Int. J. Mol. Sci. **2012**, *13*, 7015-7037; doi:10.3390/ijms13067015

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

Toward the Prediction of FBPase Inhibitory Activity Using Chemoinformatic Methods

Ming Hao, Shuwei Zhang and Jieshan Qiu *

Department of Materials Science and Chemical Engineering, Dalian University of Technology, Dalian 116023, Liaoning, China; E-Mails: dluthm@yeah.net (M.H.); zswei@dlut.edu.cn (S.Z.)

* Author to whom correspondence should be addressed; E-Mail: jqiu@dlut.edu.cn; Tel.: +86-411-84986024; Fax: +86-411-84986080.

Received: 23 April 2012; in revised form: 18 May 2012 / Accepted: 31 May 2012 / Published: 7 June 2012

Abstract: Currently, Chemoinformatic methods are used to perform the prediction for FBPase inhibitory activity. A genetic algorithm-random forest coupled method (GA-RF) was proposed to predict fructose 1,6-bisphosphatase (FBPase) inhibitors to treat type 2 diabetes mellitus using the Mold^2 molecular descriptors. A data set of 126 oxazole and thiazole analogs was used to derive the GA-RF model, yielding the significant non-cross-validated correlation coefficient r^2_{ncv} and cross-validated r^2_{cv} values of 0.96 and 0.67 for the training set, respectively. The statistically significant model was validated by a test set of 64 compounds, producing the prediction correlation coefficient r^2_{pred} of 0.90. More importantly, the building GA-RF model also passed through various criteria suggested by Tropsha and Roy with r^2_o and r^2_m values of 0.90 and 0.83, respectively. In order to compare with the GA-RF model, a pure RF model developed based on the full descriptors was performed as well for the same data set. The resulting GA-RF model with significantly internal and external prediction capacities is beneficial to the prediction of potential oxazole and thiazole series of FBPase inhibitors prior to chemical synthesis in drug discovery programs.

Keywords: FBPase inhibitor; chemoinformatics methods; genetic algorithm; random forest

1. Introduction

Diabetes is one of the most prevalent diseases worldwide, and the incidence of this continues to grow, which causes a global public health burden. It is estimated that by 2025, India, China and the United States will possess the largest number of people with diabetes [1]. The more prevalent form, type 2 diabetes, accounts for more than 90% of cases. The pathogenesis of type 2 diabetes is complex, involving progressive development of insulin resistance and a relative deficiency in insulin secretion, which may lead to overt hyperglycemia [2]. Type 2 diabetes is associated with the metabolic syndrome that comprises a set of alterations that include glucose intolerance, truncal obesity, hypertension and dyslipidemia [3]. It has been reported that the metabolic syndrome is associated with a markedly increased risk of coronary artery disease [4]. The risk of myocardial infarction in patients with diabetes and no history of cardiac disease roughly equates the risk in non-diabetic patients with known cardiac diseases [5]. In addition, diabetes often conspires to cause various complications such as eye diseases. It has been documented that diabetes is the primary reason for loss of vision [6]. Patients with diabetes suffer from great mental and physical pain. Consequently, it is necessary to develop an effective drug for the treatment of this disease. Since hyperglycemia leads to severe microvascular and macrovascular complications, the primary treatment goal is to reduce the glucose level. Several popular classes of oral diabetes therapies on the market include sulfonylureas, peroxisome proliferator-activated receptor-γ (PPAR-γ) agonists, metformin and so forth [7,8], which lower glucose by increasing glucose metabolism either via enhanced insulin secretion or improved insulin sensitivity. However, therapies with these drugs still have several drawbacks. For example, sulfonylurea therapy is usually associated with weight gain [9] and metformin therapy is forbidden in patients with renal and hepatic diseases, respiratory insufficiency and alcohol abuse [10]. As a consequence, there is a need for novel, more effective drugs with more safe profiles and fewer side effects to the treatment of diabetes.

Fructose 1,6-bisphosphatase (FBPase), a highly regulated enzyme that catalyzes the second to last step in gluconeogenesis, draws attention as a potential therapeutic target to treat type 2 diabetes mellitus [10,11]. FBPase enables the inhibition of gluconeogenesis from all the corresponding substrates while avoiding direct effects on glycogenolysis, glycolysis and the tricarboxylic acid cycle. In addition, evidence from clinical research suggests that FBPase inhibitors may show an adequate safety margin [11]. Several classes of agents against FBPase have recently been reported including anilinoquinazolines [12], benzoxazole benzenesulfonamides [13], MDL-29951 [14], adenosine 5'-monophosphate (AMP) mimics, *etc*. Some of these series of drugs, however, were explored without successfully achieving acceptable oral bioavailability, thus, it is still required to research novel drugs with desirable characteristics.

Chemoinformatic methods, as a complementary approach of experiment technology, have found wide utility and acceptance, and these methods are playing a central role in drug design [15]. In view of this, previous reports [16] have performed a computational study based on a series of FBPase inhibitors. However, these methods were developed and tested on just a few compounds. In the current work, a larger dataset was used to derive a statistical model with a high prediction power.

Random forest (RF), a new regression tool, has been reported to be a combination of relatively high prediction accuracy and a collection of desired features that makes RF uniquely suited for modeling in chemoinformatics [17] based on a quantitative description of the compound's molecular structure. RF

can show an excellent performance even when most predictive variables are noise, it can be used when the number of variables is much larger than the number of observations, and it returns measures of variable importance. It is well known that an ideal regression model should have a high performance with few descriptors. Thus in the present work, to optimize the Mold^2 molecular descriptor subset [18], with the statistical performance and efficiency of the model being simultaneously enhanced, the genetic algorithm-random forest coupled method is selected to perform a regression task to investigate whether the proposed GA-RF method can construct an ideal prediction model for the FBPase inhibitors. In addition, the derived optimal model was checked by Y-randomization to ensure that the prediction model was not obtained by chance correlation. Moreover, the rigorous statistical criteria suggested by Thropsha and Roy [19,20] were used to validate the model as well. For comparison with the GA-RF, the pure RF model, which means that the model was developed in the full descriptors without variable selection, was also applied using the same dataset.

2. Results and Discussion

2.1. Descriptor Calculation and Preprocessing

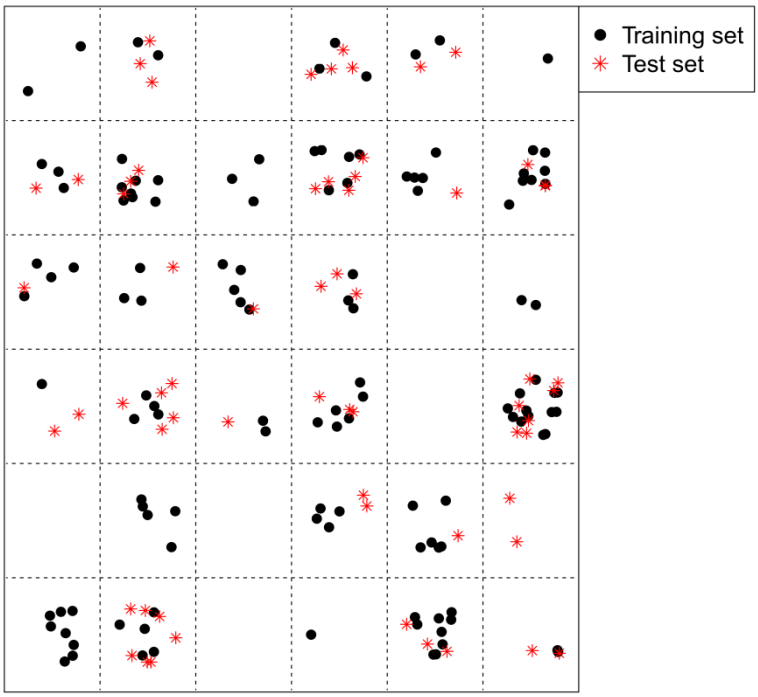
For the quantitative structure-activity relationship (QSAR) investigations, one of the important factors affecting the quality of the model is the choice of molecular descriptors used to obtain the structural information suitable for model development. The software Mold^2 [18] enables a rapid calculation of a large and diverse set of descriptors encoding two-dimensional chemical structure information. A comparative analysis of Mold^2 descriptors with those calculated by commercial tools such as Cerius^2 [21], Dragon [22] on several data sets demonstrated that Mold^2 descriptors can convey a similar amount of information as those widely-used software packages [18]. Although acting as a freely available tool, Mold^2 has been proved suitable not only for the QSAR research [23–25], but also for virtual screening of large databases in drug development [18]. In the present work, a total of 777 Mold^2 descriptors were calculated based on the SDF file format of all the studied FBPase inhibitors. All these descriptors were then preprocessed as follows: (1) descriptors containing values greater than 85% zero were removed; (2) zero- and near zero-variance predictors were removed, as descriptors like this may cause the model to crash or the fit to be unstable; and (3) one of the two descriptors that had absolute correlations above 0.75 was omitted. After these steps, the number of the descriptors was reduced to 108 for further research.

2.2. Split of the Training and Test Sets

Rational selection of training and test sets is also one of the important and challenging steps for the development of validated QSAR models. Self-organizing maps (SOM) can be employed for data survey, which has been successfully applied to split datasets [26,27]. In the current study, in order to probe the descriptor space, a total of 108 Mold^2 descriptors were used to obtain a SOM map. A small Kohonen network with $6 \times 6 = 36$ neurons was employed, producing a map with 36 positions. All the compounds were placed onto the 36 positions of the Kohonen map. Figure 1 demonstrates the distribution of the compounds, where the black dot denotes the training set, while the red asterisk stands for the compounds from the test set. As can be seen from this figure, firstly, the representative

points of the test set are close to those of the training set, and secondly, the training and test sets uniformly fill the whole chemical space, indicating a rational selection of the training and test inhibitors in the present work [28]. The training set was applied for the development of the model and the external test set was used for the assessment of the built model. The training set and test sets include 126 and 64 compounds, respectively.

Figure 1. Self-organizing map (SOM) analysis for fructose 1,6-bisphosphatase (FBPase) inhibitors, where the black dot denotes the training set and the red asterisk stands for the test set.



2.3. Set Parameters of GA-RF Algorithm

In the present work, all the genetic parameters are set as follows: The number of maximum generations is set to 200 and the number of individuals to 50. Individuals are then selected from the population using the stochastic universal sampling algorithm, with a generation gap of 0.9. The double-point crossover is adopted with the probabilities of 0.7. The mutation operation is performed based on the default value included in the genetic algorithm toolbox developed by the Evolutionary Computation Research Team at The University of Sheffield, UK. Herein, the minimum out-of-bag (OOB) mean squared error (MSE) is used as the fitness function to obtain the optimal individual.

2.4. Statistical Results

Apart from the quality of the used data sets, the selection of proper descriptors relevant to the FBPase inhibitory activity is crucial for optimizing the prediction system by reducing the noise in a statistical learning process. After GA-RF, the final 40 Mold^2 descriptors are selected. Table 1 illustrates the names of these selected descriptors and the corresponding definitions [18].

Table 1. Molecular descriptors selected from genetic algorithm-random forest coupled method (GA-RF) for the FBPase inhibitors.

Name	Definition	Name	Definition
D004	Number of 05-membered rings	D543	Lowest eigenvalue from Burdex matrix weighted by van der Waals order-4
D016	Number of double bonds	D545	Lowest eigenvalue from Burdex matrix weighted by van der Waals order-6
D152	Mean atomic polarizability scaled on carbon-SP3	D547	Lowest eigenvalue from Burdex matrix weighted by van der Waals order-8
D164	Index of terminal vertex matrix	D557	Lowest eigenvalue from Burden matrix weighted by polarizabilities order-2
D237	Kier 3-path index	D561	Lowest eigenvalue from Burden matrix weighted by polarizabilities order-6
D279	Total information content order-4 index	D562	Lowest eigenvalue from Burden matrix weighted by polarizabilities order-7
D309	Sum eigenvalue weighted by mass distance matrix	D563	Lowest eigenvalue from Burden matrix weighted by polarizabilities order-8
D455	Geary topological structure autocorrelation length-1 weighted by atomic van der Waals volumes	D571	Highest eigenvalue from Burden matrix weighted by masses order-8
D458	Geary topological structure autocorrelation length-4 weighted by atomic van der Waals volumes	D582	Highest eigenvalue from Burden matrix weighted by electronegativities Sanderson-scale order -3
D462	Geary topological structure autocorrelation length-8 weighted by atomic van der Waals volumes	D589	Highest eigenvalue from Burden matrix weighted by polarizabilities order-2
D465	Geary topological structure autocorrelation length-3 weighted by atomic Sanderson electronegativities	D598	Number of total tertiary carbon-SP3
D470	Geary topological structure autocorrelation length-8 weighted by atomic Sanderson electronegativities	D647	Number of group primary amines (aliphatic)
D473	Geary topological structure autocorrelation length-3 weighted by atomic polarizabilities	D715	Number of group CH2R2
D476	Geary topological structure autocorrelation length-6 weighted by atomic polarizabilities	D719	Number of group CH2RX
D491	Moran topological structure autocorrelation length-5 weighted by atomic van der Waals volumes	D729	Number of group =CHR
D492	Moran topological structure autocorrelation length-6 weighted by atomic van der Waals volumes	D731	Number of group =CHX
D499	Moran topological structure autocorrelation length-5 weighted by atomic Sanderson electronegativities	D746	Number of group H attached to C0(sp3) no X attached to next C

Table 1. *Cont.*

Name	Definition	Name	Definition
D506	Moran topological structure autocorrelation length-4 weighted by atomic polarizabilities	D754	Number of group O=
D523	Mean molecular topological order-3 charge index	D756	Number of group Al-O-Ar or Ar-O-Ar or R-O-C=X
D541	Lowest eigenvalue from Burden matrix weighted by van der Waals order-2	D775	Hydrophilic factor index

Based on the determined optimal parameters by GA, the GA-RF model presents an (root-mean-square error) RMSE of 0.25 and 0.34 for the training and test sets, respectively. The determined coefficient r^2_{ncv} reaches a value as high as 0.96 with r^2_{cv} = 0.67 for the training set. The model predictability is evaluated by an external prediction set, which illustrates r^2_{ts} and r^2_{pred} values of 0.91 and 0.90, respectively. It is well known that the random forest algorithm can manipulate the data set even with a large number of descriptors [17]. Thus we compare the GA-RF with pure RF, which means that the latter model is built based on the whole 108 descriptors. It can be seen from Table 2 that the pure RF performs comparably but relatively low statistics compared with GA-RF. The scatter plots of the experimental *versus* predicted FBPase inhibitory activity based on the GA-RF and RF models are shown in Figure 2, where the proposed GA-RF model presents a relatively better performance than RF. For the former, the data points of training and test sets distribute more closely in a straight line ($y = x$), indicating that GA-RF exhibits both the inner and external prediction power. Table 3 gives the experimental and predicted results for both models.

Table 2. Statistical performances of GA-RF and RF models [a].

Model	Training Set			Test Set		
	r^2_{ncv}	r^2_{cv}	RMSE	r^2_{ts}	r^2_{pred}	RMSE
GA-RF	0.96	0.67	0.25	0.91	0.90	0.34
RF	0.96	0.59	0.28	0.87	0.85	0.42

[a] r^2_{cv} from OOB estimation; m_{try} is equal to 13 and 36 for GA-RF and RF, respectively.

Figure 2. The scatter plots of actual and predicted activity by GA-RF and RF models.

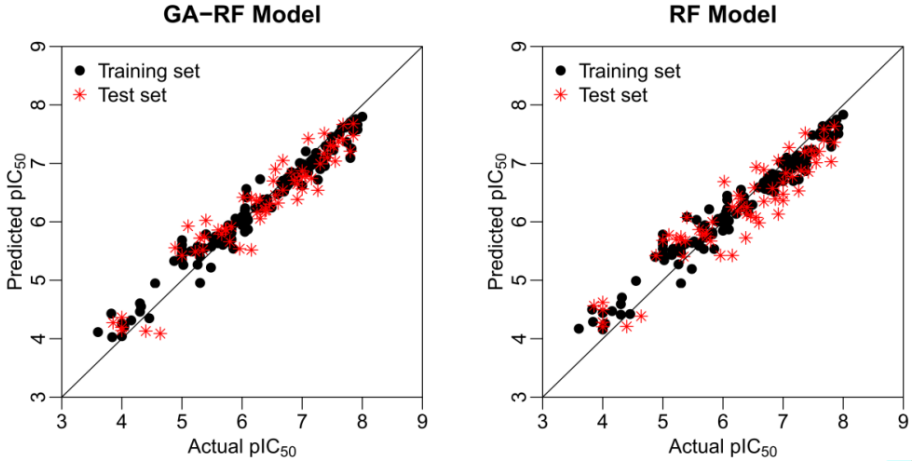
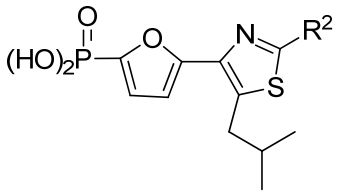
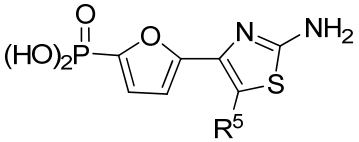


Table 3. Compounds with their chemical names, observed and predicted activities by GA-RF and RF for the FBPase inhibitors.

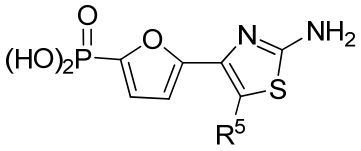


No.	R^2	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
1	Me	7.00	6.62	6.70	[29]
2 *	Et	6.40	6.20	6.25	[29]
3	vinyl	5.92	5.99	6.05	[29]
4	CH_2OH	6.66	6.63	6.61	[29]
5 *	H	6.30	6.05	6.27	[29]
6	Cl	6.74	6.61	6.64	[29]
7	Br	7.10	6.85	6.89	[29]
8	SMe	6.05	6.23	6.16	[29]
9	CN	5.70	5.65	5.73	[29]
10 *	NH_2	7.60	7.38	7.20	[29]
11	NHMe	6.00	5.95	6.08	[29]
12	NHAc	5.00	5.69	5.66	[29]
13	$CONH_2$	5.56	5.75	6.03	[29]
14 *	$CSNH_2$	6.30	6.38	6.39	[29]
15	Ph	4.87	5.33	5.40	[29]
16 *	2-thienyl	5.10	5.93	5.78	[29]
17	3-pyridyl	5.30	5.40	5.55	[29]

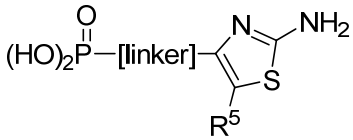


No.	R^5	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
18	H	6.35	6.38	6.42	[29]
19	Me	6.92	6.84	6.80	[29]
20	$HOCH_2$	6.30	6.73	6.55	[29]
21 *	*n*-Pr	7.52	7.29	7.13	[29]
22 *	*i*-Pr	7.55	7.04	7.01	[29]
23	CF_3CH_2	7.24	6.99	7.14	[29]
24	neopentyl	7.92	7.58	7.51	[29]
25	cyclobutyl	7.72	7.61	7.54	[29]
26 *	cyclopentyl	7.68	7.67	7.58	[29]
27	cyclohexyl	8.00	7.80	7.83	[29]
28	cyclopropyl-CH_2	7.70	7.62	7.53	[29]
29	cyclopentyl-CH_2	7.74	7.36	7.44	[29]
30	cyclohexyl-CH_2	7.23	7.18	7.08	[29]
31	$PhCH_2$	6.82	6.85	6.82	[29]
32 *	morpholinyl-CH_2	6.25	6.16	6.45	[29]

Table 3. *Cont.*

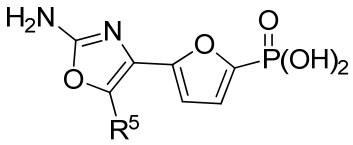


No.	R^5	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
33	Cl	7.15	7.03	6.97	[29]
34 *	Br	7.30	6.99	6.88	[29]
35 *	I	7.00	6.87	6.36	[29]
36	1-morpholinyl	7.80	7.09	7.29	[29]
37	EtS	7.48	7.32	7.24	[29]
38 *	*n*-PrS	7.80	7.21	7.03	[29]
39	*i*-PrS	7.62	7.50	7.46	[29]
40	*t*-BuS	7.62	7.52	7.53	[29]
41 *	PhS	6.52	6.70	6.58	[29]
42	$CONMe_2$	5.77	5.94	6.22	[29]
43	CO_2Et	7.85	7.55	7.48	[29]
44	CO_2Bn	7.82	7.25	7.43	[29]
45	*n*-PrSO	6.07	6.56	6.45	[29]
46 *	Ph	7.85	7.68	7.64	[29]
47 *	2-MeO-Ph	7.37	7.51	7.52	[29]
48	3-MeO-Ph	7.68	7.60	7.62	[29]
49	4-MeO-Ph	7.66	7.61	7.64	[29]
50 *	4-MeS-Ph	7.68	7.41	7.40	[29]
51	4-*t*-Bu-Ph	7.06	7.21	7.10	[29]
52 *	4-MeO_2C-Ph	7.85	7.48	7.36	[29]
53	4-F-Ph	7.80	7.71	7.68	[29]
54	4-Cl-Ph	7.89	7.76	7.75	[29]
55	4-Ac-Ph	7.49	7.45	7.48	[29]
56	4-$MeSO_2$-Ph	7.39	7.30	7.00	[29]
57 *	4-Ph-Ph	7.47	7.31	7.23	[29]
58	2-nathphyl	7.92	7.66	7.61	[29]
59	2-furanyl	7.40	7.12	7.22	[29]
60 *	2-thienyl	7.36	7.17	7.20	[29]

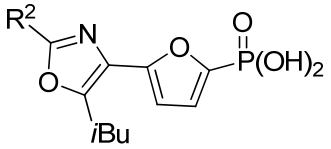


No.	[linker]	R^5	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
61	2,5-furanyl	H	5.00	5.41	5.78	[29]
62	-CH_2OCO-	*n*-Pr	7.30	6.90	6.92	[29]
63 *	-CH_2NHCO-	2-thienyl	6.02	6.42	6.69	[29]
64	2,6-pyridyl	H	5.70	5.74	5.94	[29]
65	1,3-phenyl	H	5.89	6.06	6.01	[29]
66 *	1,3-phenyl-(6-Me)	*n*-Pr	6.87	6.71	6.39	[29]
67 *	1,3-phenyl-(6-OMe)	*i*-Pr	6.68	7.05	6.89	[29]
68 *	1,3-phenyl-(6-F)	Ph	7.10	7.42	7.27	[29]

Table 3. *Cont.*

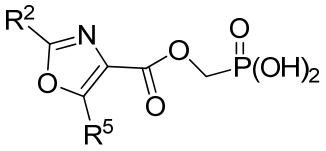


No.	R^5	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
69 *	*i*-Bu	6.92	6.38	6.13	[30]
70	H	5.00	5.60	5.43	[30]
71	Allyl	6.85	6.70	6.51	[30]
72	*n*-Bu	6.77	6.64	6.54	[30]
73*	*n*-Pentyl	6.68	6.54	6.35	[30]
74	-CH_2-cyclohexyl	6.49	6.24	6.29	[30]
75	Ph	6.80	6.78	6.80	[30]
76	Bn	6.05	6.23	6.12	[30]
77	-CH_2-(2-thienyl)	6.59	6.47	6.59	[30]
78	*n*-PrS	7.15	6.97	6.91	[30]
79	*i*-PrS	6.96	7.01	7.02	[30]
80 *	*t*-BuS	6.92	6.64	7.05	[30]
81	PhS	5.40	5.79	6.08	[30]
82	-CO_2Me	7.17	7.06	6.70	[30]
83 *	-CO_2Et	7.42	7.10	6.85	[30]
84	-CO_2Pr-*i*	7.40	7.13	7.14	[30]
85	-CO_2Bn	7.07	6.95	6.91	[30]
86	-COSEt	7.52	7.23	7.20	[30]
87	-COBu-*t*	6.07	6.10	6.22	[30]

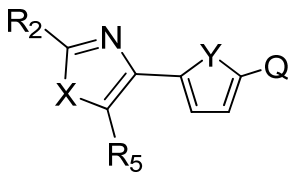


No.	R^2	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
88	Me	6.22	6.24	6.14	[30]
89	HO	5.00	5.48	5.43	[30]
90 *	H	5.72	5.87	5.80	[30]
91	Me_2N-	5.68	5.61	5.54	[30]
92 *	*i*-Pr-	5.66	5.79	5.78	[30]
93	MeHN-	5.37	5.55	5.62	[30]
94	Et	6.02	6.09	5.94	[30]
95 *	EtHN-	5.00	5.43	5.68	[30]
96	vinyl	5.17	5.49	5.54	[30]

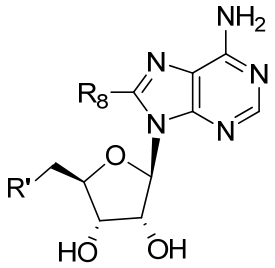
Table 3. *Cont.*



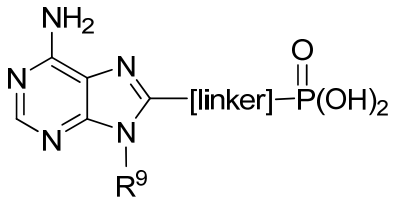
No.	R^2	R^5	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
97	H_2N-	H	5.15	5.50	5.43	[30]
98 *	H_2N-	Me	6.38	6.19	5.72	[30]
99	H_2N-	Et	6.42	6.39	6.17	[30]
100 *	H_2N-	*n*-Pr	6.55	6.46	6.08	[30]
101 *	H_2N-	*i*-Pr	6.24	6.36	6.19	[30]
102 *	H_2N-	*n*-Bu	6.60	6.32	5.98	[30]
103 *	H_2N-	*n*-Pent	6.46	6.30	6.10	[30]
104	Me	CF_3	5.00	5.45	5.54	[30]
105	H	Ph	5.00	5.35	5.45	[30]



No.	X	Y	Q	R^2	R^5	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
106	NH	O	PO_3H_2	NH_2	*i*Bu	5.30	5.55	5.47	[31]
107	S	O	PO_3H_2	H	H	5.26	5.58	5.56	[31]
108	CH=CH	O	PO_3H_2	NH_2	Ph	7.38	6.95	6.87	[31]



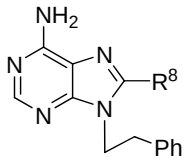
No.	R^8	R'	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
109 *	-$NH(CH_2)_2PO_3H_2$	OH	4.00	4.36	4.62	[32]
110 *	-$NH(CH_2)_2OPO_3H_2$	OH	3.85	4.27	4.56	[32]
111	-$NH(CH_2)_2PO_3H_2$	H	4.00	4.27	4.44	[32]



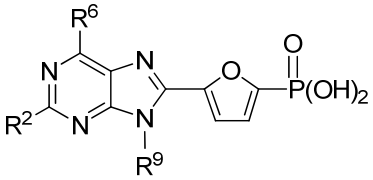
No.	[linker]	R^9	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
112	-$NH(CH_2)_2$-	Bn	4.04	4.20	4.26	[32]
113 *	-$NH(CH_2)_2$-	$Ph(CH_2)_2$-	4.00	4.14	4.20	[32]
114	-$NH(CH_2)_2$-	2-naphthyl-CH_2-	4.46	4.35	4.42	[32]
115 *	-$CONHCH_2$-	$Ph(CH_2)_2$-	4.00	4.19	4.50	[32]

Table 3. *Cont.*

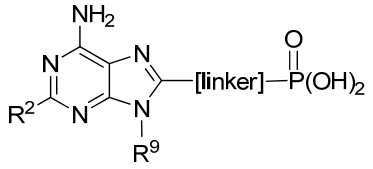
116	-$(CH_2)_3$-	Ph$(CH_2)_2$-	4.00	4.04	4.16	[32]
117 *	-CH=$CHCH_2$-	Ph$(CH_2)_2$-	4.00	4.19	4.28	[32]
118	-$S(CH_2)_2$-	Ph$(CH_2)_2$-	3.84	4.03	4.29	[32]
119 *	-CH_2OCH_2-	Ph$(CH_2)_2$-	4.64	4.09	4.38	[32]
120	-2,5-furanyl-	Ph$(CH_2)_2$-	5.30	4.95	4.95	[32]
121	-2,5-thienyl-	Ph$(CH_2)_2$-	4.32	4.55	4.71	[32]



No.	R^8	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
122 *	-$(CH_2)_2$-$OPO(OH)_2$	4.40	4.13	4.21	[32]
123	-2,5-furanyl-SO_3H	3.82	4.43	4.50	[32]

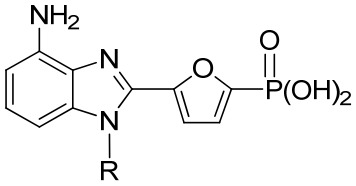


No.	R^2	R	R^9	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
124	H	-$N(Me)_2$	-$(CH_2)_2$Ph	3.60	4.11	4.17	[32]
125	H	-NHMe	-$(CH_2)_2$Ph	4.30	4.46	4.41	[32]
126	H	Cl	-$(CH_2)_2$Ph	4.30	4.61	4.59	[32]
127	H	-NH_2	-$CH_2CH(Ph)_2$	4.15	4.31	4.47	[32]
128	H	-NH_2	-$(CH_2)_2$(cyclohexyl)	5.85	5.54	5.54	[32]
129	H	-NH_2	-(CH_2)(2-naphthyl)	5.48	5.22	5.19	[32]
130	H	-NH_2	cyclopropyl	5.82	5.70	5.78	[32]
131	H	-NH_2	cyclopentyl	5.70	5.69	5.76	[32]
132	H	-NH_2	Et	5.74	5.65	5.76	[32]
133	H	-NH_2	isobutyl	5.82	5.81	5.82	[32]
134	H	-NH_2	neopentyl	6.10	5.87	5.87	[32]
135 *	-SMe	-NH_2	isobutyl	6.15	5.52	5.42	[32]
136	-SO_2Me	-NH_2	isobutyl	4.55	4.95	4.99	[32]

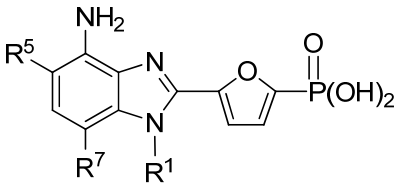


No.	R^2	R^9	[linker]	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
137 *	H	-$CH_2C(Me)_2CH_2OH$	2,5-furanyl	5.35	5.52	5.40	[32]
138	H	-$CH_2C(Me)_2CH_2Cl$	2,5-furanyl	6.05	5.83	5.91	[32]
139 *	H	-$CH_2C(Me)_2CMe_3$	2,5-furanyl	5.80	5.66	5.67	[32]
140 *	H	-$CH(Me)CMe_3$	2,5-furanyl	5.30	5.72	5.74	[32]
141	-NH_2	-CH_2CMe_3	2,5-furanyl	5.26	5.27	5.27	[32]
142 *	-SMe	-CH_2CMe_3	2,5-furanyl	5.96	5.54	5.42	[32]
143 *	H	-CH_2CMe_3	2,5-(3,4-di-Cl)furanyl	4.89	5.56	5.42	[32]

Table 3. *Cont.*

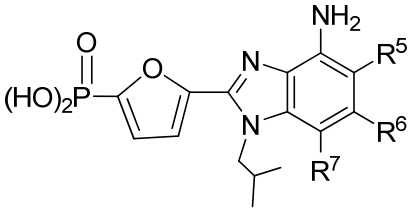


No.	R	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
144 *	Me	5.22	5.49	5.70	[33]
145	Et	5.65	5.80	5.88	[33]
146	*n*Pr	5.96	6.00	6.03	[33]
147 *	*i*Bu	5.82	5.91	6.00	[33]
148	cyclllopropyl-CH_2-	6.10	6.03	6.02	[33]
149	cyclobutyl-CH_2-	6.10	6.04	6.01	[33]
150	cyclopentyl-CH_2-	5.82	5.87	5.81	[33]
151	cyclohexyl-CH_2-	5.60	5.64	5.63	[33]
152	cycloheptyl-CH_2-	5.49	5.57	5.64	[33]
153	norbornyl	6.00	5.94	5.85	[33]
154 *	benzyl	5.30	5.72	5.70	[33]
155	4-*t*Bu-benzyl	5.02	5.26	5.34	[33]
156	4-CF_3-benzyl	5.15	5.50	5.51	[33]
157	4-Ph-benzyl	5.60	5.63	5.59	[33]
158 *	3-furanyl-CH_2-	5.38	5.74	5.68	[33]
159 *	3-HO-benzyl	5.73	5.87	5.75	[33]
160 *	3-thienyl-CH_2-	5.40	6.02	6.08	[33]

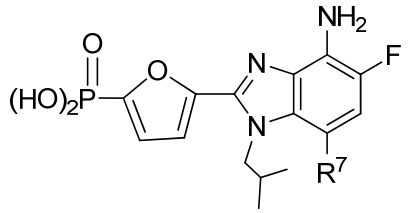


No.	R^1	R^5	R^7	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
161 *	*i*Bu	Et	H	5.60	5.86	5.87	[33]
162	*i*Bu	*n*Pr	H	5.52	5.69	5.66	[33]
163 *	*i*Bu	MeO	H	6.15	6.41	6.25	[33]
164	*i*Bu	OH	H	6.30	6.23	6.24	[33]
165	*i*Bu	Cl	H	6.70	6.52	6.56	[33]
166	*i*Bu	H	Cl	6.05	6.19	6.11	[33]
167	*i*Bu	Br	H	6.40	6.32	6.29	[33]
168 *	*i*Bu	H	Br	6.40	6.21	6.14	[33]
169 *	*i*Bu	F	H	7.00	6.56	6.47	[33]
170 *	$(Et)_2CHCH_2$-	F	H	6.82	6.83	6.57	[33]
171	$(Et)_2CH$-	F	H	6.07	6.42	6.38	[33]
172 *	*c*Pr-CH_2-	F	H	7.26	6.54	6.53	[33]

Table 3. *Cont.*



No.	R^5	R^6	R^7	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
173	Br	H	Br	6.00	6.09	6.03	[33]
174	Cl	H	Cl	6.35	6.34	6.30	[33]
175 *	F	H	Cl	7.00	6.77	6.67	[33]
176	F	H	Br	6.89	6.75	6.67	[33]
177	F	Cl	H	6.65	6.66	6.59	[33]
178	Br	Cl	Cl	5.00	5.53	5.60	[33]
179 *	F	H	vinyl	6.55	6.90	6.94	[33]
180	F	H	*c*Pr	7.22	7.08	7.10	[33]



No.	R^7	Obs. pIC_{50}	GA-RF	RF	Ref. [a]
181 *	Ph	7.05	6.79	6.83	[33]
182	4-F-Ph	6.74	6.74	6.75	[33]
183	4-Cl-Ph	7.05	6.94	6.81	[33]
184	Et	7.26	7.07	7.06	[33]
185	*n*Pr	7.00	6.99	7.02	[33]
186	tBu$(CH_2)_2$-	6.68	6.70	6.67	[33]
187	$(Me)_2CH(CH_2)_3$-	7.00	6.92	6.99	[33]
188	$HO(CH_2)_3$-	7.10	6.97	7.05	[33]
189	$(Me)_2N(CH_2)_3$-	7.26	6.72	6.73	[33]
190 *	$Cl(CH_2)_4$-	7.15	6.74	6.78	[33]

* test set; [a] from the corresponding references.

2.5. Further Test for the External Prediction Power

To validate firmly the performances of the prediction, the squared correlation coefficient values between the observed and predicted values for the test set compounds with intercept (r^2_{ts}) and without intercept (r^2_o) are also calculated. Table 4 presents the values of the parameters for all models in the present work. According to references [19,34], models are considered acceptable if they satisfy all the following conditions: (1) $r^2_{pred} > 0.5$, (2) $r^2_{ts} > 0.6$, and (3) r^2_o is close to r^2_{ts}, such that the $[(r^2_{ts} - r^2_o)/r^2_{ts}] < 0.1$ and $0.85 \leq k \leq 1.15$. When the predicted values of the test set compounds (x axis) are plotted against the observed values of the compounds (y axis) with the intercept set to zero, the slope of the fitted line gives the value of k, with the corresponding correlation coefficient r^2_o. Herein, r^2_o is a quantity characterizing linear regression with the y-intercept set to zero, which can be illustrated by

y = kx; while r^2_{ts} is the conventional coefficient of determination for the best fit linear regression (*i.e.*, denoted by y = ax + b) in the test set [19,28]. It can be noticed that the developed GA-RF and pure RF models fully satisfy all the requirements, but the latter is relatively less accurate than GA-RF.

Table 4. External predictability of GA-RF model.

Model	r^2_{ts}	r^2_{pred}	r^2_o	$(r^2_{ts} - r^2_o)/r^2_{ts}$	k	r^2_m
GA-RF	0.91	0.90	0.90	0.01	1.01	0.83
RF	0.87	0.85	0.85	0.02	1.01	0.76

Roy *et al*. have reported [35] that the r^2_{pred} may, sometimes, not truly reflect the predictive capability of a model on a new dataset. Also, the squared regression coefficient r^2_{ts} between the observed and predicted values of the test set compounds does not necessarily mean that the predicted values are very near to the observed activities. To better evaluate the external predictive capacity of a model, a modified r^2_{ts} term, r^2_m, is suggested as follows [20]:

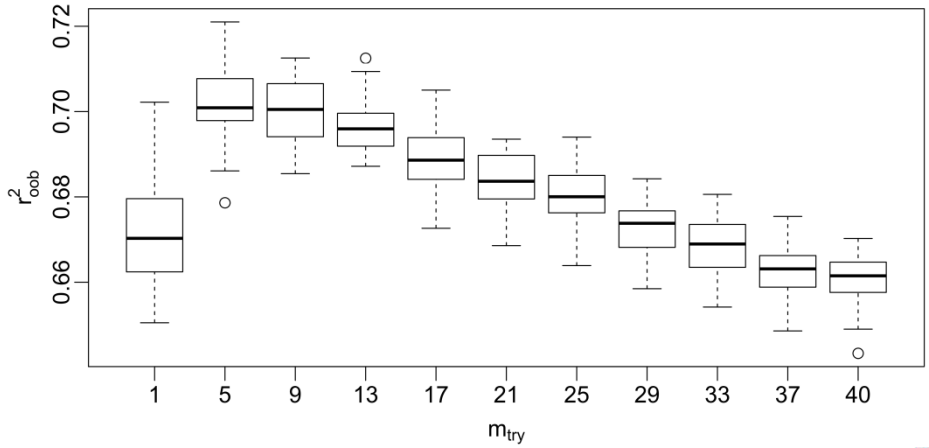
$$r_m^2 = r_{ts}^2 \times (1 - \sqrt{r_{ts}^2 - r_o^2}) \quad (1)$$

In case of good external prediction capacity, predicted values will be very close to the observed ones and thus the r^2_{ts} will be very near to the r^2_o. In the best case r^2_m may be equal to r^2_{ts}, whereas in the worst case the r^2_m value could be zero. Herein, the built GA-RF model achieves a better r^2_m value of 0.83 than RF (0.76), which illustrates that the current model possesses a highly predictive power.

2.6. Investigation of Parameter Turning on the GA-RF Model

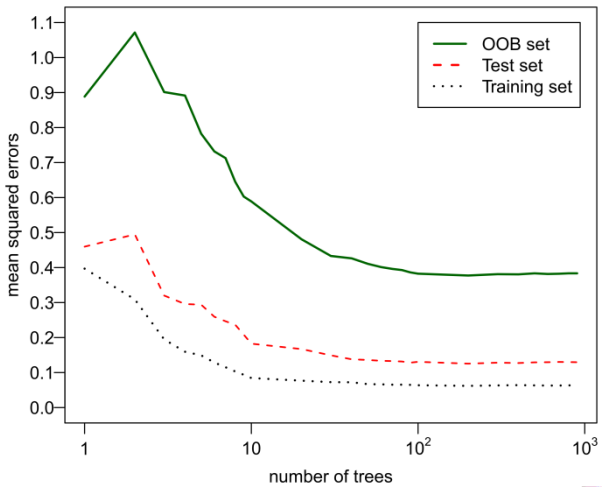
As seen from Tables 2 and 4, since the proposed GA-RF model illustrates relatively better statistical results than the pure RF model, the following analysis is only restrict to this model. Generally, RF has effectively only one tuning parameter, m_{try}, which is the number of the descriptors randomly sampled as candidates for splitting at each node during the tree induction. It ranges from 1 to p, the total number of descriptors available, in which p is equal to 40. Although it has been reported [17] that RF still performs well using the default m_{try} value (p/3), one still expects to investigate the effect of parameter turning. Herein, 50 replications of OOB estimation (r^2_{oob}) based on the FBPase inhibitors are performed, with the purpose to assess the correlation between the actual and predicted data with a range of m_{try} values, including the default value, which is equal to 13 for the current study. Figure 3 shows the boxplot of these correlations. This plot suggests that m_{try} is optimal when near five with a median value of 0.701, while the default m_{try} gives a median value of 0.696. Both results are comparable. It is also observed that the worst statistical results are derived from m_{try} = 1 and = 40. The observation is in agreement with the previous report [17]. From this Figure, one can notice that it is necessary to perform a moderate parameter tuning to get the optimal one, although at most times, RF can give the optimal model by using default parameters.

Figure 3. Boxplot of 50 replications of OOB estimation (r^2_{oob}) at various values of m_{try}. Horizontal lines inside the boxes are the median correlation.



Besides the m_{try}, the number of trees also has an effect on the RF performance [17]. One principle of building RF model is to ensure that there are sufficient trees in the forest in order to get enough training of each sample. To illustrate this, the performances of OOB set, test set and training set are compared with the increase of the number of trees. Figure 4 shows that similar tendency exists for the tracks of the OOB mean squared errors, the test set and the training set ones, once there are a sufficient number of trees. In the present work, 100 trees are enough to build RF model. The information obtained from this Figure is that mean squared errors of the test and OOB do not increase after the mean squared errors of training set reach the minimum; instead, they converge to their asymptotic values which are also close to their minimum. In this sense, it can be concluded that RF does not overfit, which has been supported by the previous reports [17,23].

Figure 4. Comparison of mean squared errors from out-of-bag (OOB) set, test set and training set as the number of trees increases for FBPase inhibitors.



2.7. Y-Randomization Check

Presently, the Y-randomization check [34] is implemented for further assurance of the robustness of the optimal GA-RF model. The dependent variable is randomly shuffled and a new QSAR model is

developed using the original independent variable matrix. The new QSAR models (after several repetitions) are expected to possess low r^2_{ncv}, r^2_{cv}, r^2_{ts}, r^2_{pred}, r^2_m and high RMSE for the training and test sets, respectively. If the opposite happens, then an acceptable QSAR model cannot be obtained for the specific modeling method and data. In the current work, the Y-randomization check is repeated 500 times and the resulting statistics are compared with the prediction statistics without such checks, with the average values reported in Table 5. As shown in this Table, the correlation coefficients have a significant decline while the RMSE values sharply increase, which indicates that the proposed GA-RF model possesses a real prediction power, and the result is not due to a chance correlation.

Table 5. Comparison with and without Y-randomization check of the optimal GA-RF model.

Model	Training Set			Test Set			
	r^2_{ncv}	r^2_{cv}	RMSE	r^2_{ts}	r^2_{pred}	r^2_m	RMSE
GA-RF [a]	0.96	0.67	0.25	0.91	0.90	0.83	0.34
GA-RF [b]	0.01	−0.14	1.27	0.06	−0.10	0.04	1.13

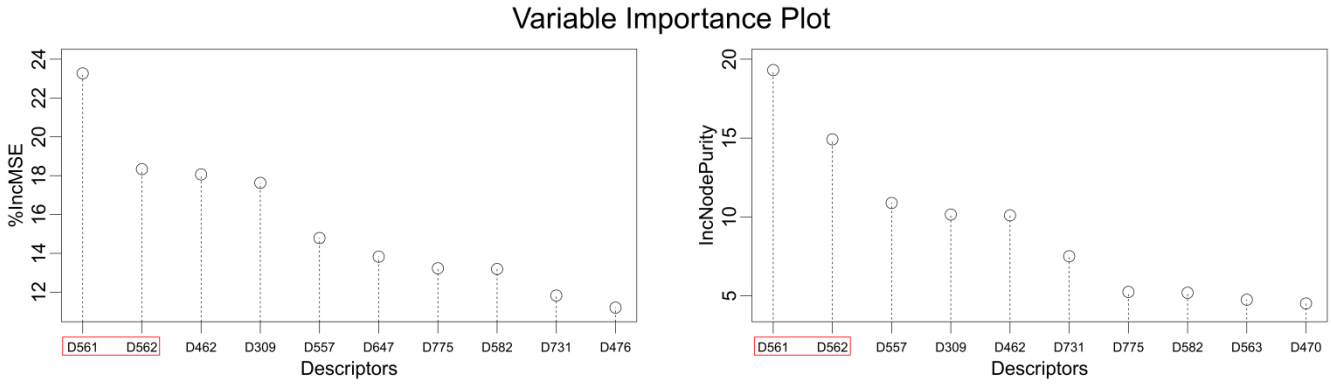
[a] without Y-randomization check; [b] with Y-randomization check.

2.8. Explanation of the Selected Descriptors

The ideal QSAR model would be robust, sparse, predictive, and interpretable. In many cases such an ideal is not achievable with current descriptors and the corresponding variable mapping methods, although much effort is being expended in approaching this ideal. In most QSAR researches, a full direct explanation for all the descriptors is difficult, where most similar reported works all give few detailed analyses of the descriptors involved in their model development, thus only a few descriptors in this work are explained. Generally speaking, QSAR can be classified into two categories: Interpretative QSAR and predictive QSAR. For the current work, it belongs to the latter. In spite of this fact, we still attempt to offer some rational explanations for the major descriptors using RF built-in variable importance measure technology. Figure 5 depicts the variable importance plot of the GA-RF model. Herein, there are two parameters that give the definitions of the variable importance measures: (1) Mean Decrease Accuracy (%IncMSE) and Mean Decrease Gini (IncNodePurity). The higher values of these two parameters represent the higher variable importance. For more details about these parameters, please refer to the corresponding literature [36,37]. In Figure 5, it can be seen that the first two most important descriptors are D561 and D562 surrounded by the red frames. D561 refers to the lowest eigenvalue from Burden matrix weighted by polarizabilities order-6, while D562 stands for the lowest eigenvalue from Burden matrix weighted by polarizabilities order-7. Both descriptors illustrate that molecular polarizabilities play a central role in FBPase inhibitory activity, which can be supported by the experimental results [29–33]. Previous literature has illustrated that the phosphate group forms a constellation of hydrogen bond interactions that are essential for binding affinity [32]. In the present work, it can be observed that all the studied compounds possess the phosphate group and most of them have oxazole and thiazole groups as well as -NH_2 substituents, which increase the molecular polarity. In addition, previous reports have also depicted the similar conclusion that the polar groups are favored to increase the FBPase inhibitory activity [16,38]. As suggested in the literature that the FBPase binding pocket presents the hydrophilic nature [11] and the polar groups of inhibitors will bind to the

site, leading to potent inhibition of the enzyme. The information obtained by the current work, to some degree, provides an insight into the structural features of both oxazole and thiazole FBPase inhibitors from a theoretical point of view, which should be helpful to design new FBPase inhibitors of this series for the treatment of type 2 diabetes. It should be pointed out that herein we just present a representative explanation of the selected descriptors. However, in terms of developing a highly predictive model, the proposed GA-RF model in this work could implement this task.

Figure 5. Variable importance plot from GA-RF. The first two important descriptors are surrounded by red frames.



3. Experimental Section

3.1. Dataset

A large, diverse dataset of 190 FBPase inhibitors were collected from the literature [29–33] published by Dang and co-workers after removing duplicated and undesirable compounds. Here the converted molar pIC_{50} ($-\log IC_{50}$) values, ranging from 3.60 to 8.00 M, were used as the dependent variables in the QSAR regression analysis to improve the normal distribution of the experimental data points. The whole data set was divided into training (126) and test (64 molecules) sets, respectively. All structures and the corresponding activity values of the dataset as well as their belongings to the training and test sets are listed in Table 3.

3.2. Descriptor Calculation

A rational design of novel lead drug is getting more and more popular [39]. QSAR, one of the most frequent drug design methods, can build a bridge between molecular descriptors and statistical methods to predict the new compounds. In the present work, all two-dimensional structures of the dataset were built with the ISIS/Draw 2.3 program, and converted SDF format by the Open Babel software package [40]. The final structures were transferred into Mold^2 [18], a free program available to the public, to calculate molecular descriptors. The Mold^2 software package can calculate 777 molecular descriptors solely from 2D chemical structures. Hong *et al*. have reported that the models generated using Mold^2 descriptors were comparable to those obtained using descriptors from the commercial software packages. In our work, all original 777 molecular descriptors were calculated.

3.3. Computational Methods

GA: Genetic algorithm is derived from Darwin's theory of natural selection and evolution. Based on the Darwinian principle of survival of the fittest, GA obtains the optimal solution after a series of iterative computations including selection or reproduction, crossover or recombination, and mutation. Due to its highly efficient optimization algorithm, GA has already been successfully applied in many QSAR analyses [41–44] to perform variable selection. In the present work, the binary coding form of each chromosome was adopted with 1 and 0 representing selected and non-selected descriptors, respectively. The double point crossover was employed in our study [45–47], which allowed new solution regions in the search space to be explored in order to get the global optimum. In binary code genes, the code may be changed from 0 to 1 or vice versa through mutation operation. As a last step, the old population was replaced by children, and a new generation was produced. The evolutionary process operated for many generations until the termination condition was satisfied [47]. For the detailed methodology about GA, please refer to the corresponding literature [48,49].

RF: Random forest is an ensemble of single decision trees, producing a corresponding number of outputs, and the outputs of all trees are aggregated to obtain one final prediction. The training algorithm of RF for regression can be briefly summarized as follows [17,25]: (1) draw N bootstrap samples from the original training set; (2) construct an un-pruned tree T_p (p = 1, …, N) with each training set B_p. At each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and then choose the best split from among those variables. The tree is grown to the maximum size and not pruned back; (3) predict the N trees by average for regression. The tree growing algorithm used in RF is CART which is efficient especially when the number of descriptors is very large, with the reason being that RF only tests the m_{try} of the descriptors rather than the whole one, where the default m_{try} is one-third of the number of descriptors for regression. Since the number of m_{try} is very small, the search can finish quickly.

RF possesses its own reliable statistical characteristics based on OOB set prediction, which could be used for validation and model selection with no cross-validation performed. It was shown that the prediction accuracy of an OOB set and a five-fold cross validation procedure was nearly the same [17]. Although RF performs relatively well "off the shelf" without expending much effort on the parameter tuning or variable selection [17], it is also of importance for carrying out some tentative investigations on the changes of m_{try} or descriptor selection to optimize the performance of RF. Herein, we just present a brief introduction of RF, for more details please see the corresponding literature [17,50]. It has been reported that RF can show excellent performance even when most predictive variables are noise, and it can be used when the number of variables is much larger than the number of observations, and it returns measures of variable importance [17,50]. However, to obtain an ideal regression model, a variable selection process is still required. To achieve the above objective, in this work, the GA variable selection method using OOB MSE as the fitness function was carried out to achieve the regression task for the current FBPase inhibitors in order to yield a high prediction model.

3.4. Statistical Validation

In the current study, the selected descriptors served as independent variables and the pIC_{50} values as dependent variables in the RF regression analysis. The inner predictive values of the models were evaluated first by a cross-validation process [51,52]. The cross-validated coefficient, r^2_{cv}, was calculated using Equation (2):

$$r^2_{cv} = 1 - \frac{\sum_{i=1}^{train}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{train}(y_i - \bar{y}_{tr})^2} \quad (2)$$

where y_i, $\hat{y}_i$, and $\bar{y}_{tr}$ are the observed, predicted, and mean values of the target property, respectively, for the training set. Herein, $\sum_{i=1}^{train}(y_i - \hat{y}_i)^2$ is the predictive residual sum of squares (PRESS). The optimal number of components obtained from the cross-validation was used to derive the final QSAR model. Then, a non-cross-validation analysis was carried out; and the Pearson coefficient (r^2_{ncv}) and RMSE were calculated.

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

where n denotes the number of the studied compounds.

It has been reported [19] that although the low value of r^2_{cv} for the training set can exhibit a low predictive ability of a model, the opposite is not necessarily true. That is, a high r^2_{cv} is necessary, but not sufficient, for a model with a high predictive power. Therefore, the external validation must be estimated to establish a reliable and predictive QSAR model. The predictive coefficient r^2_{pred} listed in the following equation was used to check the models. In addition, various criteria suggested by Tropsha and Roy [19,20] were also performed to validate the predictive power of the current built models.

$$r^2_{pred} = 1 - ("PRESS"/SD) \quad (4)$$

where SD is the sum of the squared deviations between the actual activity of the compounds in the test set and the mean activity in the training set, and "PRESS" is the sum of the squared deviations between predicted and observed activity for each compound in the test set.

4. Conclusions

In the present work, a GA-RF algorithm is successfully proposed as an efficient chemoinformatic method to predict FBPase inhibitory activity. The GA-RF model went through all rigorous examinations suggested by Tropsha and Roy with r^2_{pred} of 0.90 and r^2_m of 0.83, exhibiting its feasibility and reliability to derive a highly predictive model for FBPase inhibitors. In addition, results from a Y-randomization check illustrate that the GA-RF model possesses real prediction power not due to chance correlation. Explanation of the selected descriptors by GA-RF suggests that the polar factors play a central role in the FBPase inhibition. Thus, the proposed model is useful for predictive tasks to screen for new and potent oxazole and thiazole series of FBPase inhibitors in early drug development.

Acknowledgments

This work was partly supported by the NSFC (No. 20836002).

References

1. King, H.; Aubert, R.E.; Herman, W.H. Global burden of diabetes, 1995–2025: Prevalence, numerical estimates, and projections. *Diabetes Care* **1998**, *21*, 1414–1431.
2. Zhang, B.B.; Moller, D.E. New approaches in the treatment of type 2 diabetes. *Curr. Opin. Chem. Biol.* **2000**, *4*, 4614–4667.
3. Reaven, G.M. Pathophysiology of insulin resistance in human disease. *Physiol. Rev.* **1995**, *75*, 4734–4786.
4. Haffner, S.M.; Lehto, S.; Rönnemaa, T.; Pyörälä, K.; Laakso, M. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. *N. Engl. J. Med.* **1998**, *339*, 2292–2234.
5. Moller, D.E. New drug targets for type 2 diabetes and the metabolic syndrome. *Nature* **2001**, *414*, 8218–8227.
6. Klein, R. Hyperglycemia and microvascular and macrovascular disease in diabetes. *Diabetes Care* **1995**, *18*, 2582–2568.
7. Siconolfi-Baez, L.; Banerji, M.; Lebovitz, H. Characterization and significance of sulfonylurea receptors. *Diabetes Care* **1990**, *13*, 2–8.
8. Bailey, C.J.; Turner, R.C. Metformin. *N. Engl. J. Med.* **1996**, *334*, 5745–5779.
9. Kelley, D.E. Effects of weight loss on glucose homeostasis in NIDDM. *Diabetes Rev.* **1995**, *3*, 366–377.
10. Howlett, H.C.S.; Bailey, C.J. A risk-benefit assessment of metformin in type 2 diabetes mellitus. *Drug Saf.* **1999**, *20*, 489–503.
11. Erion, M.D.; van Poelje, P.D.; Dang, Q.; Kasibhatla, S.R.; Potter, S.C.; Reddy, M.R.; Reddy, K.R.; Jiang, T.; Lipscomb, W.N. MB06322 (CS-917): A potent and selective inhibitor of fructose 1,6-bisphosphatase for controlling gluconeogenesis in type 2 diabetes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7970–7975.
12. Wright, S.W.; Anthony, A.; Carty, M.D.; Danley, D.E.; Hageman, D.L.; Karam, G.A.; Levy, C.B.; Mansour, M.N.; Mathiowetz, A.M.; McClure, L.D. Anilinoquinazoline inhibitors of fructose 1, 6-bisphosphatase bind at a novel allosteric site: Synthesis, *in vitro* characterization, and X-ray crystallography. *J. Med. Chem.* **2002**, *45*, 3865–3877.
13. Lai, C.; Gum, R.J.; Daly, M.; Fry, E.H.; Hutchins, C.; Abad-Zapatero, C.; von Geldern, T.W. Benzoxazole benzenesulfonamides as allosteric inhibitors of fructose-1,6-bisphosphatase. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1807–1810.
14. Wright, S.W.; Carlo, A.A.; Danley, D.E.; Hageman, D.L.; Karam, G.A.; Mansour, M.N.; McClure, L.D.; Pandit, J.; Schulte, G.K.; Treadway, J.L.; *et al*. 3-(2-Carboxy-ethyl)-4,6-dichloro-1*H*-indole-2-carboxylic acid: An allosteric inhibitor of fructose-1,6-bisphosphatase at the AMP site. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2055–2058.
15. Gasteiger, J. The central role of chemoinformatics. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 200–209.

16. Lan, P.; Wu, Z.W.; Chen, W.N.; Sun, P.H.; Chen, W.M. Molecular modeling studies on phosphonic acid-containing thiazole derivatives: Design for fructose-1,6-bisphosphatase inhibitors. *J. Mol. Model.* **2011**, *18*, 973–990.
17. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
18. Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold^2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337–1344.
19. Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
20. Roy, P.; Roy, K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb. Sci.* **2008**, *27*, 302–313.
21. *Cerius*^{2}, version 4.6; Accelrys, Inc.: San Diego, CA, USA, 2001.
22. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **2006**, *56*, 237–248.
23. Hao, M.; Li, Y.; Wang, Y.; Zhang, S. Prediction of PKCθ inhibitory activity using the random forest algorithm. *Int. J. Mol. Sci.* **2010**, *11*, 3413–3433.
24. Hao, M.; Li, Y.; Wang, Y.H.; Zhang, S.W. A classification study of respiratory syncytial virus (RSV) inhibitors by variable selection with random forest. *Int. J. Mol. Sci.* **2011**, *12*, 1259–1280.
25. Hao, M.; Li, Y.; Wang, Y.; Zhang, S. Prediction of $P2Y_{12}$ antagonists using a novel genetic algorithm-support vector machine coupled approach. *Anal. Chim. Acta* **2011**, *690*, 53–63.
26. Li, J.Z.; Li, S.Y.; Lei, B.L.; Liu, H.X.; Yao, X.J.; Liu, M.C.; Gramatica, P. A new strategy to improve the predictive ability of the local lazy regression and its application to the QSAR study of melanin-concentrating hormone receptor 1 antagonists. *J. Comput. Chem.* **2010**, *31*, 973–985.
27. Hao, M.; Li, Y.; Wang, Y.; Yan, Y.; Zhang, S. Combined 3D-QSAR, molecular docking, and molecular dynamics study on piperazinyl-glutamate-pyridines/pyrimidines as potent $P2Y_{12}$ antagonists for inhibition of platelet aggregation. *J. Chem. Inf. Model.* **2011**, *51*, 2560–2572.
28. Golbraikh, A.; Shen, M.; Xiao, Z.Y.; Xiao, Y.D.; Lee, K.H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* **2003**, *17*, 241–253.
29. Dang, Q.; Liu, Y.; Cashion, D.K.; Kasibhatla, S.R.; Jiang, T.; Taplin, F.; Jacintho, J.D.; Li, H.; Sun, Z.; Fan, Y.; *et al*. Discovery of a series of phosphonic acid-containing thiazoles and orally bioavailable diamide prodrugs that lower glucose in diabetic animals through inhibition of fructose-1,6-bisphosphatase. *J. Med. Chem.* **2010**, *54*, 153–165.
30. Dang, Q.; Kasibthatla, S.R.; Jiang, T.; Taplin, F.; Gibson, T.; Potter, S.C.; van Poelje, P.D.; Erion, M.D. Oxazole phosphonic acids as fructose 1,6-bisphosphatase inhibitors with potent glucose-lowering activity. *Med. Chem. Commun.* **2011**, *2*, 287–290.
31. Dang, Q.; Kasibhatla, S.R.; Reddy, K.R.; Jiang, T.; Reddy, M.R.; Potter, S.C.; Fujitaki, J.M.; van Poelje, P.D.; Huang, J.; Lipscomb, W.N.; Erion, M.D. Discovery of potent and specific fructose-1,6-bisphosphatase inhibitors and a series of orally-bioavailable phosphoramidase-sensitive prodrugs for the treatment of type 2 diabetes. *J. Am. Chem. Soc.* **2007**, *129*, 15491–15502.

32. Dang, Q.; Brown, B.S.; Liu, Y.; Rydzewski, R.M.; Robinson, E.D.; van Poelje, P.D.; Reddy, M.R.; Erion, M.D. Fructose-1,6-bisphosphatase inhibitors. 1. Purine phosphonic acids as novel AMP mimics. *J. Med. Chem.* **2009**, *52*, 2880–2898.
33. Dang, Q.; Kasibhatla, S.R.; Xiao, W.; Liu, Y.; DaRe, J.; Taplin, F.; Reddy, K.R.; Scarlato, G.R.; Gibson, T.; van Poelje, P.D.; *et al*. Fructose-1,6-bisphosphatase inhibitors. 2. Design, synthesis, and structure-activity relationship of a series of phosphonic acid containing benzimidazoles that function as 5'-adenosinemonophosphate (AMP) mimics. *J. Med. Chem.* **2010**, *53*, 441–451.
34. Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
35. Roy, K.; Mandal, A. Predictive QSAR modeling of CCR5 antagonist piperidine derivatives using chemometric tools. *J. Enzyme Inhib. Med. Chem.* **2009**, *24*, 205–223.
36. Ma, Q.; Wyszynski, D.F.; Farrell, J.J.; Kutlar, A.; Farrer, L.A.; Baldwin, C.T.; Steinberg, M.H. Fetal hemoglobin in sickle cell anemia: Genetic determinants of response to hydroxyurea. *Pharmacogenomics J.* **2007**, *7*, 386–394.
37. Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinforma.* **2008**, *9*, 400–418.
38. Hao, M.; Zhang, X.; Ren, H.; Li, Y.; Zhang, S.; Luo, F.; Ji, M.; Li, G.; Yang, L. *In silico* identification of structure requirement for novel thiazole and oxazole derivatives as potent fructose 1,6-bisphosphatase inhibitors. *Int. J. Mol. Sci.* **2011**, *12*, 8161–8180.
39. Mavromoustakos, T.; Durdagi, S.; Koukoulitsa, C.; Simcic, M.; Papadopoulos, M.G.; Hodoscek, M.; Golic Grdadolnik, S. Strategies in the rational drug design. *Curr. Med. Chem.* **2011**, *18*, 2517–2530.
40. O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33–46.
41. Taha, M.; Qandil, A.; Zaki, D.; AlDamen, M. Ligand-based assessment of factor Xa binding site flexibility via elaborate pharmacophore exploration and genetic algorithm-based QSAR modeling. *Eur. J. Med. Chem.* **2005**, *40*, 701–727.
42. Mazzatorta, P.; Cronin, M.T.D.; Benfenati, E. A QSAR study of avian oral toxicity using support vector machines and genetic algorithms. *QSAR Comb. Sci.* **2006**, *25*, 616–628.
43. Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: Application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous). *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328–1334.
44. Hemmateenejad, B.; Miri, R.; Akhond, M.; Shamsipur, M. QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 91–99.
45. Gao, H. Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402–407.
46. Fatemi, M.H.; Jalali-Heravi, M.; Konuze, E. Prediction of bioconcentration factor using genetic algorithm and artificial neural network. *Anal. Chim. Acta* **2003**, *486*, 101–108.

47. Huang, C.; Wang, C. A GA-based feature selection and parameters optimizationfor support vector machines. *Expert. Syst. Appl.* **2006**, *31*, 231–240.
48. Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267–281.
49. Leardi, R.; González, A. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195–207.
50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
51. Durdagi, S.; Mavromoustakos, T.; Chronakis, N.; Papadopoulos, M.G. Computational design of novel fullerene analogues as potential HIV-1 PR inhibitors: Analysis of the binding interactions between fullerene inhibitors and HIV-1 PR residues using 3D QSAR, molecular docking and molecular dynamics simulations. *Bioorg. Med. Chem.* **2008**, *16*, 9957–9974.
52. Zaheer-ul-Haq; Lodhi, M.A.; Nawaz, S.A.; Iqbal, S.; Khan, K.M.; Rode, B.M.; Atta-ur-Rahman; Choudhary, M.I. 3D-QSAR CoMFA studies on bis-coumarine analogues as urease inhibitors: A strategic design in anti-urease agents. *Bioorg. Med. Chem.* **2008**, *16*, 3456–3461.