

Supplementary data

QSAR studies of the Bioactivity of Hepatitis C Virus (HCV) NS3/4A Protease

Inhibitors by Multiple Linear Regression (MLR) and Support Vector Machine (SVM)

Zijian Qin, Maolin Wang, Aixia Yan*

State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, P. R. China.

* Corresponding author. Tel.: +86 10 64421335; fax: +86 10 64416428.

E-mail addresses: yanax@mail.buct.edu.cn

1. the brief principle of multiple linear regression (MLR)

The brief principle of MLR is shown via the equation (1).

According to equation (1), the aim of MLR is to find the optimum c_i and c_0 coefficients.

$$y = \sum_{i=1}^n (c_i \times d_i) + c_0 \quad (1)$$

where y represents the regression value (calculated value), i represents the index of descriptors, n represents the number of the descriptors used in this MLR model, c_i represents the linear regression coefficient of the descriptor i , d_i represents the descriptor i , c_0 represents the constant of this MLR model.

2. the brief principle of support vector machine (SVM)

The brief principle of SVM is shown via the equation (2) and (3).

$$\forall \varepsilon > 0, \exists f(x) = \omega^T \cdot x + \omega_0 \quad (2)$$

$$\text{subject to: } |y_i - f(x_i)| - \varepsilon \leq 0 \quad (3)$$

where ε represents the insensitive loss parameter of a support vector machine, ω and ω_0 represent the weights of a support vector machine, i represents the index of input data, x_i and y_i represent the attributes and a label of input data, respectively.

The epsilon-SVR (ε - support vector regression) which was one of the SVM-regression methods was adopted in this work and the aim of epsilon-SVR is shown via the equation (4) and (5).

$$\min_{\omega, b, \xi, \xi^*} \left[\frac{1}{2} \omega^T \omega + C \sum_{i=1}^l (\xi + \xi^*) \right] \quad (4)$$

$$\text{subject to } \begin{cases} y_i - f(x_i) \leq \varepsilon + \xi \\ f(x_i) - y_i \leq \varepsilon + \xi^* \\ \xi, \xi^* \geq 0 \end{cases} \quad (5)$$

where C represents the loss parameter, ξ and ξ^* represent the relaxation factors of a support vector machine, respectively.

The kernel functions are very important in SVM-regression. There four kernel functions, which include linear function, polynomial function, radial basis function (RBF) and sigmoid function, are available in LibSVM toolbox.

According to equation (2) - (5), there are four important parameters, C , g (kernel

function parameters), ε and kernel function type in epsilon-SVR analysis. The optimization of these four parameters can be taken out by the python script “gridregression.py” which is available at the LibSVM website. The “gridregression.py” uses the ergodic search with a cross validation method to optimize parameters C, g and ε in a given kernel function type.

3. r^2 , sd and MAE

Their equations for calculating r^2 , sd and MAE are shown as equation (6) to (8).

$$r^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_{cal,i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

$$sd = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - m - 1}} \quad (7)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - y_{cal,i}|}{n} \quad (8)$$

where y_i and $y_{cal,i}$ represent the experimental and the calculated pIC_{50} value of the compound i , \bar{y} represents the mean experimental pIC_{50} value of all the compounds, n represents the number of compounds, m represents the number of descriptors.

4. “gridregression.py” for optimum parameters

The type of kernel functions and parameters C, g, ε were optimized by the python script “gridregression.py” with ergodic search method based on the training set of each model at first. With a ergodic step of every parameter is 2^1 , parameter C was from 2^{-1} to 2^6 (2^{-1} , 2^0 , 2^1 , ..., 2^6), and parameter g was from 2^{-8} to 2^0 (2^{-8} , 2^{-7} , 2^{-6} , ..., 2^0), and parameter ε was from 2^{-8} to 2^{-1} (2^{-8} , 2^{-7} , 2^{-6} , ..., 2^{-1}), and there were 4 types of kernel functions, so we got $8 \times 9 \times 8 \times 4 = 2304$ groups of parameters. Every group of parameters was used for building models by a 5-fold validation method on a corresponding training set to calculate mean squared error (MSE). At last, only one group of parameters with the minimal MSE value was used as optimum parameters of this SVM model. Thus, we got eight groups of optimum parameters to develop models.

5. how to use our models for predicting the bioactivity of new compounds

Step 1: Preparing a SD file which contains the molecules that one wants to be predicted. The molecules with or without 3D molecular structures are the same for our models.

Step 2: Calculating the molecular descriptors (global and 2D) by using the program CORINA Symphony. Deleting the descriptors which the model is not used.

Step 3: According to the maximum and the minimum values of corresponding training set (Table S6 and S7), scaling every descriptor value to 0.1~0.9 by using the following equation (9):

$$X_i^* = \frac{X_i - X_{min}}{X_{max} - X_{min}} \times 0.8 + 0.1 \quad (9)$$

where X_i^* represents the scaled value, X_i represents the original value, X_{max} and X_{min} represent the maximum and minimum values of the corresponding descriptors in the training set, respectively.

Step 4: For MLR models, the predicted pIC_{50} value can be calculated by the correlation equation. Each correlation equation of the best sub- and whole dataset MLR models is shown in the corresponding subfolder in Supplementary data.

For SVM models, the corresponding training set file can be used to predict the test set file by the LibSVM toolbox. Each parameter and file of the best sub- and whole dataset SVM models is shown in the corresponding subfolder in Supplementary data.

Table S1Distribution of pIC₅₀ value of the dataset inhibitors.

Ranges of IC ₅₀ values	Ranges of pIC ₅₀ values	The number of compounds
IC ₅₀ ≤ 1nM	pIC ₅₀ ≥ 9	59
1nM < IC ₅₀ ≤ 10nM	8 ≤ pIC ₅₀ < 9	169
10nM < IC ₅₀ ≤ 100nM	7 ≤ pIC ₅₀ < 8	85
100nM < IC ₅₀ ≤ 1000nM	6 ≤ pIC ₅₀ < 7	78
1000nM < IC ₅₀ ≤ 10μM	5 ≤ pIC ₅₀ < 6	70
IC ₅₀ > 10μM	pIC ₅₀ < 5	51

Table S2

The descriptions of all 21 selected descriptors.

No.	Descriptor	Description
global		
1	Bonds	Number of bonds (including hydrogen atoms)
2	BondsRot	Number of open-chain, single rotatable bonds
3	Complex	Molecular complexity by Hendrickson
4	HDon	Total number of hydrogen bonding donors derived from the sum of nitrogen and oxygen atoms in the molecule
5	HDonO	Number of hydrogen bonding donors derived from the sum of oxygen atoms in the molecule
6	LogS	Solubility of the molecule in water in [log units]
7	Ro5Viol	Number of violations of the Lipinski's rule of 5 (Weight > 500, XlogP > 5, HDon > 5, HAcc > 10)
8	Ro5ViolExt	Number of violations of the extended Lipinski's rule of 5 (additional rule: number of rotatable bonds > 10)
9	Stereo	Number of atoms (usually carbon atoms) bonded to four different ligands (atoms or groups of atoms), in a spatial arrangement which is not superimposable on its mirror image
2DACorr:		
10	Ident_4	Atom identities ($d_{ij} = 3$) ^a
11	LpEN_5	Lone pair electronegativities ($d_{ij} = 4$) ^a
12	LpEN_11	Lone pair electronegativities ($d_{ij} = 10$) ^a
13	PiChg_3	π atom charges ($d_{ij} = 2$) ^a
14	PiChg_4	π atom charges ($d_{ij} = 3$) ^a
15	PiChg_9	π atom charges ($d_{ij} = 8$) ^a
16	PiChg_11	π atom charges ($d_{ij} = 10$) ^a
17	PiEN_1	π atom electronegativities ($d_{ij} = 0$) ^a
18	PiEN_10	π atom electronegativities ($d_{ij} = 9$) ^a
19	Polariz_7	Effective atom polarizabilities ($d_{ij} = 6$) ^a
20	SigChg_8	σ atom charges ($d_{ij} = 7$) ^a
21	TotChg_3	Total atom charges ($d_{ij} = 2$) ^a

^a d_{ij} represents the topological distance.

Table S3List of inter-correlations between 21 selected molecular descriptors and the bioactivity pIC₅₀ value.

	pIC ₅₀	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
D1	0.32	1																			
D2	-0.39	0.40	1																		
D3	-0.47	0.28	0.82	1																	
D4	-0.46	0.06	0.55	0.82	1																
D5	-0.44	0.32	0.61	0.62	0.44	1															
D6	-0.47	0.29	0.76	0.62	0.41	0.85	1														
D7	-0.43	0.22	0.41	0.47	0.45	0.46	0.49	1													
D8	0.52	0.83	0.15	0.05	-0.17	0.11	0.07	0.10	1												
D9	-0.54	-0.48	0.36	0.52	0.56	0.12	0.18	0.11	-0.58	1											
D10	0.56	0.80	-0.04	-0.13	-0.26	0.05	-0.02	0.09	0.95	-0.73	1										
D11	-0.18	0.23	0.57	0.55	0.37	0.37	0.31	0.10	0.21	0.31	0.07	1									
D12	0.54	0.26	-0.17	-0.22	-0.34	-0.27	-0.37	-0.43	0.54	-0.27	0.51	0.23	1								
D13	0.36	-0.06	-0.19	-0.33	-0.27	-0.37	-0.25	-0.46	0.08	0.06	0.04	-0.11	0.23	1							
D14	-0.50	-0.02	0.38	0.52	0.70	0.39	0.35	0.43	-0.39	0.32	-0.36	0.15	-0.50	-0.50	1						
D15	-0.27	0.01	0.50	0.59	0.54	0.21	0.23	0.14	-0.07	0.60	-0.25	0.45	-0.07	0.12	0.18	1					
D16	0.74	0.00	-0.62	-0.67	-0.55	-0.64	-0.62	-0.53	0.21	-0.37	0.28	-0.38	0.47	0.50	-0.51	-0.34	1				
D17	0.48	0.58	0.17	0.03	-0.23	0.03	0.04	-0.27	0.80	-0.37	0.68	0.30	0.61	0.37	-0.57	0.08	0.24	1			
D18	0.61	0.60	-0.09	-0.20	-0.27	-0.07	-0.10	-0.16	0.82	-0.72	0.83	-0.01	0.52	0.08	-0.40	-0.33	0.36	0.72	1		
D19	0.72	0.68	-0.23	-0.30	-0.30	-0.16	-0.23	-0.23	0.76	-0.63	0.82	-0.04	0.50	0.33	-0.37	-0.26	0.54	0.62	0.74	1	
D20	0.42	-0.02	-0.34	-0.51	-0.71	-0.31	-0.31	-0.48	0.17	-0.42	0.22	-0.16	0.53	0.20	-0.52	-0.40	0.42	0.33	0.23	0.24	1
D21	0.47	0.05	-0.36	-0.32	-0.11	-0.30	-0.33	-0.29	0.22	-0.16	0.19	-0.16	0.25	0.35	-0.31	-0.09	0.59	0.25	0.37	0.34	-0.09

D1: Bonds; D2: BondsRot; D3: HDon; D4: HDonO; D5: Ro5Viol; D6: Ro5ViolExt; D7: Stereo; D8: Complex; D9: LogS; D10: Ident_4; D11: LpEN_5; D12: LpEN_11;

D13: PiChg_3; D14: PiChg_4; D15: PiChg_9; D16: PiChg_11; D17: PiEN_1; D18: PiEN_10; D19: Polariz_7; D20: SigChg_8; D21: TotChg_3

Table S4

Performances of sub-dataset models.

	sub-dataset	Descriptors selection method	Number of descriptors	training/test	Modeling method	training set			test set		
						r ²	sd	MAE	r ²	sd	MAE
SubModel_LA1	linear	Stepwise linear regression	10	243/112	MLR	0.77	0.78	0.59	0.77	0.77	0.59
SubModel_LA2		(Group sub-LA)		243/112	SVM	0.86	0.62	0.44	0.85	0.63	0.46
SubModel_LB1		ReliefFAttributeEval	10	243/112	MLR	0.60	1.03	0.82	0.63	0.98	0.76
SubModel_LB2		(Group sub-LB)		243/112	SVM	0.87	0.59	0.43	0.85	0.63	0.45
SubModel_MC1	macrocyclic	Stepwise linear regression	7	109/48	MLR	0.58	0.57	0.41	0.47	0.66	0.47
SubModel_MC2		(Group sub-MC)		109/48	SVM	0.66	0.52	0.38	0.53	0.61	0.44
SubModel_MD1		ReliefFAttributeEval	7	109/48	MLR	0.30	0.73	0.51	0.27	0.76	0.52
SubModel_MD2		(Group sub-MD)		109/48	SVM	0.76	0.45	0.28	0.67	0.50	0.35

Table S5

Summary of selected descriptors of sub-dataset models.

Description	Sub-dataset	Descriptors selection method	Number of descriptors	Selection results
Group sub-LA	linear	Stepwise linear regression	10 (global: 2 2D: 8)	global: Atoms, Ro5Viol 2DACorr: Ident_3, Ident_10, LpEN_2, PiChg_4, PiChg_11, PiEN_9, TotChg_4, TotChg_11
Group sub-LB	linear	ReliefFAttributeEval	10 (global: 2 2D: 8)	global: HDonO, Stereo 2DACorr: LpEN_2, LpEN_5, LpEN_7, LpEN_11, PiEN_9, PiEN_11, SigChg_6, TotChg_5
Group sub-MC	macrocyclic	Stepwise linear regression	7 (global: 0 2D: 7)	2DACorr: LpEN_2, LpEN_6, PiChg_3, PiChg_5, PiEN_3, SigEN_4, TotChg_7
Group sub-MD	macrocyclic	ReliefFAttributeEval	7 (global: 2 2D: 5)	global: HAcc, HAccN 2DACorr: LpEN_2, LpEN_3, LpEN_8, LpEN_11, PiChg_10

Table S6

The minimum, maximum and average values of molecular descriptors employed by the best whole dataset models.

Model C2 (MLR)				Model D4 (SVM)			
Descriptors	min	max	average	Descriptors	min	max	average
BondsRot	3	32	13.73	HDonO	0	4	0.73
Ro5ViolExt	0	5	3.06	Stereo	3	9	5.39
PiChg_4	-0.35	0.06	-0.06	Ident_4	47	120	89.62
PiChg_9	-0.14	0.28	0.01	LpEN_5	0	182.01	42.35
PiChg_11	-0.37	0.26	0.04	LpEN_11	0	275.33	94.19
PiEN_1	591.74	2798.65	1480.23	PiChg_3	-0.24	0.16	-0.06
PiEN_10	134.09	2592.34	1478.76	PiChg_4	-0.35	0.06	-0.06
Polariz_7	3764.34	13109.00	8249.88	PiChg_9	-0.14	0.28	0.01
SigChg_8	-1.58	0.36	-0.48	PiChg_11	-0.37	0.26	0.04
TotChg_3	-0.55	2.07	0.41	PiEN_10	134.09	2592.34	1478.76
				Polariz_7	3764.34	13109.00	8249.88
pIC ₅₀	3.08	9.52	7.23	pIC ₅₀	3.08	9.52	7.23

Table S7

The minimum, maximum and average values of molecular descriptors employed by the best sub-dataset models.

Linear: SubModel LA1 (MLR)				Linear: SubModel LB2 (SVM)				Macrocyclic: SubModel MC1 (MLR)				Macrocyclic: SubModel MD2 (SVM)			
Descriptors	min	max	mean	Descriptors	min	max	mean	Descriptors	min	max	mean	Descriptors	min	max	mean
Atoms	71	145	102.98	HDonO	0	4	0.88	LpEN_2	0	32.56	1.32	HAcc	11	17	14.16
Ro5Viol	0	4	2.50	Stereo	3	9	5.73	LpEN_6	133.74	319.27	196.83	HAccN	4	9	6.10
Ident_3	49	111	81.35	LpEN_2	0	33.99	0.86	PiChg_3	-0.18	0.16	-0.04	LpEN_2	0	32.56	1.32
Ident_10	32	173	100.69	LpEN_5	0	182	43.03	PiChg_5	-0.18	0.18	0.00	LpEN_3	78.11	199.46	134.28
LpEN_2	0	33.99	0.86	LpEN_7	45.59	403.80	155.59	PiEN_3	661.30	1856.13	1338.50	LpEN_8	6.30	142.82	62.05
PiChg_4	-0.27	0.06	-0.04	LpEN_11	0	256.63	71.53	SigEN_4	6437.78	10457.40	8681.64	LpEN_11	6.41	275.33	144.71
PiChg_11	-0.37	0.26	-0.01	PiEN_9	271.77	2661.41	1497.65	TotChg_7	-1.46	0.53	-0.26	PiChg_10	-0.32	0.07	-0.09
PiEN_9	271.77	2661.41	1497.65	PiEN_11	214.72	2660.40	1418.29								
TotChg_4	-0.60	1.80	0.56	SigChg_6	-0.30	1.79	0.68								
TotChg_11	-2.98	1.85	-0.25	TotChg_5	-3.61	-0.29	-1.51								
pIC ₅₀	3.08	9.05	6.70	pIC ₅₀	3.08	9.05	6.70	pIC ₅₀	4.60	9.52	8.42	pIC ₅₀	4.60	9.52	8.42

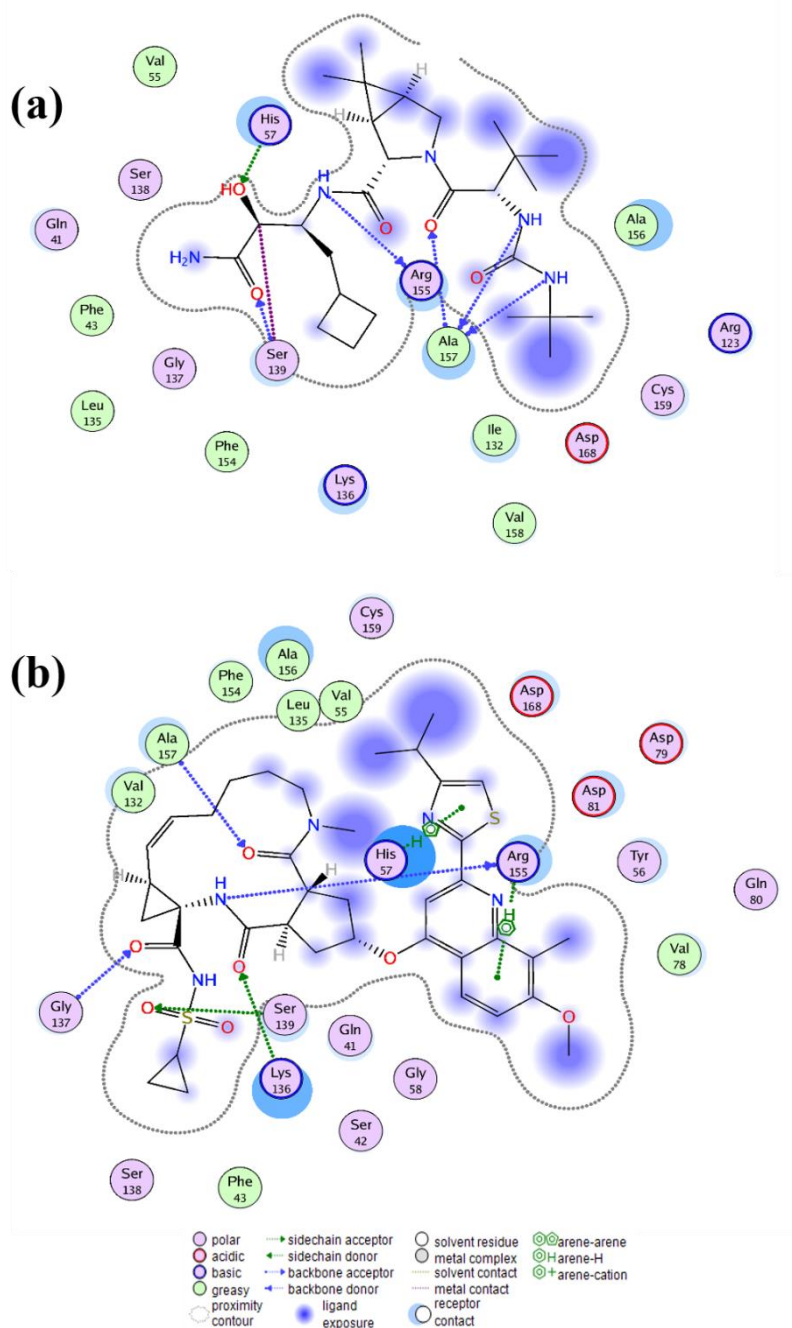


Figure S1. Receptor-ligand interaction systems between the inhibitors and HCV NS3/4A protease. (a) Representative hydrogen binding and covalently binding between the linear inhibitor Boceprevir and the HCV NS3/4A protease (PDB code: 2OC8) generated by using MOE software. (b) Representative hydrogen binding, hydrophobic forces and arene systems between the macrocyclic inhibitor Simeprevir and HCV NS3/4A protease (PDB code: 3KEE) generated by using MOE software.