Journal of Computational Science &Engineering 55 (2021) 1271-1282

Available at http://www.asocse.org ISSN 1710-4068

QSAR Studies on Kinase Inhibitors of Fibroblast Growth Factor Receptor 4 Basis on Particle Swarm Optimization

Jize Wu[1], Junkang Liu[1], Xinyi Zhang[1], Honglin Zhai[2]

[1]Department of Mathematics and Statistics, Qingdao University, Qingdao, 266071, China

[2]Department of Chemistry, Lanzhou University, Lanzhou 730000, China

Abstract: **Hepatocellular carcinoma (HCC) is the sixth most prevalent cancer worldwide, typically developing in patients with underlying viral infections, metabolic disorders, or a history of alcohol abuse. Due to the variety adverse reactions to currently available treatments, there remains a clear requirement for agents that can provide improved outcomes in HCC. It is reported that fibroblast growth factor 19 (FGF19), signaling through fibroblast growth factor receptor 4 (FGFR4) and the coreceptor β-klotho (KLB), is implicated as the oncogenic driver in a subset of HCC, making selective FGFR4 inhibition an attractive treatment opportunity. In this work, quantitative structure-activity relationship (QSAR) studies were conducted on a series of FGFR4 inhibitors to shed light on the molecular requirements for inhibitor affinity and selectivity. Heuristic method (HM) is used to screen out five descriptor parameters, on this basis, multiple linear regression model was built. For the best linear model, the square of correlation coefficient (R^2) was 0.67 and the square of cross validation correlation coefficient (R^2_{cv}) was 0.58. For nonlinear model, support vector machine (SVM) was used, with R^2 and mean square error (S^2) of 0.95 and 0.01 respectively for the training set, 0.84 and 0.04 respectively for the test set. It is apparent that the QSAR model based on SVM has better forecasting stability of inhibitor efficacy, and these findings should be useful for the design of FGFR4 inhibitors as highly selective first-in-class clinical candidate.**

Keywords: Hepatocellular carcinoma; Fibroblast growth factor receptor 4; Quantitative structure-activity relationship; Support vector machine

Received 15 Jan 2021, Revised 20 May 2021, Accepted 22 June 2021

1.Introduction

Liver cancer is the sixth most common cancer worldwide [1]. In 2020, there are more than 900,000 new liver cancer cases and over 830,000 people die of liver cancer, with the death toll closing to the number of new cases. Hepatocellular carcinoma is the most common type of liver cancer and accounts for approximately 90% of cases. The highest incidence and mortality of HCC are observed in East Asia and Africa, although HCC incidence and mortality are increasing in different parts of Europe and in the USA [2]. Hepatitis B virus infection is the most prominent risk factor for HCC development, accounting for around 60% of HCC cases in Asia and Africa and 20% of cases in the West [3][-4]. Additionally, non-alcoholic steatohepatitis, associated with metabolic syndrome or diabetes mellitus, is becoming the fastest growing etiology of HCC, particularly in Europe [2].

The traditional treatment of HCC mainly includes surgery, local ablation and transcatheter arterial chemoembolization. However, most patients are in advanced stage when they are diagnosed with no indication for surgery, so it is urgent to develop novel treatment. In 2007, the launch of sorafenib, the first generation of molecular targeted drug, has made breakthrough in targeted therapy and opened the prelude of systemic therapy. Its high selectivity and low toxicity make it the first-line therapy for systemic treatment of advanced HCC. Other targeted drugs such as lenvatinib and regorafenib have also brought multiple options for the treatment of advanced HCC.

A longstanding challenge to systemic therapy in advanced HCC has been the limited tolerability to treatment, owing in part to adverse events from the treatments themselves. In the pivotal clinical trials investigating multi- kinase inhibitors such as sorafenib, lenvatinib or regorafenib, the rate of treatment- related adverse events of grade ≥3 generally exceeded 50% [5][6][7][-8]. The common adverse reactions include fatigue, nausea, hypertension and hand-foot syndrome. However, dose reductions to delay treatment- related adverse events are prevalent but may interfere with efficacy. Additionally, it is reported that about 70% of patients develop drug resistance 10 years later after clinical introduction of sorafenib [9]. In this context, the identification of new drugs with high specificity and low adverse reactions based on structural biology is highly needed.

FGFR4 belongs to the FGFR family of four receptor tyrosine kinases that interact with fibroblast growth factors (FGFs) to regulate key developmental and physiological processes [10][-11]. Aberrant FGFR signaling is implicated as the driver in a number of cancers and growth disorders, including FGFR1 in non-small-cell lung cancer; FGFR2 in gastric cancer; FGFR3 in bladder cancer; FGFR4 in rhabdomyosarcoma and hepatocellular carcinoma [12][-13]. As a result, numerous efforts have been made to identify FGFR inhibitors and a number of multitargeted kinase inhibitors have the FGFR family

2021@ The American Computational Science Society. All rights reserved.

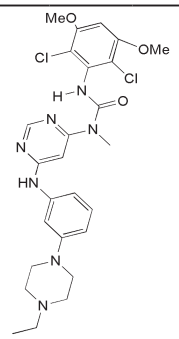
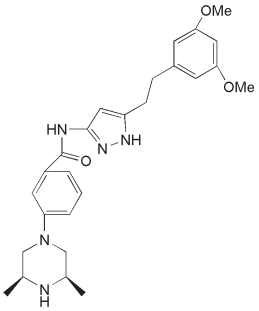
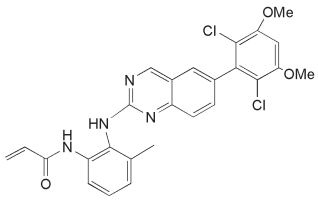
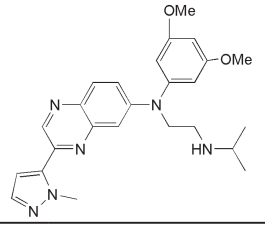
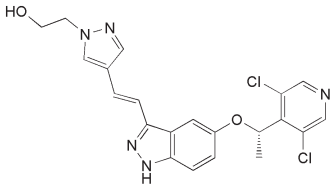
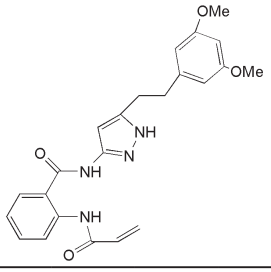
among the list of kinases which they inhibit (examples include derazantinib, dovitinib, lenvatinib, regorafenib, sorafenib). In recent years, studies on FGFR inhibitors have achieved great success. In 2019, erdafitinib, the first FGFR pan inhibitor, was approved by FDA. At present, the research on FGFRs inhibitors focuses on achieving high selectivity of subtypes.

In fact, we have noticed that FGF19, signaling through the FGFR4/β-klotho, has been observed being implicated as the oncogenic driver in a subset of HCC. FGF19 is amplified in a number of HCCs, where it likely activates FGFR4 in an autocrine fashion. To address this opportunity, it is important to conduct prediction study on selective FGFR4 inhibitors as an improved approach to the already described selective pan-FGFR inhibitors. Furthermore, a selective FGFR4 inhibitor would be better tolerated and enable a higher or a more sustained level of inhibition.

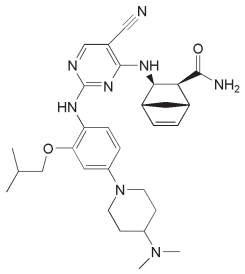
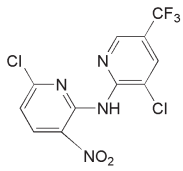
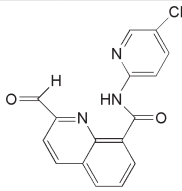
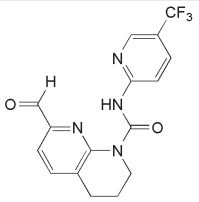
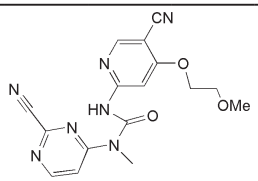
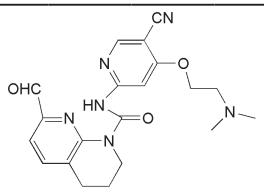
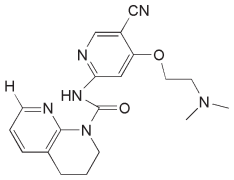
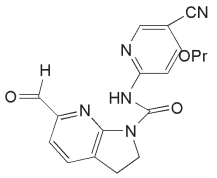
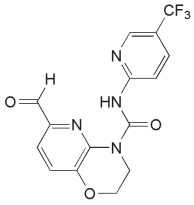
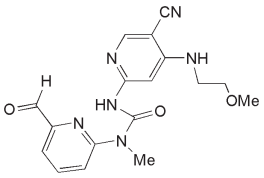
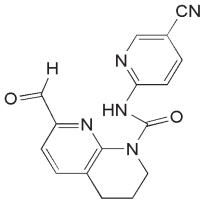
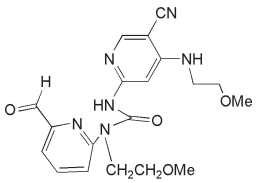
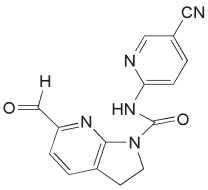
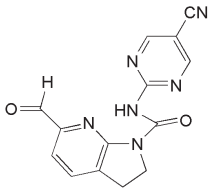
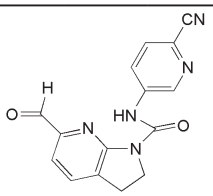
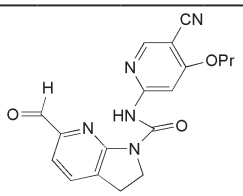
In the processes of traditional drug development, lead compounds should be screened through biological experiments, and then preliminary structure-activity relationship is obtained by the combination of the activity data and compound structure. Due to the long exploitation cycle and high investment of conventional drug development, QSAR method can be used as an effective tool to design novel inhibitors by establishing computer-aided models, which could also reduce the use of chemical reagents so that decrease the risk of environmental pollution.

QSAR have been widely used as an attractive strategy for facilitating the design of inhibitors. This method uses statistical algorithms to unveil significant correlation between structure and biological activities of compounds, and then optimizes lead molecules to screen and discover new drugs. In this study, linear regression heuristic method and nonlinear support vector machine were applied to build QSAR model. And then the structure–activity relationships of a series of compound was investigated, leading to the selective inhibition of the FGFR4. The QSAR model exhibited relatively high accuracy and could reasonably predict the inhibitory activity against FGFR4 inhibitors.

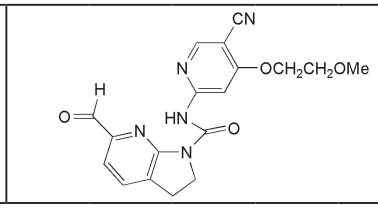
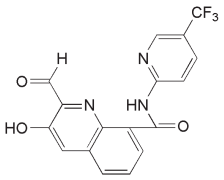
Table 1 compounds and their half maximal inhibitory concentration（IC50）

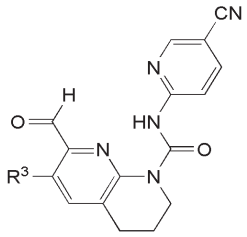
Index	Structure	IC50	Index	Structure	IC50
1a*		71	**1b**		56
1c		14	**1d**		3.8
1e		5.0	**5***		1.5

2021@ The American Computational Science Society. All rights reserved.

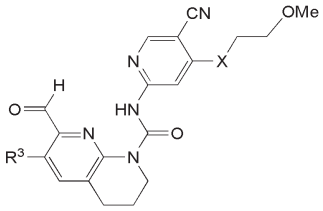
6*		2.2	**7**		53
8		57	**9**		13
15*		8700	**16***		11
17		4400	**19**		11
20		42	**21**		6.5
22*		660	**23**		7.4
24*		16	**25**		45
26*		33	**27**		0.7

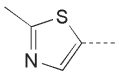
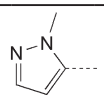
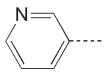
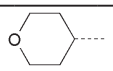
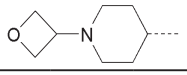
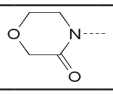
2021@ The American Computational Science Society. All rights reserved.

28		1.3	**32**		58

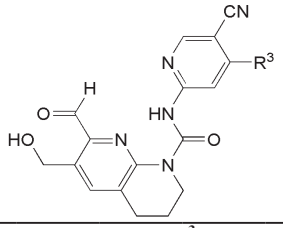


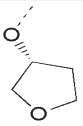
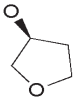
Index	**R^3**	**IC50**	**Index**	**R^3**	**IC50**
33	Me	33	**34***	NH_2	630
35	OMe	387	**36**	CH_2OH	62
37*	Cl	23	**38**	cPr	20
39	CH_2NMe_2	800	**40**	CH_2CO_2H	28

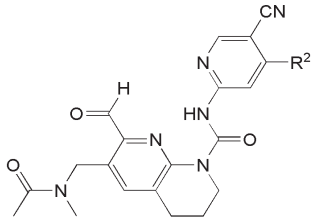


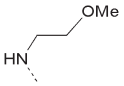
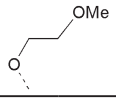
Index	**X**	**R^3**	**IC50**
41	O	Br	2.0
42	O	CH_2OH	3.2
43	NH	CH_2OH	2.9
44	O	CH_2OMe	20
45*	O	CF_2H	2.4
46	NH	CF_2H	0.9
47	O	CF_3	5.8
48	NH		3.0
49	NH		12
50	NH		14
52*	NH		2.5
53	NH		8.2
54	NH		8.8

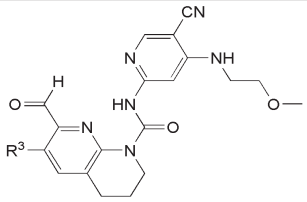
2021@ The American Computational Science Society. All rights reserved.

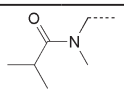
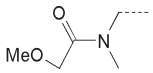
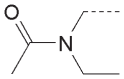
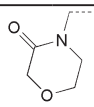
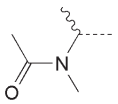


Index	R^3	IC50
57		4.5
58		4.1

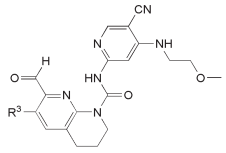


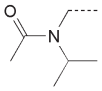
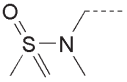
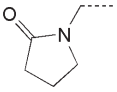
Index	R^2	IC50
59	H	35
60	N(H)^{i}Pr	0.4
61		1.3
62		2.1

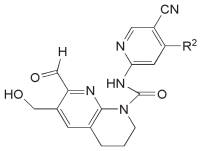


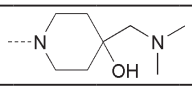
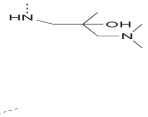
Index	R^3	IC50
63		1.2
64		0.3
65		0.9
69		0.4
70		58

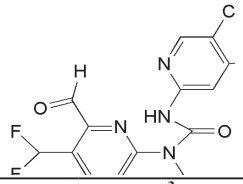
2021@ The American Computational Science Society. All rights reserved.

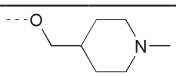
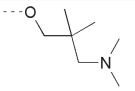


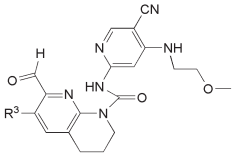
Index	R^3	IC50
66*		1.1
67		0.8
68		0.7

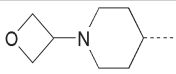
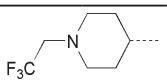
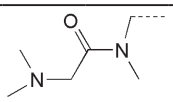


Index	R^3	IC50
71		18
72		18

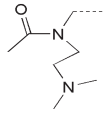
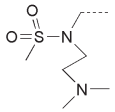
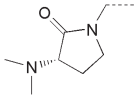
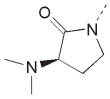
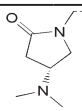
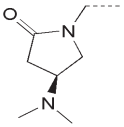
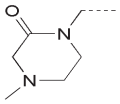
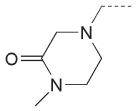
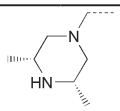
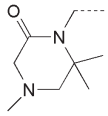
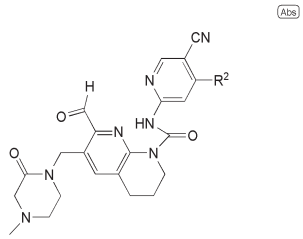
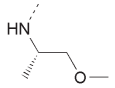
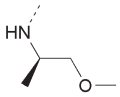


Index	R^3	IC50
73		3.4
74		3.1



Index	R^3	IC50
75		8.1
76		1.2
77		0.9

2021@ The American Computational Science Society. All rights reserved.

78		2.0
79		2.3
80		0.9
81		1.6
82*		0.2
83		0.6
84		0.9
85*		1.3
86		2.6
88		1.4
		
Index	**R^2**	**IC50**
89		2.4
90		1.5

* is the test set of of SVM

2021@ The American Computational Science Society. All rights reserved.

2.Methods

2.1 Data Set

A series of 78 reversible-covalent compounds were selected as FGFR4 inhibitors from the literature[14]. The IC50 values of compounds were converted into lnIC50 and used as the dependent variable in this work. The compounds were divided into a training set and test set at a 4:1 ratio for the construction and internal validation of the SVM models, respectively. The sets contained structural and biological representatives of the whole data set (Table 1). The lnIC50 values are reasonably distributed across the range of values.

2.2 Calculation of the Descriptors

In QSAR studies, compounds must be denoted by molecular descriptors, which refers to the measurement of a molecule's properties in a certain aspect [15]. A wide variety of descriptors have been reported for QSAR analysis, such as constitutional, electrostatic, topological, geometrical and quantum-chemical descriptors. In our study, each molecule was described by five descriptors [16][17][-18].

2.3 Linear Model by HM

HM in CODESSA software can fully search a large number of molecular descriptors to establish the best linear regression equation. Since bringing in new variables may make the introduced ones less significant, the non-significant descriptors should be eliminated. The main processes of HM include weak orthogonal hypothesis and parameter preprocessing [19]. The quality of the model is determined by R^2, F test value (F) and S^2. The stability of the model depends on whether it has good internal consistency, and in this work, the value of R^2_{CV} of Leave-One-Out (Loo) cross-test was used to check it.

2.4 Nonlinear Models by SVM

Vapnik et al.[20] began to study the basic theory of machine learning in the case of small samples in the late 1960s, and put forward the support vector machine algorithm on the basis of statistical learning theory in the mid-1990s. The basic idea of support vector regression (SVR) is to map the data x to a high-dimensional feature space F through a nonlinear constructed function Ö. And then establishing an optimal classification hyperplane, which minimizes the distance from all sample points to the hyperplane [21][22].

The mathematical model of SVM is:

$$y = \min \frac{1}{2} \| \omega \|^2 + C \sum_i (\xi_i + \xi_i^*)$$

Subject to:

$$\begin{cases} y_i - (w \cdot x_i) - b \le \varepsilon + \xi_i \\ (w \cdot x_i) + b - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

where ε is error function, C is penalty parameter, and ξ_i, ξ_i^* are slack variables. The optimal hyperplane linear regression function can be obtained:

$$f(x) = (w \cdot \Psi(x)) + b = \sum_{i=1}^{L} (a_i - a_i^*) K(x \cdot x_i) + b$$

where a_i, a_i^* and b are obtained by constraint conditions.

Selecting the appropriate kernel function can avoid complex operation in high dimensional space. The typical kernel functions include linear, polynomial, Gaussian and sigmoid functions [23]. In this paper, Gaussian function is selected as the kernel function, whose expression is shown as follows:

$$K(x_i, x_j) = \exp\left(\frac{-x_i - x_j^2}{2\sigma^2}\right), \sigma \in R - \{0\}$$

where σ is the width parameter.

All SVM models, in our present study, were implemented using the python software.

2.5 Particle Swarm Optimization (PSO)

The selection of parameters C and σ determines the accuracy of SVM regression and prediction. Therefore, to improve the generalization capability of SVM, in this work, PSO is used to optimize the parameters.

PSO was first proposed by Eberhart and Kennedy [24] in 1995, which originated in the bird flock foraging behavior. The purpose of PSO is to make all particles find the optimal solution in multi-dimensional hypervolume. First, all particles in space are assigned initial random positions and velocities. Then, the position of each particle is pushed successively according to the speed of each particle, the optimal global position and the known optimal position of the particle. With the number of iterations increasing, particles gather around one or more best points by utilizing known advantageous positions in the search space. The algorithm process is as follows: (Figure1)

(1) Initialization. Suppose that the particle swarm size is n, the search dimension is m. The maximum number of iterations, and the maximum velocity of particles are

2021@ The American Computational Science Society. All rights reserved.

set. Position information is the entire search space. The speed and position are initialized randomly in the speed interval and search space.

(2) Individual extremum and global optimal solution. Define fitness function, the individual extremum is the optimal solution found by each particle, and the global optimal solution is the global value from these individual extremums.

(3) Update velocities and positions. Suppose $x_i = (x_{i1}, x_{i2}, ..., x_{im})$ is defined as the position of the particle i, $v_i = (v_{i1}, v_{i2}, ..., v_{im})$ is the speed of the particle i, $p_{besti} = (p_{besti1}, p_{besti2}, ..., p_{bestim})$ is the optimal position of the particle i, and $g_{best} = (g_{best1}, g_{best2}, ..., g_{bestD})$ is the global optimal position. Particle velocity and position are updated as follows:

$$v_{ij}^{t+1} = \omega v_{ij}^{t} + c_1 r_1 \left(p_{bestij} - x_{ij}^{t} \right) + c_2 r_2 \left(g_{bestj} - x_{ij}^{t} \right)$$

$$x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1}$$

where c_1 and c_2 are $\omega = \omega_{\max} - t \times (\omega_{\max} - \omega_{\min}) / N$ learning factors, and in this paper $c_1 = c_2 = 2$. r_1 and r_2 are random numbers of logistic chaotic map on [0,1]. ω is the inertia weight coefficient, here, the upper limit $\omega_{max} = 0.8$ and the lower limit $\omega_{min} = 0.2$. t is the t generation; N is the total number of iterations.

(4) Termination conditions

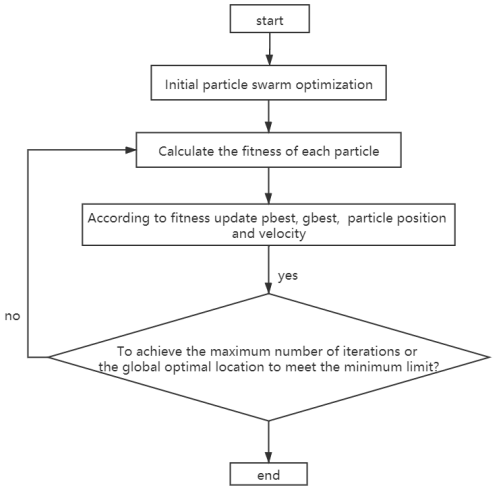


Figure 1 Algorithm process of PSO

3. Results and Discussion

3.1 Calculation Results of HM

One of the major goals of QSAR studies is the determination of the factors influencing the activity of the studied compounds. They contribute to the comprehension of modes of action composed on their biological targets and guide the synthesis toward compounds with optimal activity. In this study, a total of 500 descriptors were obtained by CODESSA software. HM algorithm automatically selected the best descriptors to build a linear model for predicting lnIC50 value. In order to avoid over-parameterization, the model based on different number of descriptors was studied, and the result showed that with the increase of the selected parameters, the value of R^2 increased gradually while the relative error (s^2) decreased (Figures 2 and 3). Therefore, five parameters related to IC50 value were selected to establish QSAR model. The correlation coefficients between any two descriptors are lower than 0.7, indicating that they are independent of each other (Tables 2 and 3). Then the linear equation model based on these descriptor parameters is as follows:

$$lnIC_{50} = -3.9568 \times AIC + 7.2555 \times ABON - 23.564 \times MCI$$

$$+558.2 \times HADH - 1.1506 \times MRCH + 138.4$$

2021@ The American Computational Science Society. All rights reserved.

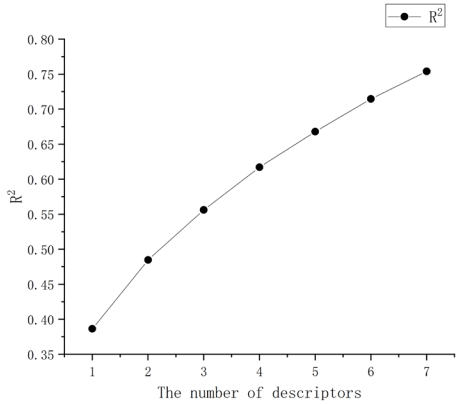


Figure 1 The influence of the number of descriptors on R^2

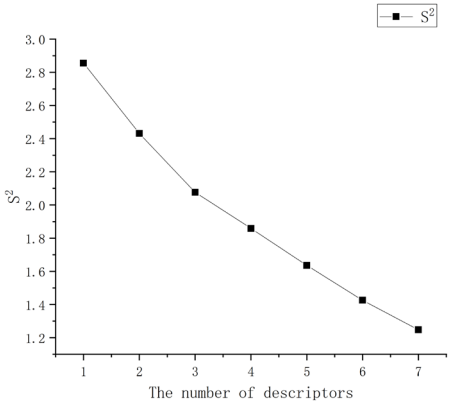


Figure 2 The influence of the number of descriptors on S^2 of linear models

Table 2 The physical-chemical meaning, coefficient, DX and t-test values for molecular descriptors of the best linear model by HM

Symbol	Physical-chemical Meaning	X	DX	t-test
Intercept	Intercept	138.480	20.987	6.598
AIC	Average Information context	-3.957	0.679	-5.829
ABON	Avg bond order of a N atom	7.256	2.413	3.007
MCI	Max coulombic interaction for a C-H bond	-23.564	4.077	-5.780
HADH	HA dependent HDCA-1/TMSA [Zefirov's PC]	558.200	119.110	4.686
MRCH	Min e-e repulsion for a C-H bond	-1.151	0.304	-3.787

Table 3 The correlation matrix of the descriptors in the best linear model

	AIC	ABON	MCI	HADH	MRCH
AIC	1	-0.015	0.255	0.679	-0.016
ABON		1	-0.355	-0.106	0.176
MCI			1	0.392	-0.431
HADH				1	0.122
MRCH					1

Figure 3 The plot of calculated values versus experimental values for all compounds by HM

2021@ The American Computational Science Society. All rights reserved.

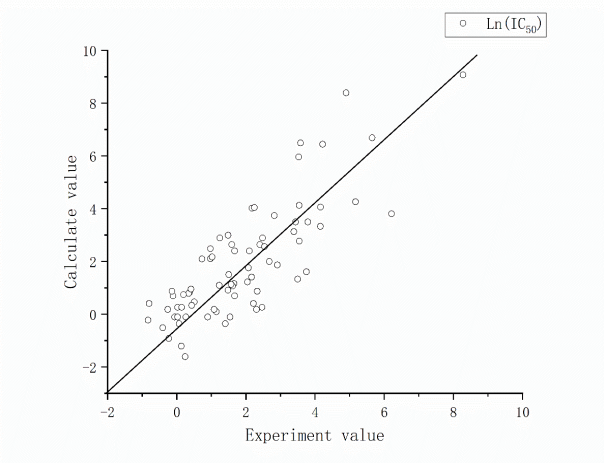


Figure 3 The plot of calculated values versus experimental values for all compounds by HM

The statistical data are: $R^2 = 0.6679$, $R_{cv}^2 = 0.5777$, and $S^2 = 1.6362$. The F value of this model is 27.3544, with the t value of each variable are much larger than the critical value, showing that the model is reliable and each variable are significantly correlated with the target value. The correlation between IC50 experimental value and predicted value of multiple linear equation model is shown in Figure 4. In general, the result is satisfactory but not good enough. Thus, the nonlinear models with corresponding descriptors and lnIC50 value were developed using SVM.

3.2 Results of SVM

In order to be comparable with HM, the test set, training set and parameters selected in SVM are same as HM. The training of the SVM model included the selection of capacity parameter C, ε and the corresponding parameters of the kernel function σ.

Firstly, the selection of kernel function plays a key role in SVM, which determines the sample distribution in the mapping space. Generally, using RBF kernel function will yield better prediction performance, and it was used as the SVM model's kernel in this study accordingly. The radial basis function used is:

$$K\left(x_i, x_j\right) = \exp\left(\frac{-x_i - x_j^{\ 2}}{2\sigma^2}\right), \sigma \in R - \{0\};$$

where, σ is a constant, x_i and x_j are two independent variables, and the generalization ability of SVM is controlled by σ adjusting the shape of RBF function. Through the PSO calculation, the error is the smallest when $\sigma = 0.083$

Secondly, the insensitive loss function ε is introduced into support vector regression. When the distance between all sample points and the hyperplane is not greater than ε, the problem of finding the optimal hyperplane is transformed into a quadratic convex programming problem. The optimal value ε depends on the type of data, and the support vector should also be taken into consideration. Parameter e-insensitive prevents the entire training set meeting boundary conditions so the selection of ε is of great essential for SVM. In the experiment, the error reduces at first and increases then with the increasing of ε, and it reaches the smallest when $\varepsilon = 0.1.$

Lastly, the effect of capacity parameter C was tested. It controls the trade-off between maximizing the margin and minimizing the training error. If C is too small then insufficient stress will be placed on fitting the training data. If C is too large then the algorithm will overfit the training data. The calculation results of PSO show that $C = 280.$

Through the above adjustment, the best parameters σ, ε, C of the model are 0.083, 0.1 and 280, respectively, and at this time the prediction ability is the best while the model is the most stable. As can be seen in Figure 5, in the SVM prediction results, R2 is 0.9340 for training set and 0.8235 for test set. S2 is 0.18 and 1.72 for training set and test set respectively.

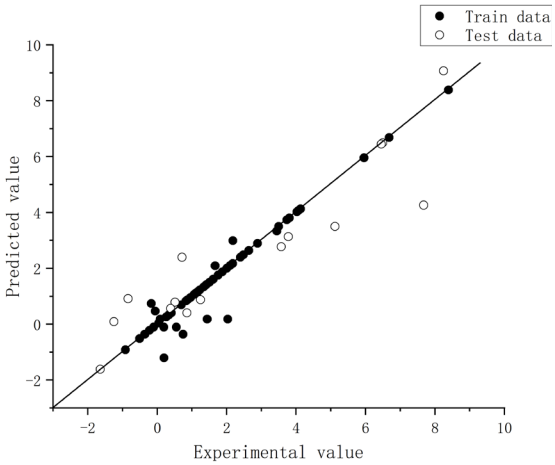


Figure 4 The plot of calculated values versus experimental values for all compounds by HM

3.3 Internal validation

The main goal of any QSAR modeling is that the developed model should be robust enough to be capable

2021@ The American Computational Science Society. All rights reserved.

of making accurate and reliable predictions of biological activities of new compounds. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and confidence of prediction can thus be judged. In this work, bootstrap method is used for internal validation. Bootstrap method is a kind of Monte Carlo simulation first introduced by Efron [25]. Many resamples are drawn from the sample with replacement and the statistic is calculated each time. The distribution of the statistic calculated from resamples forms the bootstrap distribution of the statistic. In this study, 15 samples were randomly selected from 78 samples each time, to get the validation set and calculate the correlation coefficient. Then the samples were put back in sample set. The above steps were repeated five times and the average value of the decorrelation coefficient was 0.72, showing that the SVM model have good performance.

4.Conclusion

This study proposes a method to predict the activity of FGFR4 inhibitors based on HM and SVM. The linear model intuitively reflects the relationship between structural features compounds activities. In order to deeply understand the importance of descriptors, by sorting the five descriptors based on the absolute value of coefficient, the effect of descriptors on lnIC50 was: AIC > MCI > HADH > MRCH > ABON. In addition, we find that PSO has many advantages in optimizing SVM parameters, with fast calculation speed and strong global search capability. Therefore, it can be used as alternative tool in building the QSAR model. Comparing the performance of HM and SVM, the obtained results show that SVM has better qualities and better generalization capabilities than linear regression method, which would seem to have great potential for determining quantitative structure–FGFR4 inhibitor activity relationships and as such be a valuable tool for the chemist.

References

[1] https://www.iarc.who.int/

[2] The Lautenberg Center for Immunology and Cancer Research, IMRIC, The Hebrew University of Jerusalem, Jerusalem, Israel.

[3] Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco (UCSF), CA, USA.

[4] M.J. Sheu, F.W. Liang, C.Y. Lin, T.H, Lu, Population Health Metrics, 19 (2021) 1-8.

[5] J.M. Llovet, S. Ricci, V. Mazzaferro, P. Hilgard, E. Gane, J.F. Blanc, A.C. de Oliveira, A. Santoro, J.L. Raoul, A. Forner, M. Schwartz, C. Porta, S. Zeuzem, L. Bolondi, T.F. Greten, P.R. Galle, J.F. Seitz, I. Borbath, D. Haussinger, T. Giannaris, M. Shan, M. Moscovici, D. Voliotis, J. Bruix, New England Journal of Medicine, 359 (2008) 378-390.

[6] M. Kobayashi, M. Kudo, N. Izumi, S. Kaneko, M. Azuma, R. Copher, G. Meier, J. Pan, M. Ishii, S. Ikeda, Journal of Jastroenterology, 54 (2019) 558-570.

[7] Y.A. Heo, Y.Y. Syed, Drugs, 78 (2018) 951-958.

[8] A.L. Cheng, Annals of Oncology, 21 (2010) 5-5.

[9] M. Pinter, M. Peck-Radosavljevic, Aliment Pharmacol Therapy, 48 (2018) 598-609.

[10] N. Turner, R. Grose, Nature Reviews Cancer, 10 (2010) 116-129.

[11] Beenken, A.; Mohammadi, M, Nature Reviews Drug Discovery, 8 (2009) 235-253.

[12] M. Katoh, Nature Reviews Clinical Oncology, 16 (2019) 105-122.

[13] G.C. Ghedini, R. Ronca, M. Presta, A. Giacomini, Anticancer Therapy 18 (2018) 861-872.

[14] R.A. Fairhurst, T. Knoepfel, N. Buschmann, C. Leblanc, R. Mah, M. Todorov, P. Nimsgern, S. Ripoche, M. Niklaus, N. Warin, V.H. Luu, M. Madoerin, J. Wirth, D. Graus-Porta, A. Weiss, M. Kiffe, M. Wartmann, J. Kinyamu-Akunda, D. Sterker, C. Stamm, F. Adler, A. Buhles, H. Schadt, P. Couttet, J. Blank, I. Galuba, J. Trappe, J. Voshol, N. Ostermann, C. Zou, J. Berghausen, A.D. Espinola, W. Jahnke, P. Furet, Journal of Medicinal Chemistry, 63 (2020) 12542-12573.

[15] W. Ren, D.X. Kong. Computers and Applied Chemistry, 26 (2009) 1455-1458.

[16] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, A.D. Hu, B. T. Fan, Journal of Chemical Information and Computer Sciences, 43 (2003) 1288-1296.

[17] J.L. Mullin, B.P. Huddle, Abstracts of Papers of the American Chemical Society, 209 (1995) 112-112.

[18] M.J. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, Journal of the American Chemical Society, 107 (1985) 3902-3909.

[19] Q.Y. Li, F,R, You, B.X. Jia, H.Z. Si, X.J. Yao, BioMed Research International, 2014 (2014) 1-8.

[20] V. Cherkassky, IEEE Transactions on Neural Networks, 8 (1997) 1564-1564.

[21] J. Hong, Department of Chemistry, Zhejiang University, Hangzhou, 2010.

[22] C. Cortes, V. Vapnik, Machine Learning, 20 (1995) 273-293.

[23] V. Mercer, Philosophical Transactions of the Royal Society A, 209 (1909) 415-446.

[24] J. Kennedy, R. Eberhart, Piscataway, NJ, USA, 1995.

[25] B. Efron, The Annals of Statistics, 7 (1979) 1-26.

2021@ The American Computational Science Society. All rights reserved.