Structural Chemistry (2020) 31:831–839
https://doi.org/10.1007/s11224-019-01468-w

ORIGINAL RESEARCH

The index of ideality of correlation: A statistical yardstick for better QSAR modeling of glucokinase activators

Manisha Nimbhal[1] · Kiran Bagri[1] · Parvin Kumar[2] · Ashwani Kumar[1]

Received: 29 October 2019 / Accepted: 21 November 2019 / Published online: 11 December 2019

Abstract

Glucokinase is an enzyme which is responsible for the conversion of glucose to glucose-6-phosphate through ATP-dependent phosphorylation and has a significant role in glycogen synthesis and hepatic glucose production. Allosteric activators of glucokinase could be an attractive approach for the treatment of T2DM (type 2 diabetes mellitus). Recently, an innovative standard "Index of Ideality of Correlation" has been introduced for the estimation of QSAR (quantitative structural activity relationship) model's potential. In the present work, QSAR models for activators of glucokinase have been developed with target function TF_1 and TF_2 using index of ideality of correlation (IIC). Along with this, prediction of calibration sets for different QSAR models generated for different splits is also categorized as correct and wrong. Moreover, dispersion in the different runs of same split is also explained. The values of criteria R^2 and IIC for best split prepared with target function TF_1 are 0.6554 and 0.7912 and that for TF_2 are 0.9531 and 0.9758, respectively. The models developed with index of ideality of correlation are better than the models generated without index of ideality of correlation. The IIC could be a better criteria option for predictability of QSAR model for glucokinase activators.

Keywords Glucokinase · Target function · QSAR · Index of ideality of correlation · Dispersion

Introduction

Increased hepatic glucose production and dysfunction of the pancreatic β-cells are mainly responsible for the whole-body insulin resistance and hyperglycemia, which are related to type 2 diabetes mellitus [1]. It is a chronic metabolic disease influencing about 150 million people throughout the world [2]. In the developing world, it is assumed as one of the primary causes of death, and from recent data of IDF Diabetes Atlas, it is specified as chief obstacle in the universal development [3]. At present, there is not a single oral antidiabetic drug available through which we can achieve permanent glycemic control. In reality, the utilization of combination therapy is assumed as better option than monotherapy, although combination therapy also has several unwanted side effects. Thus, to overcome the crisis related with T2D therapies, the demand of more effective and safe novel antidiabetic drugs is also rising [4].

Glucokinase (GK) can be suggested as a better target option for the treatment of T2D because of having activity in multiple organs which helps in control of whole-body glucose level [5]. It is related to the hexokinase family also known as hexokinase D/hexokinase IV [6]. It is involved in the first step of glycolysis and is accountable for the ATP-dependent phosphorylation of glucose. GK is present in pancreatic β-cells and acts like as detector for secretion of insulin. It maintains the glucose homeostasis due to unique kinetic features [7].

Quantitative structure activity relationships (QSARs)/quantitative structure property relationships (QSPRs) play an important function in screening and development of the novel biomolecules with effectiveness [8]. In QSAR/QSPR, mathematical models are developed through which we can relate the physiochemical or biological property of compounds with their chemical structures [9]. After the development of a QSAR model, actual predictive potential of the QSAR model

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11224-019-01468-w) contains supplementary material, which is available to authorized users.

✉ Ashwani Kumar
ashwanijangra@ymail.com

[1] Department of Pharmaceutical Sciences, Guru Jambheshwar University of Sciences & Technology, Hisar 125001, India

[2] Department of Chemistry, Kurukshetra University, Kurukshetra, India

is corroborated with distinct decisive factors. Development of these criterions is not an easy task. Some matrices have been explained in the literature to explain the predictability. Recently, a new criteria known as the Index of Ideality of Correlation (IIC) has been suggested. The IIC estimates the predictive potential of QSAR model which is not only based on the correlation coefficient but also depends on the residual values of endpoint and arrangement of the dots image related to the diagonal [10]. The purpose of the current research is to compare the IIC with other different well-known criteria of predictive potential of QSAR models for activators of glucokinase in T2D.

Materials and methods

In the present study, a data set consisting of 67 benzamide derivatives was used for QSAR model development. The experimental values for EC_{50} data were retrieved from literature reports [11–13]. Then, these experimental values were changed into negative decimal logarithm (pEC_{50}) which was considered as the dependent variable for QSAR model generation [14]. 3D arrangement of the glucokinase activators were sketched with Marvin Sketch [15], and further, Open Babel [16] was used to convert them into the SMILES depiction. Three different splits were prepared by random distribution of molecules into training, invisible training, calibration, and external validation sets [17]. The training, invisible training sets are like the manufacturer and inspector of the correlation weights, and the external validation set is the indicator of the true predictive potential of the correlation weights [18]. OECD guidelines were precisely followed in QSAR model development [19]. The percentage of the identical distribution of compounds into splits was determined with the well-known method [20], and it is summarized in Table 1. From this table, nonidentical nature of splits can be confirmed.

Optimal descriptors

The CORAL QSAR modeling depends on the concept, described in the following Eq. 1 [21]:

$$\text{Endpoint} = \text{F (Molecular Structure)} \quad (1)$$

Simplified molecular-input line-entry system (SMILES) notation is regarded as the most suitable depiction for the molecular structures of compounds [22]. In QSAR modeling, molecular optimal descriptor (DCW) is defined as the function of the molecule's SMILES notation, described in Eq. 2 [23].

$$\text{DCW} = \text{F (SMILES)} \quad (2)$$

Molecular structures of compounds can be shown as SMILES and molecular graph; in several cases, hybrid representation is also used [24]. In hybrid form, both SMILES and molecular graph are employed for model development in QSAR modeling. The CORAL method depends on correlation weights of structural attributes obtained from hydrogen-suppressed graph (HSG), hydrogen-filled graph (HFG), and graph of atomic orbitals (GAO). There are two types of molecular features, named as local and global which are extracted from HSG [25]. The hybrid descriptor based on Monte Carlo simulation of the activators of glucokinase was computed with the following equation [26]:

$$\text{DCW (T*N*)} = \text{DCW}_{\text{graph}}\text{ (T*N*)} + \text{DCW}_{\text{SMILES}}\text{ (T*N*)} \quad (3)$$

Table 1 Percentage of identical distribution of compounds into the training set, invisible training set, calibration set, and validation set

Splits	Sets	Split 1	Split 2	Split 3
1	Training set	100	33	07
	Invisible training set	100	11	35
	Calibration set	100	08	00
	Validation set	100	10	20
2	Training set		100	29
	Invisible training set		100	22
	Calibration set		100	09
	Validation set		100	00
3	Training set			100
	Invisible training set			100
	Calibration set			100
	Validation set			100

To measure (%) of nonidentity of splits into the training, invisible training, calibration, and validation set, examined in this work

Identity (%) = Ni.j/0.5 (Ni + Nj)*100

where

Ni is the number of compounds which are distributed into the set for i-th split

Nj is the number of compounds which are distributed into the set for j-th split

Index of ideality of correlation (IIC)

In Monte Carlo optimization, sets of correlation weights CW (x) are the coefficients which result in production of the target function with higher value. Different target functions can be calculated for available optimization method by changing the value of parameter W_{IIC}. Here, two versions of the target function were evaluated. The target function is defined as [27]:

$$TF_1 = R_{training} + R_{invisible\ training} - \left|R_{training} - R_{invisible\ training}\right| \times 0.1 \quad (4)$$

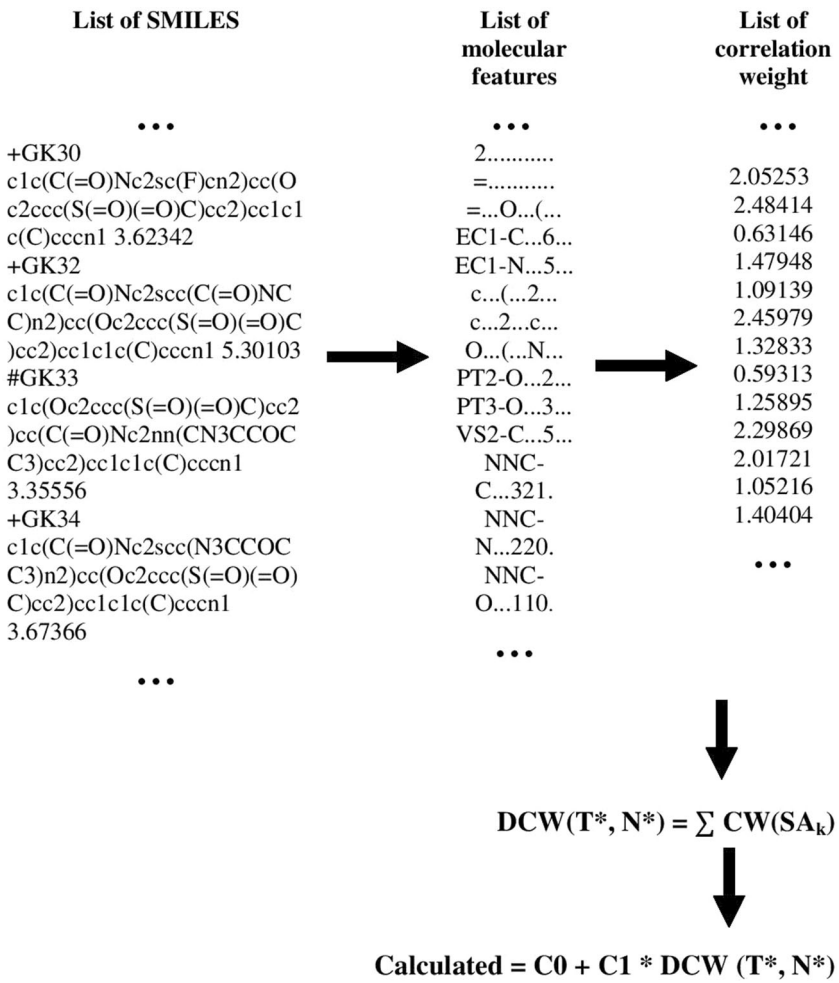

Fig. 1 The general scheme for building up of QSAR model by means of Monte Carlo method

where $R_{training}$ and $R_{invisible\ training}$ are the correlation coefficients between the observed and predicted endpoints for the training and invisible training sets, respectively.

$$TF_2 = TF_1 + IIC \times W_{IIC} \tag{5}$$

The IIC is described as the index of ideality of correlation. The W_{IIC} is an experimental coefficient; generally its value is considered as zero in the Monte Carlo optimization. But in the case of modified version, the value of W_{IIC} is taken as greater than zero, but too large value of W_{IIC} can also ruin the optimization process.

Index of ideality of correlation is defined as:

$$IIC = R_{calibration} \times \frac{min\left(-MAE_{calibration,} + MAE_{calibration}\right)}{max\left(+MAE_{calibration,}\right)} \tag{6}$$

$$Here, -MAE_{calibration} = \frac{1}{N^-} \times \sum_{K=1}^{N^-} \left(Y_{obs} - Y_{pred}\right); where, \left(Y_{obs} - Y_{pred}\right) < 0 \tag{7}$$

$$Here, +MAE_{calibration} = \frac{1}{N^-} \times \sum_{K=1}^{N^+} \left(Y_{obs} - Y_{pred}\right); where, \left(Y_{obs} - Y_{pred}\right) > 0 \tag{8}$$

In Eqs. 7 and 8, the parameters Y_{obs} and Y_{pred} are correspondingly observed and calculated values of pEC_{50} for the calibration set. If we use IIC as a replacement of the conventional correlation coefficient, the statistical parameters of any inferior models could be improved. Hence, the IIC can be taken as an alternative option to check the characteristic of developed model. The application of the IIC becomes impossible if [27]:

$$-_{MAE} = +_{MAE} = 0$$

How to rate different criteria of predictive potential as correct or wrong? [28]

$$\text{If } X_{CLB}\ [1] > X_{CLB}\ [2] \text{ and } R^2_{VLD}\ [1] > R^2_{VLD}\ [2] \tag{9}$$

(then the rating is given as correct)

$$\text{If } X_{CLB}\ [2] > X_{CLB}\ [1] \text{ and } R^2_{VLD}\ [2] > R^2_{VLD}\ [1] \tag{10}$$

Table 2 Statistical characteristics of three runs for split 1 of glucokinase activators with Monte Carlo optimization

Run	TF	WIIC	Set	n	R^2	CCC	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Rm^2	IIC
1	TF1	0.0	CLB	12	0.2635	0.4531	− 0.1290	0.1158	− 0.3855	0.5265	0.1324	0.4075
			VLD	10	0.7209							
	TF2	0.2	CLB	12	0.7190	0.8100	0.5950	0.7668	0.6345	0.8751	0.6092	0.8476
			VLD	10	0.7936							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**
2	TF1	0.0	CLB	12	0.4186	0.6285	0.0130	0.4056	0.0678	0.6817	0.2625	0.5732
			VLD	10	0.6261							
	TF2	0.2	CLB	12	0.6706	0.7200	0.5578	0.6309	0.4217	0.8024	0.5492	0.8189
			VLD	10	0.7525							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**
3	TF1	0.0	CLB	12	0.7327	0.8213	0.6334	0.7657	0.6329	0.8745	0.6277	0.1515
			VLD	10	0.4978							
	TF2	0.2	CLB	12	0.6644	0.7561	0.5302	0.6802	0.4988	0.8287	0.5422	0.8151
			VLD	10	0.5543							
		Rating			**Wrong**	**Wrong**	**Wrong**	**Wrong**	**Wrong**	**Wrong**	**Wrong**	**Correct**

Where CLB represents the calibration set, VLD is the validation set, n is number of molecules in set, R^2 is regression coefficient, CCC is concordance correlation coefficient, Q^2 is cross-validation correlation coefficient, Rm^2 is criteria of predictability, and IIC is index of ideality of correlation

And

$$\text{If } X_{CLB}\,[1] > X_{CLB}\,[2] \text{ and } R^2_{VLD}\,[1] < R^2_{VLD}\,[2] \text{ (then rating is given as wrong)} \quad (11)$$

The rating is given as "correct" if the values of the criteria for both calibration and validation set for model 1 are higher than model 2 or values or the parameters of both calibration and validation set related to model 2 are more than 1. But in comparison of model 1 with model 2, if the value of the X [1] increases for calibration set and the value of R^2 decreases for validation set, then rating is given as "wrong." The X_{CLB} [1] and X_{CLB} [2] demonstrated the values of criteria R^2, q^2, q^2_{F1}, q^2_{F2}, Q^2_{F3}, Rm^2, CCC, and IIC for model 1 and model 2, where models 1 and 2 were prepared with TF_1 and TF_2 with W_{IIC} values 0 and 0.2, respectively.

In CORAL QSAR modeling, the dispersion in several runs of the same split with same optimization procedure can be

Table 3 Statistical characteristics of three runs for split 2 of glucokinase activators with Monte Carlo optimization

Run	TF	WIIC	Set	n	R^2	CCC	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Rm^2	IIC
1	TF1	0.0	CLB	12	0.5959	0.6932	0.3778	0.1831	0.1810	0.6006	0.4577	0.3380
			VLD	10	0.5138							
	TF2	0.2	CLB	12	0.7819	0.8758	0.7240	0.7606	0.7600	0.8829	0.6921	0.8841
			VLD	10	0.9492							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**
2	TF1	0.0	CLB	12	0.7863	0.7354	0.7136	− 0.1586	− 0.1615	0.4336	0.3341	0.4706
			VLD	10	0.5659							
	TF2	0.2	CLB	12	0.8661	0.9228	0.8123	0.8573	0.8569	0.9302	0.8044	0.9239
			VLD	10	0.9254							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**
3	TF1	0.0	CLB	12	0.7103	0.8345	0.5723	0.6548	0.6540	0.8312	0.5997	0.5997
			VLD	10	0.7745							
	TF2	0.2	CLB	12	0.8854	0.7718	0.8487	0.7084	0.7077	0.8574	0.2907	0.9398
			VLD	10	0.4672							
		Rating			**Wrong**	**Wrong**	**Correct**	**Correct**	**Correct**	**Correct**	**Wrong**	**Correct**

Explanation of terms used is same as given in footnote of Table 2

Table 4 Statistical characteristics of three runs for split 3 of glucokinase activators with Monte Carlo optimization

Run	TF	WIIC	Set	n	R^2	CCC	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Rm^2	IIC
1	TF1	0.0	CLB	11	0.6554	0.7912	0.4098	0.5017	0.4976	0.4940	0.5221	0.7912
			VLD	10	0.6421							
	TF2	0.2	CLB	11	0.9531	0.9468	0.9341	0.9143	0.9136	0.9130	0.7400	0.9758
			VLD	10	0.8314							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**
2	TF1	0.0	CLB	11	0.8476	0.9129	0.7337	0.8347	0.8333	0.8321	0.7793	0.7883
			VLD	10	0.6480							
	TF2	0.2	CLB	11	0.8729	0.9274	0.7421	0.8421	0.8408	0.8397	0.8148	0.9341
			VLD	10	0.8439							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**
3	TF1	0.0	CLB	11	0.7888	0.8734	0.6259	0.7241	0.7219	0.7199	0.7015	0.6301
			VLD	10	0.6405							
	TF2	0.2	CLB	11	0.8893	0.9286	0.7903	0.8311	0.8298	0.8285	0.7638	0.9417
			VLD	10	0.7168							
		Rating			**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**	**Correct**

Explanation of terms used is same as given in footnote of Table 2

explained and calculated with the standard deviation. Along with this, the developed splits could be categorized as correct and uncertain. Firstly, the average values of a criteria X1 for model-1 and X2 for model-2 were calculated, and then their standard deviations Δ1 and Δ2 were determined which were developed with target function TF_1 and TF_2. On the basis of following inequality, it can be defined as uncertain or correct [29].

$$\overline{X_1} - \overline{X_2} \leq Max(\Delta_1, \Delta_2) \quad (12)$$

Max $(\Delta_1, \Delta_2) = \Delta_1 > \Delta_2$, otherwise Δ_2 is taken

Then, in the standard deviations of a criteria Δ_1, Δ_2 maximum standard deviation value is determined. Suppose if the value of Δ_1 is greater than Δ_2, then Δ_1 is considered as "Max Δ." Further, if the difference between the average values of criteria was lower than the max Δ value, then it is recommended as "uncertain," and opposite of above statement is supposed to be the "correct."

Building of CORAL model

Three steps involved in the development of the CORAL QSAR models were [30] the following:

1. The total data set was divided into the training, invisible training, calibration, and validation sets, and different splits were generated by running the CORAL SEA 2019 with the search for preferable number of epochs (N*) and threshold (T); ranges of T and N_{epoch} were selected from 1 to 10 and 1 to 50, respectively.
2. Then the models were developed with preferable number of threshold (3) and N_{epoch} (25), and molecular features for all compounds were computed by mean of CORAL.
3. Correlation weights were extracted for all molecular features related to QSAR models.

Figure 1 represents the general scheme used of CORAL model development with Monte Carlo method [31].

Table 5 Equations for QSAR models with target function TF_2

Developed splits	No. of run	Equations
Split 1	Run 1	Endpoint = −2.90926 (± 0.12386) + 0.05615 (± 0.00103) × DCW(2,13)
	Run 2	Endpoint = −2.01484 (± 0.06762) + 0.04111 (± 0.00047) × DCW(2,5)
	Run 3	Endpoint = −2.17933 (± 0.12822) + 0.06355 (± 0.00130) × DCW(3,3)
Split 2	Run 1	Endpoint = −2.16134 (± 0.09073) + 0.06071 (± 0.00093) × DCW(2,4)
	Run 2	Endpoint = −3.80739 (± 0.12969) + 0.05631 (± 0.00090) × DCW(2,3)
	Run 3	Endpoint = −3.79536 (± 0.11795) + 0.06700 (± 0.00101) × DCW(1,5)
Split 3	Run 1	Endpoint = −2.58645 (± 0.14207) + 0.06141 (± 0.00126) × DCW(3,3)
	Run 2	Endpoint = −4.49770 (± 0.16462) + 0.07441 (± 0.00138) × DCW(1,3)
	Run 3	Endpoint = −4.09191 (± 0.16438) + 0.08455 (± 0.00166) × DCW(1,3)

Table 6 Rating of recommendations provided by criteria of dispersion in the three splits of Monte Carlo optimization

Parameters	X_1	Δ_1	X_2	Δ_2	X_2-X_1	Rating according to equation
Split 1						
R^2	0.4716	0.1952	0.7412	0.0686	0.2696	Correct
CCC	0.6343	0.1504	0.8006	0.0624	0.1663	Correct
Q^2	0.1725	0.3310	0.6396	0.0907	0.4671	Correct
Q^2_{F1}	0.4290	0.2658	0.7460	0.0867	0.3170	Correct
Q^2_{F2}	0.1051	0.4166	0.6017	0.1356	0.4967	Correct
Q^2_{F3}	0.6942	0.1423	0.8640	0.0464	0.1698	Correct
Rm^2	0.3409	0.2097	0.4174	0.2302	0.0765	Uncertain
IIC	0.3774	0.1735	0.8581	0.0370	0.4807	Correct
Split 2						
R^2	0.6975	0.0783	0.8445	0.0449	0.1470	Correct
CCC	0.7544	0.0592	0.8568	0.0631	0.1024	Correct
Q^2	0.5546	0.1377	0.7950	0.0524	0.2404	Correct
Q^2_{F1}	0.2264	0.3335	0.7754	0.0617	0.5490	Correct
Q^2_{F2}	0.2245	0.3343	0.7749	0.0618	0.5504	Correct
Q^2_{F3}	0.6218	0.1630	0.8902	0.0302	0.2684	Correct
Rm^2	0.4638	0.1085	0.5957	0.2205	0.1319	Uncertain
IIC	0.4694	0.1068	0.9159	0.0234	0.4465	Correct
Split 3						
R^2	0.8187	0.1232	0.8503	0.0440	0.0316	Uncertain
CCC	0.8836	0.0668	0.9098	0.0257	0.0262	Uncertain
Q^2	0.6925	0.2160	0.7194	0.0690	0.0269	Uncertain
Q^2_{F1}	0.7502	0.1787	0.7991	0.0532	0.0489	Uncertain
Q^2_{F2}	0.7482	0.1802	0.7975	0.0536	0.0493	Uncertain
Q^2_{F3}	0.7464	0.1815	0.7960	0.0540	0.0497	Uncertain
Rm^2	0.6805	0.1131	0.7600	0.0463	0.0796	Uncertain
IIC	0.8518	0.0877	0.8353	0.1451	−0.0165	Uncertain

Table 7 Percentage of correct recommendations provided by criteria of the predictive potential of developed QSAR model

Split	Run	R^2	CCC	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	Rm^2	IIC
1	1	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1
	3	0	0	0	0	0	0	0	1
2	1	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1
	3	0	0	1	1	1	1	0	1
3	1	1	1	1	1	1	1	1	1
	2	1	1	1	1	1	1	1	1
	3	1	1	1	1	1	1	1	1
Correct recommendation		78%	78%	89%	89%	89%	89%	78%	100%
Recommendations which are done with considering the dispersion of criteria									
1		1	1	1	1	1	1	0	1
2		1	1	1	1	1	1	0	1
3		0	0	0	0	0	0	0	0
Correct recommendation		67%	67%	67%	67%	67%	67%	0%	67%

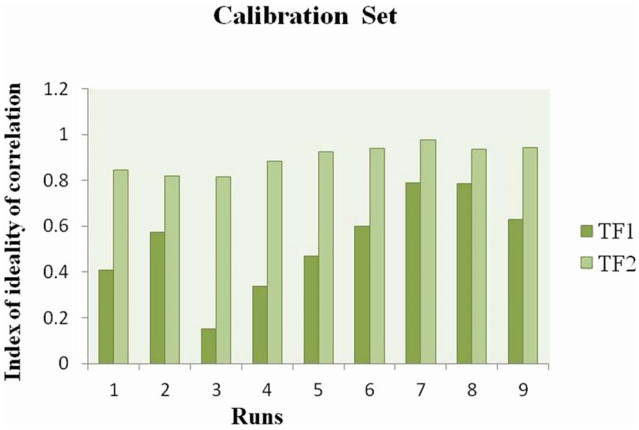

Fig. 2 Graphical representation of IIC for different splits with utilization of target functions TF_1 and TF_2

Domain of applicability

The statistical defects related to the molecular features depend on allocation of different molecular features into the training and calibration set [28].

$$d(F_K) = \frac{P_T(F_K) - P_C(F_K)}{N_T(F_K) + N_C(F_K)} \quad (13)$$

where $P_T(F_K)$ and $P_C(F_K)$ are probabilities of feature F_K to be in training set and calibration set and $N_T(F_K)$ and $N_C(F_K)$ are prevalence of feature F_K in the training set and calibration sets, respectively.

The defect of the individuals' SMILES can be calculated as:

$$d\ (SMILES) = \sum_{FK \in SMILES} d(F_K) \quad (14)$$

The addition of defect of individual SMILES results into the defect of the split related to training, invisible training, calibration, and validation set.

$$d(Split) = \sum d(SMILES) \quad (15)$$

Domain of applicability can be estimated as

$$d(SMILES) < 2 \times d\left(\overline{SMILES}\right) \quad (16)$$

where the d $\overline{(\text{SMILES})}$ is the average of the statistical defect of SMILES related to the training set.

Results and discussion

The major purpose behind the use of different criteria to predict the potential of developed QSAR models was to identify that the built models have predictability control or not. The comparison of q^2, Rm^2, CCC, and IIC provided the satisfactory outcome in terms of the predictive potential of the QSAR model because all the criteria have comparable range from zero to one. Moreover, during comparison of two models, one having larger value of criteria is assumed as superior, and this is true for all above mentioned parameters [10]. Tables 2, 3, and 4 are describing the comparison of different statistical characteristics of three runs of split 1, 2, and 3 of glucokinase activators with Monte Carlo optimization. According to the rating principle, in case of split 1, for

Table 8 Structural attributes extracted from QSAR model of best split

S. No.	Structural attributes	Correlation weights	N1	N2	N3
Promoters of endpoint increase					
1.	c...c...1...	1.4991	26	20	11
2.	c...c...2...	2.15034	26	20	11
3.	++++N---O===	1.16073	26	20	11
4.	2...(.....	2.05253	26	20	11
5.	2.....	2.48414	26	20	11
6.	=.....	0.63146	26	20	11
7.	=...O...(...	1.47948	26	20	11
8.	EC1-C...6...	1.09139	26	20	11
9.	EC1-N...5...	2.45979	26	20	11
10.	EC1-O...6...	1.32833	26	20	11
11.	c...(...2...	0.59313	26	20	11
12.	c...1...c...	1.25895	26	20	11
13.	c...2...c...	2.29869	26	20	11
14.	O...(...N...	2.01721	26	20	11
15.	PT2-O...2...	1.05216	26	20	11
16.	PT3-O...3...	2.40404	26	20	11
17.	VS2-C...5...	1.42829	26	20	11
18.	NNC-C...321.	2.32328	26	20	11
19.	NNC-N...220.	1.04267	26	20	11
20.	NNC-O...110.	1.1046	26	20	11
21.	PT2-C...2...	2.04222	26	20	11
Promoters of endpoint decrease					
1.	c...N.....	− 0.55508	26	20	11
2.	n.....	− 0.72228	26	20	11
3.	n...2.....	− 0.50163	26	20	11
4.	(...C...(...	− 0.80875	26	20	11
5.	++++N---B2==	− 0.23788	26	20	11
6.	1...c...(...	− 0.85025	26	20	11
7.	EC1-O...3...	− 0.93775	26	20	11
8.	N.....	− 0.80988	26	20	11
9.	N...c...2...	− 0.82967	26	20	11
10	O.....	− 0.53408	26	20	11
11.	O... = ...(...	− 0.44427	26	20	11
12	O... = ...	− 0.68024	26	20	11
13.	c...N...(...	− 0.33517	26	20	11

Where N1 is the number of SMILES in training set with SA; N2 is the number of SMILES in invisible training set with SA; N3 is the number of SMILES in calibration set with SA

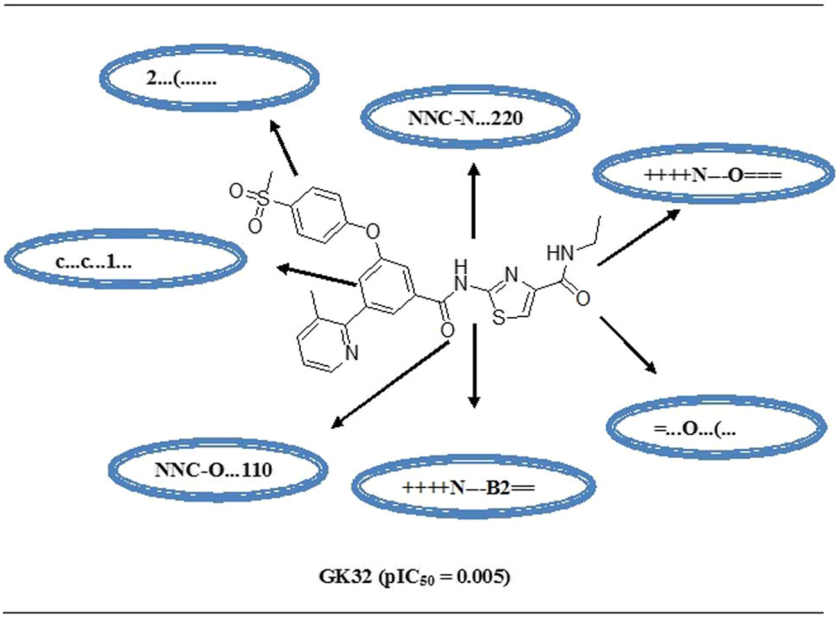

Fig. 3 SMILES attributes present in the glucokinase activator

IIC, rating was identified as correct in three run, while for R^2, CCC, Rm^2, and q^2 matrices, it was correct only for two runs. In split 2, rating was correct for IIC in all three runs, but for other criteria, it was correct only for two runs, and lastly in split 3, rating was obtained as correct for all the statistical parameters. From the interpretation of above data, it could be observed that the splits prepared with TF_2 were better than the TF_1 and the first run of split 3 was defined as the best split prepared due to having highest values of R^2 (0.9531) and IIC (0.9758). Different QSAR equations of various runs of three splits with target function TF_2 are summarized in Table 5, and the rating of recommendations provided by criteria in the three splits of glucokinase is described in Table 6. According to the criteria of standard deviation, splits 1 and 2 were correct or certain for all statistical parameters except Rm^2 matrices although split 3 was uncertain for all criteria. The percentage of correct recommendations estimated for different criteria of the predictive potential of QSAR models is listed in Table 7. The percentage of correct recommendations for IIC was calculated as highest 100% followed by q^2 matrices with 89% and lastly for the R^2, CCC, and Rm^2 with 78%. Percentage according to the standard deviation was 67% for all parameters except Rm^2 matrices. Figure 2 displays the graphical representation of the IIC versus target function TF_1 and TF_2.

Mechanistic interpretation

From the data related to the correlation weight of the developed QSAR models, different structural attributes can be framed as stable positive category, stable negative category, and undefined category [32]. Stable positive category is accountable for the enhancement of the calculated endpoint in all prepared splits, while other negative are contradictory of the above statement. Some structural attributes have not a particular role; they have both positive and negative values of descriptors in different runs, and thus, for such attributes, an accurate correlation weight cannot be expressed [33]. Structural attributes extracted from the best split (first run of split 3) are summarized in Table 8 along with their correlation weights, and Fig. 3 shows the SMILES attributes present in one of the glucokinase activators.

Conclusion

The CORAL software provided the robust and predictive QSAR models for the activators of glucokinase containing benzamide moiety. In comparison of the predictive potential of these models, the index of ideality of correlation emerged as a useful criterion. Application of IIC with target functions resulted in improvement of statistical quality of all QSAR models related to different splits. The coefficient W_{IIC} controlled the effect of the index of ideality of correlation in Monte Carlo optimization which is an empirical parameter and depends on the nature of endpoint and compounds diversity of corresponding available data. Hence, IIC can be used for prediction of glucokinase activation in a lucid way.

Acknowledgments The authors are highly indebted to Dr. Andrey A. Toropov and Dr. Alla P. Toropova for providing the CORAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict(s) of interest.

References

1. Bebernitz GR, Beaulieu V, Dale BA, Deacon R, Duttaroy A, Gao J, Grondine MS, Gupta RC, Kakmak M, Kavana M, Kirman LC, Liang J, Maniara WM, Munshi S, Nadkarni SS, Schuster HF, Stams T, Denny IS, Taslimi PM, Vash B, Caplan SL (2009) Investigation of functionally liver selective glucokinase activators for the treatment. J Med Chem 52:6142–6152
2. Bonn P, Brink DM, Fägerhag J, Jurva U, Robb GR, Schnecke V, Svensson A, Waring MJ, Westerlund C (2012) The discovery of a novel series of glucokinase activators based on a pyrazolopyrimidine scaffold. Bioorg Med Chem Lett 22:7302–7305
3. Charaya N, Pandita D, Grewal AS, Lather V (2018) Design, synthesis and biological evaluation of novel thiazol-2-yl benzamide derivatives as glucokinase activators. Comput Biol Chem 73:221–229
4. Kumari V, Li C (2008) Comparative docking assessment of glucokinase interactions with its allosteric activators. Curr Chem Genomics 2:76–89
5. Bertram LS, Black D, Briner PH, Chatfield R, Cooke A, Fyfe MCT, Murray PJ, Rasamison CM, Reynet C, Schofield KL, Shah VK, Spindler F, Taylor A, Turton R, Williams GM, Wong-kai-in P, Yasuda K (2008) Pharmacokinetics, safety and efficacy of glucokinase activating 2-(4-sulfonylphenyl)-N-thiazol-2-ylacetamides : discovery of PSN-GK1. J Med Chem 51:4340–4345
6. Antoine M, Boutin JA, Ferry G (2009) Binding kinetics of glucose and allosteric activators to human glucokinase reveal multiple conformational states. Biochemistry 48:5466–5482
7. Bowler JM, Hervert KL, Kearley ML, Miller BG (2013) Small-molecule allosteric activation of human glucokinase in the absence of glucose. ACS Med Chem Lett 4:580–584
8. Begum S, Achary PGR (2015) Simplified molecular input line entry system-based: QSAR modeling for MAP kinase-interacting protein kinase (MNK1). SAR QSAR Environ Res 26(5):343–361
9. Begam BF, Kumar JS (2016) Computer assisted QSAR / QSPR approaches – a review. Ind J Sci Tech 9(8). https://doi.org/10.17485/ijst/2016/v9i8/87901
10. Toropov AA, Toropova AP (2017) The index of ideality of correlation: a criterion of predictive potential of QSPR / QSAR models ? Mutat Res Gen Tox En 819:31–37
11. Park K, Lee BM, Kim YH, Han T, Yi W, Lee DH, Choi HH, Chong W, Lee CH (2013) Discovery of a novel phenylethyl benzamide glucokinase activator for the treatment of type 2 diabetes mellitus. Bioorg Med Chem Lett 23:537–542
12. Park K, Lee BM, Hyun KH, Lee DH, Choi HH, Kim H, Chong W, Kim KB, Nam SY (2014) Discovery of 3-(4-methanesulfonylphenoxy)-N-[1-(2-methoxyethoxymethyl)-1H-pyrazol-3-yl]-5-(3-methylpyridin-2-yl)-benzamideas a novel glucokinase activator (GKA) for the treatment of type 2 diabetes mellitus. Bioorg Med Chem 22:2280–2293
13. Park K, Lee BM, Hyun KH, Han T, Lee DH (2015) Design and synthesis of acetylenyl benzamide derivatives as novel glucokinase activators for the treatment of T2DM. ACS Med Chem Lett 6:296–301
14. Toropova AP, Toropov AA, Veselinovic JB, Miljkovi FN, Veselinovic AM (2014) QSAR models for HEPT derivates as NNRTI inhibitors based on Monte Carlo method. Eur J Med Chem 77:298–305
15. Marvin Sketch v.14.11.17.0, (2014) ChemAxon, XhemAxon KFT. Budapest, Hungary
16. O'Boyle N, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J.Cheminform. 3:33
17. Kumar P, Kumar A, Sindhu J, Lal S (2019) QSAR models for nitrogen containing monophosphonate and bisphosphonate derivatives as human farnesyl pyrophosphate synthase inhibitors based on Monte Carlo method. Drug Res 69:159–167
18. Toropova AP, Toropov AA (2017) The index of ideality of correlation: a criterion of predictability of QSAR models for skin permeability ? Sci Total Environ 586:466–472
19. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models. Available at: http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf
20. Kumar A, Chauhan S (2016) Use of the Monte Carlo method for OECD principles-guided QSAR modeling of SIRT1 inhibitors. Arch Pharm Chem Life Sci 349:1–9
21. Zivkovic JV, Truti NV, Veselinovic JB, Nikoli GM, Veselinovic AM (2015) Monte Carlo method based QSAR modeling of maleimide derivatives as glycogen synthase kinase-3 β inhibitors. Comp Biol Med. https://doi.org/10.1016/j.compbiomed.2015.07.004
22. Kumar A, Chauhan S (2016) QSAR differential model for prediction of SIRT1 modulation using Monte Carlo method. Drug Res 67(3):156–162
23. Sokolović D, Aleksić D, Milenković V, Karaleić S, Mitić D, Kocić J, Mekić B, Veselinović JB, Veselinović AM (2016) QSAR modeling of bis-quinolinium and bis-isoquinolinium compounds as acetylcholine esterase inhibitors based on the Monte Carlo method–the implication for myasthenia gravis treatment. Med Chem Res 25:2989–2998
24. Manisha, Chauhan S, Kumar P, Kumar A (2019) Development of prediction model for fructose-1,6-bisphosphatase inhibitors using the Monte Carlo method. SAR QSAR Environ Res 30:145–159
25. Kumar A, Chauhan S (2018) Use of simplified molecular input line entry system and molecular graph based descriptors in prediction and design of pancreatic lipase inhibitors. Future Med Chem 10:1603–1622
26. Kumar P, Kumar A (2017) Monte Carlo method based QSAR studies of Mer kinase inhibitors in compliance with OECD principles. Drug Res 68(04):189–195
27. Toropov AA, Carbó-dorca R, Toropova AP (2017) Index of ideality of correlation : new possibilities to validate QSAR : a case study. Struct Chem 29(1):33–38
28. Toropov AA, Toropova AP (2018) Use of index of ideality of correlation to improve predictive potential for biochemical endpoints. Toxicol Mech Methods. https://doi.org/10.1080/15376516.2018.1506851
29. Toropova AP, Toropov AA (2019) Does the index of ideality of correlation detect the better model correctly ? Mol Inf. https://doi.org/10.1002/minf.201800157
30. Toropova AP, Toropov AA (2018) The index of ideality of correlation : improvement of models for toxicity to algae of models for toxicity to algae. Nat Prod Res. https://doi.org/10.1080/14786419.2018.1493591
31. Toropova AP (2018) The index of ideality of correlation : hierarchy of Monte Carlo models for glass transition temperatures of polymers. J Polym Res. https://doi.org/10.1007/s10965-018-1618-z
32. Gaikwad R, Ghorai S, Amin SA, Adhikari N, Patel T, Das K, Jha T, Gayen S (2018) Monte Carlo based modelling approach for designing and predicting cytotoxicity of 2-phenylindole derivatives against breast cancer cell line MCF7. Toxicol Vitr 52:23–32
33. Rescifina A, Floresta G, Marrazzo A, Parenti C, Prezzavento O, Nastasi G, Amata E, Dichiara M, Amata E (2017) Development of a sigma-2 receptor affinity filter through a Monte Carlo based QSAR analysis. Eur J Pharm Sci 106:94–101