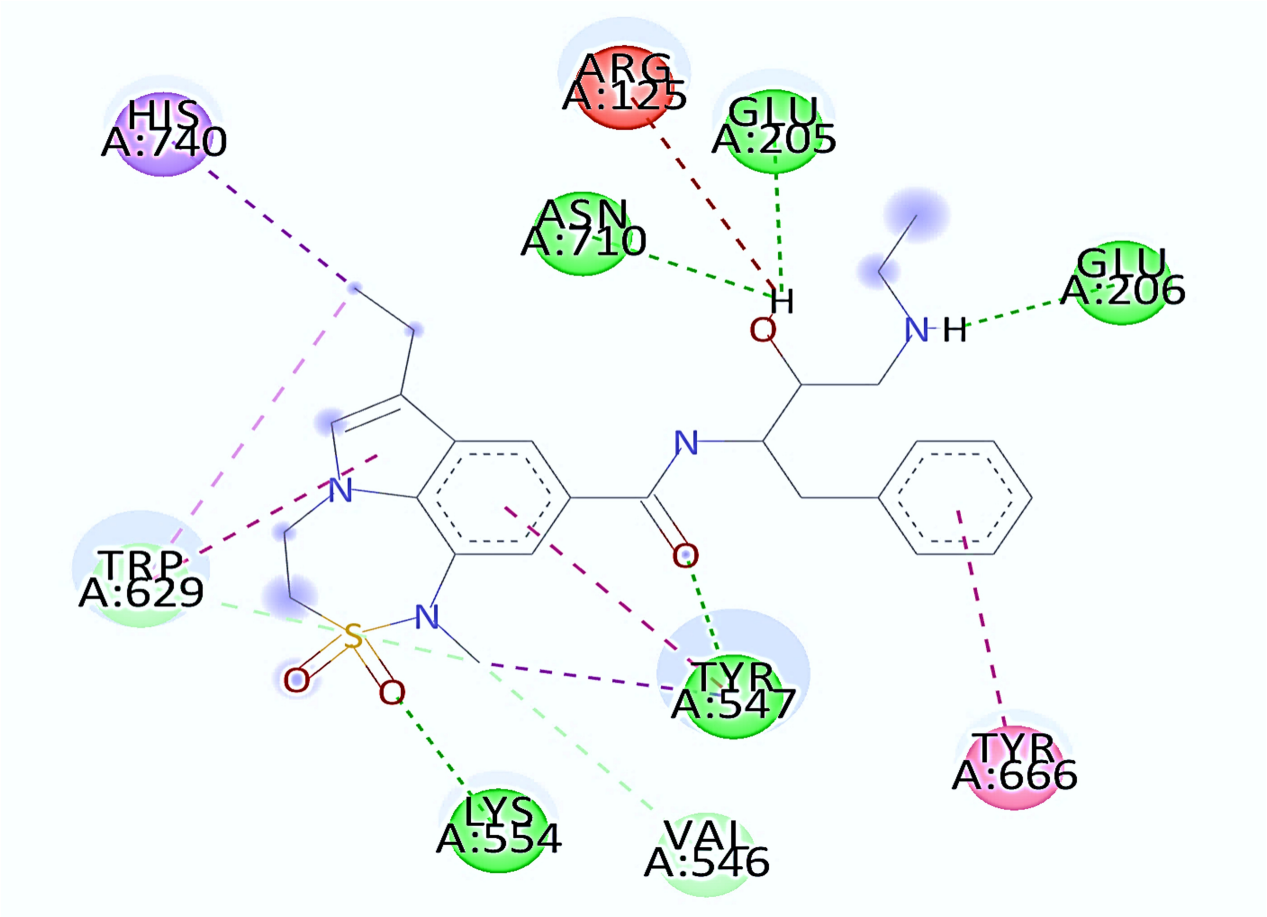


Figure 7

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

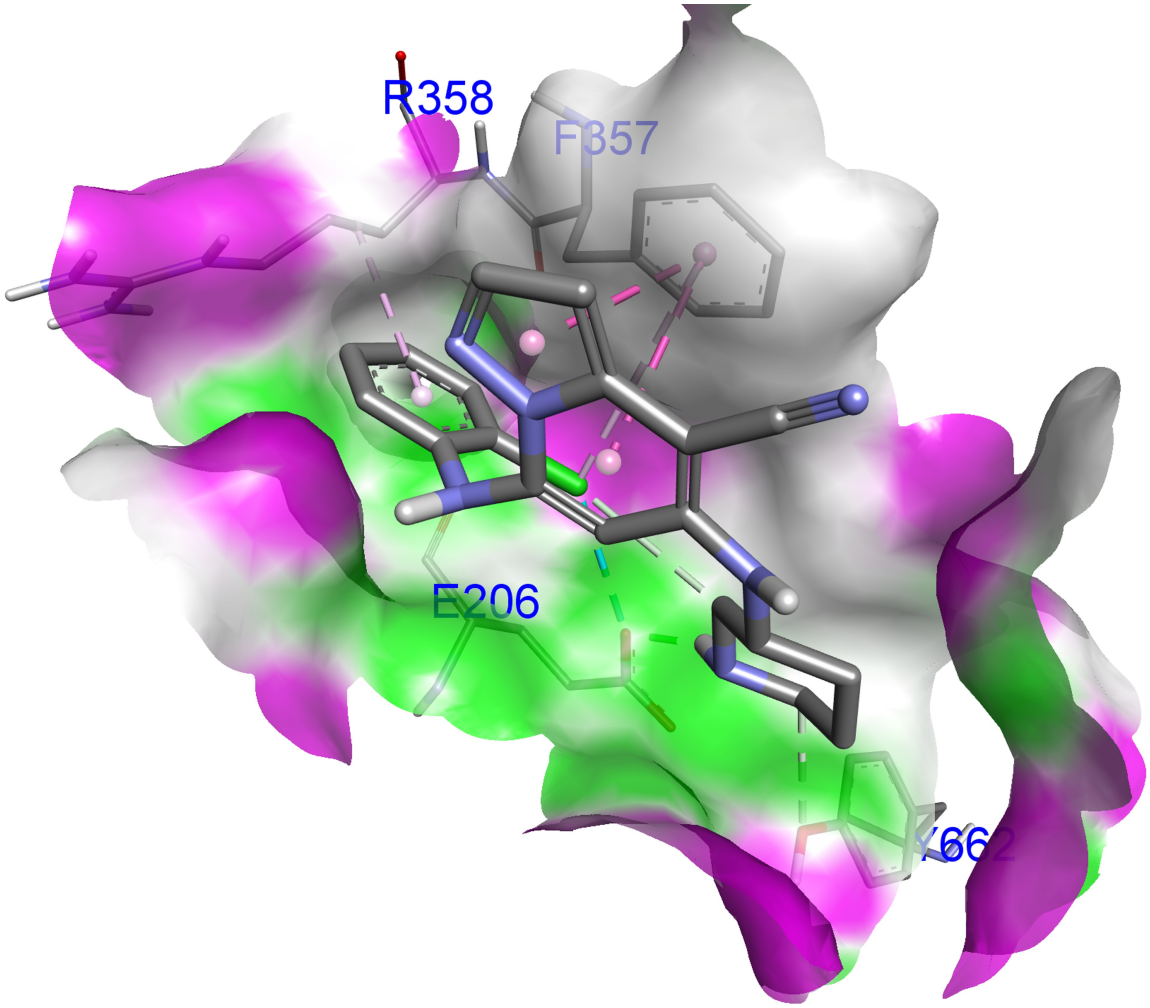


Figure 8

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.

Graphical Abstract.png
Suplementary File.rar