CP-MLR/PLS derived QSAR rationales for the GPR119 agonistic activity of the Indole-based derivatives

Parihar Raghuraj[1], Sharma Brij Kishore*[1]

1. Department of Chemistry, Government College, Bundi-323 001 (Rajasthan), India

ABSTRACT

QSAR study has been carried out on the GPR119 agonistic activity of indole-based derivatives in terms of Dragon descriptors. The derived QSAR models have revealed that the atomic Sandersons electronegativities weighted and charge accounting descriptors ATS7e, GATS1e, GATS4e and GGI8, molecular mass weighted descriptors, MATS7m and BELm5, and atomic polarizabilities weighted descriptors ATS7p and BELp8, and molecular topology accounting feature Lovasz-Pelikan index (LP1) played a pivotal role in rationalization of GPR119 agonistic activity of titled compounds. Hydrophilic factor (Hy) and certain structural fragments, such as CHR2X (C-008), R--CX--X (C-008) and H attached to heteroatom (H-050) are also predominant to explain GPR119 agonistic actions of indole-based derivatives. PLS analysis has also corroborated the dominance of CP-MLR identified descriptors. Applicability domain analysis revealed that the suggested model matches the high quality parameters with good fitting power and the capability of assessing external data and all of the compounds was within the applicability domain of the proposed model and were evaluated correctly.

Keywords: QSAR; GPR119 agonistic activity; Combinatorial protocol in multiple linear regression (CP-MLR) analysis; PLS analysis; Dragon descriptors; Indole-based derivatives.

INTRODUCTION

Diabetes mellitus, characterized by high blood glucose levels, is a metabolic disorder. Nearly 90% of all cases of diabetes belong to type 2 diabetes mellitus (T2DM). T2DM, highly associated with obesity, is due to insulin resistance and impaired pancreatic β-cell function. The various microvascular and macrovascular complications produced by chronic hyperglycemia substantially increase disease-related morbidity and mortality in patients and limit activities of daily living [1]. It is estimated in a study that nearly 350 million people suffering worldwide from diabetes [2]. At present, intake of anti-diabetic agents, diet control and exercise is the management of type 2 diabetes. A large number of T2DM patients fail to reach desired HbA1c levels due to insufficient glycemic control [3]. Thus there is a need to develop a novel glucose-lowering drug to attain better glycemic control which protect pancreatic β -cells or exerts anti-obesity effects and devoid of causing hypoglycemia and cardiovascular side effects. Incretin-related approaches including dipeptidyl peptidase-4 inhibitors and glucagon-like peptide-1 (GLP-1) receptor agonists[4-6], and GPR119 [7-12] are the potential target for anti-diabetic therapy.

GPR119, a member of rhodopsin β -group of G protein-coupled receptors, expressed primarily in pancreatic β -cells and enteroendocrine cells of the gastrointestinal tract. Stimulation of adenylate cyclase and increase in intracellular cyclic AMP levels through G β $_s$ coupling is caused by the activation of GPR119 receptor leading to insulin secretion from the pancreatic β -cells. On the other hand the secretion of incretins like GLP-1 and GIP is promoted by the activation of the GPR119 receptor in the enteroendocrine cells. GPR119 agonists may serve as novel anti-diabetic agents on the basis of the fact that secreted incretins also improve glycemic control and have potential β -cell protective effects [13,14]. The glucose concentration-dependent insulinotropic effects of GPR119 suggested that GPR119 agonists will not produce hypoglycemia. The activation of GPR119 is beneficial therapeutically for obesity [15-18]. The GPR119 agonists demonstrated safety and tolerability in humans [19-22]. Some representative agonists to this receptor are APD-668, SAR-260093 and GSK-1292263. As an attempt to develop a novel GPR119 agonist for the treatment of T2DM, a series of indoline-based compounds has been reported by Sato et al. [23].

The aim of the present communication is to establish the quantitative relationships between the reported activities and descriptors unfolding the substitutional changes in titled molecules.

MATERIALS AND METHOD

Biological actions and theoretical molecular descriptors

The reported twenty six indole-based derivatives is considered as the data set for present study [23]. The general structure of these analogous is represented in Figure 1. These derivatives were
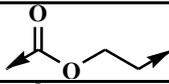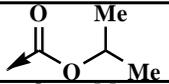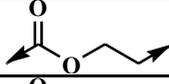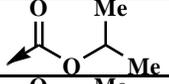
evaluated for their GPR agonist activities in the reporter gene assay using CHO cells stably co-expressing cyclic AMP response element (CRE)–luciferase reporter gene (Promega) and GPR119 and were reported as EC_{50}. The reported activity on molar basis (as pEC_{50}) along with the structural variations of these analogues is shown in Table 1. The data set was sub-divided into training set to develop models and test set to validate the models externally. The test set compounds which were selected using an in-house written randomization program, are also mentioned in Table 1.
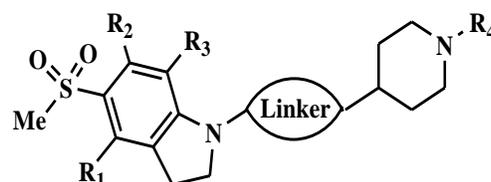
Figure 1: General structure of indole-based GPR119 agonists

Table 1: Structural variations and reported GPR119 agonistic activities of indole-based derivatives.

Cpd.	R_1	R_2	R_3	Linker	R_4	$pEC_{50}(M)^a$
1	H	H	H			5.85
2	H	H	H			6.92
3	H	H	H			7.17
4	H	H	H			5.60
5[b]	H	H	H			5.59
6	H	H	H			8.08
7	H	H	H			5.62
8	F	H	H			7.60
9[b]	Cl	H	H			6.28
10	Me	H	H			6.11
11	H	F	H			6.80
12	H	OMe	H			5.32

13	H	H	F			7.60
14	H	H	Cl			7.68
15	H	H	Me			7.36
16	H	H	OMe			5.77
17[b]	H	H	F			8.17
18	H	H	F			8.41
19	H	H	H			6.31
20[b]	H	H	H			7.06
21	H	H	H			7.77
22[c]	H	H	H			5.06
23	H	H	H			6.28
24[b]	H	H	H			7.80
25	H	H	H			7.96
26	H	H	H			8.11
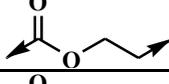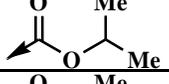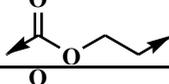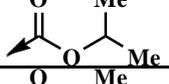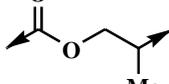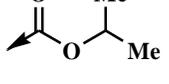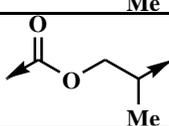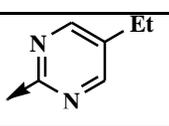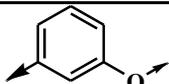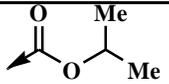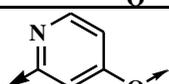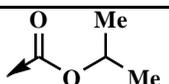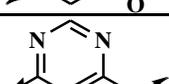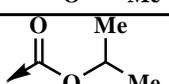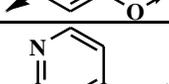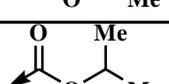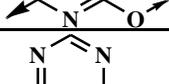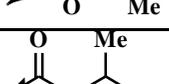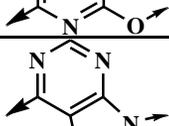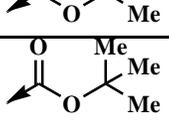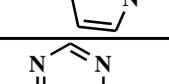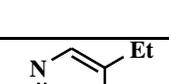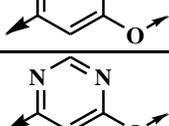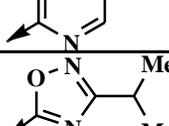
[a]EC_{50} (the concentration of the test compound required to achieve 50% of the maximal response) on molar basis, taken from reference [23]; [b] Compound included in test set; [c] Outlier compound.

The structures of all the data set compounds of Table 1, drawn in 2D Chem Draw [24], were subjected to energy minimization in the MOPAC using the AM1 procedure for closed shell system after converting these into 3D modules. The energy minimization was carried out to attain a well-defined conformer relationship among the congeners under study. Descriptors, belonging to 0D-, 1D- and 2D-classes, of titled compounds were computed using DRAGON software [25]. This software offers a large number of descriptors corresponding to ten different classes of 0D- to 2D-descriptor modules which include the constitutional, topological, molecular walk counts, modified

Burden eigenvalues, Galvez topological charge indices, 2D-autocorrelations, functional groups, atom-centered fragments, empirical descriptors and the properties describing descriptors. Characteristic structural information specific to the descriptor class is offered by these descriptors. The definition and scope of these descriptor's classes is given in Table 2.

Table 2: Descriptor classes used for modeling the hGPR119 agonistic activity of triazolopyridines.

S. No.	Descriptor Class (Acronyms)[a]	Definition and Scope
1	Constitutional (**CONST**)	Dimensionless or 0D descriptors; independent from molecular connectivity and conformations
2	Topological (**TOPO**)	2D-descriptor from molecular graphs and independent conformations
3	Molecular walk counts (**MWC**)	2D-descriptors representing self-returning walk counts of different lengths
4	Modified Burden eigenvalues (**BCUT**)	2D-descriptors representing positive and negative eigenvalues of the adjacency matrix, weights of the diagonal elements and atoms
5	Galvez topological charge indices (**GALVEZ**)	2D-descriptors representing the first 10 eigenvalues of corrected adjacency matrix
6	2D-autocorrelatons (**2D-AUTO**)	Molecular descriptors calculated from the molecular graphs by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag)
7	Functional groups (**FUN**)	Molecular descriptors based on the counting of the chemical functional groups
8	Atom centered fragments (**ACF**)	Molecular descriptors based on the counting of 120 atom centered fragments, as defined by Ghose-Crippen
9	Empirical (**EMP**)	1D-descriptors represent the counts of non single bonds, hydrophilic groups and ratio of the number of aromatic bonds and total bonds in an H-depleted molecule
10	Properties (**PROP**)	1D-descriptors representing molecular properties of a molecule

[a] Reference [25].

A total number of 485 descriptors, belonging to 0D- to 2D- modules, computed by Dragon software have been utilized to obtain most appropriate models describing the biological activity. The descriptors pool has been reduced by eliminating those descriptors which are inter-correlated beyond 0.90 (descriptor versus descriptor, $r > 0.9$) and showing a correlation of less than 0.1 with the biological endpoints (descriptor versus activity, $r < 0.1$), prior to model development procedure. In doing so 122 descriptors were obtained and will be employed to explain the biological actions of titled compounds.

Development and validation of model

QSAR models have been developed, in the present study, using the combinatorial protocol in multiple linear regression (CP-MLR) [26-30] and partial least squares (PLS) [31-33] procedures. CP-MLR is a "filter"-based variable selection procedure and the embedded filters make the variable selection process efficient and lead to unique solution. The fear of existence of "chance correlations" in using large descriptor pools for multilinear QSAR/QSPR studies [34,35] overcome by randomization test [36, 37] in which each cross-validated CP-MLR recognized model has been subjected to repeated randomization (100 simulation runs) of the biological responses. The datasets with randomized response vector have been reassessed by multiple regression analysis. The resulting regression equations, if any, with correlation coefficients better than or equal to the one corresponding to unscrambled response data were counted. This has been used as a measure to express the percent chance correlation of the model under scrutiny.

Validation of the derived model is necessary to test its prediction and generalization within the study domain. A number of statistical parameters such as r (the multiple correlation coefficient), s (the standard deviation), F (the F ratio between the variances of calculated and observed activities), and Q^2_{LOO} (the cross-validated index from leave-one-out procedure) have been obtained to access its overall statistical significance, for each model derived in n data points. In case of internal validation, Q^2_{LOO} is used as a criterion of both robustness and predictive ability of the model. A value greater than 0.5 of Q^2 index suggests a statistically significant model. The predictive power of derived model is based on test set compounds. The model obtained from training set has a reliable predictive power if the value of the r^2_{Test} (the squared correlation coefficient between the observed and predicted values of compounds from test set) is greater than 0.5. Additional statistical parameters such as, the Akaike's information criterion, AIC [38, 39], the Kubinyi function, FIT [40,41] and the Friedman's lack of fit, LOF [42], have also been calculated to further validate the derived models. The AIC takes into account the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit. The FIT, closely related to the F-value, proved to be a useful parameter for assessing the quality of the models. A model which is derived in k independent descriptors, its F-value will be more sensitive if k is small while it becomes less sensitive if k is large. The FIT, on the other hand, will be less sensitive if k is small whereas it becomes more sensitive if k is large. The model that produces the lowest AIC value and highest FIT value is considered potentially the most useful and the best. The LOF factor takes into account the number of terms used in the equation and is not biased, as are other indicators, toward large number of parameters.

Applicability domain

The usefulness of a model is based on its accurate prediction ability for new congeners. A model is valid only within its training domain and new compounds must be assessed as belonging to the domain before the model is applied. The applicability domain (AD) is evaluated by the leverage values for each compound [43]. A Williams plot (the plot of standardized residuals versus leverage values (h)) is constructed, which can be used for a simple graphical detection of both the response outliers (Y outliers) and structurally influential chemicals (X outliers) in the model. In this plot, the AD is established inside a squared area within $\pm x$ standard deviations and a leverage threshold h^*, which is generally fixed at $3(k + 1)/n$ (n is the number of training set compounds and k is the number of model parameters), whereas $x = 2$ or 3. If the compounds have a high leverage value ($h > h^*$), then the prediction is not trustworthy. On the other hand, when the leverage value of a compound is lower than the threshold value, the probability of accordance

between predicted and observed values is as high as that for the training set compounds.

RESULTS AND DISCUSSION

QSAR results

A derived model equation(s), using a pool of descriptors of different descriptor classes, provides an opportunity to unravel the phenomenon under study i.e. the concepts embedded in the descriptor classes relate the biological actions revealed by the compounds. Primary observation of the data set revealed that one compound (S. No. 22, Table 1) does not fit in the trend of data set. Thus this compound has been excluded in deriving QSAR models. There are many reasons for their occurrence in QSAR studies; for example, chemicals might be acting by a mechanism different from that of the majority of the data points. It is also likely that outlier might be a result of a random experimental error that could be significant when analyzing a large data set.

For the purpose of modeling study, 05 (one fifth of total active) compounds have been included in the test set for the validation of the models derived from remaining 20 training set compounds. A total number of 122 relevant descriptors from 0D- to 2D- classes, which were obtained after the reduction of descriptor data set, have been subjected to CP-MLR analysis with default "filters" set in it. Statistical models in two and three descriptors have been explored to achieve the best relationship correlating GPR119 agonistic activity. All the models obtained in two descriptors were having the r^2_{Test} value less than 0.5. The obtained, all the three models, in three descriptors are given below through Eqs (1) to (3). These models (with 122 descriptors) were identified in CP-MLR by successively incrementing the filter-3 with increasing number of descriptors (per equation). For this, the optimum r-bar value of the preceding level model (=0.699, r-bar value of

the two parameter model having highest r^2_{Test}) has been used as the new threshold of filter-3 for the next generation.

$pEC_{50} = 7.980 + 3.264(0.936)GGI8 - 2.038(0.500)ATS7e - 1.702(0.485)GATS1e$

$n = 20$, $r = 0.853$, $s = 0.559$, $F = 14.290$, $Q^2_{LOO} = 0.549$, $Q^2_{L5O} = 0.689$

$r^2_{Test} = 0.788$, $FIT = 1.478$, $LOF = 0.511$, $AIC = 0.470$　　　　　　　　　　(1)

$pEC_{50} = 7.458 + 3.461(1.038)GGI8 - 2.003(0.536)ATS7e - 3.475(1.173)Hy$

$n = 20$, $r = 0.830$, $s = 0.598$, $F = 11.846$, $Q^2_{LOO} = 0.531$, $Q^2_{L5O} = 0.510$

$r^2_{Test} = 0.700$, $FIT = 1.225$, $LOF = 0.584$, $AIC = 0.537$　　　　　　　　　　(2)

$pEC_{50} = 8.151 + 2.585(0.891)LP1 - 1.993(0.444)BELp8 - 2.535(0.562)MATS7m$

$n = 20$, $r = 0.826$, $s = 0.604$, $F = 11.477$, $Q^2_{LOO} = 0.575$, $Q^2_{L5O} = 0.612$

$r^2_{Test} = 0.538$, $FIT = 1.187$, $LOF = 0.597$, $AIC = 0.548$　　　　　　　　　　(3)

where n, r, s and F represent respectively the number of data points, the multiple correlation coefficient, the standard deviation and the F-ratio between the variances of calculated and observed activities. In above and all follow-up regression equations, the values given in the parentheses are the standard errors of the regression coefficients. The signs of the regression coefficients suggest the direction of influence of explanatory variables in the models. The positive regression coefficient associated to a descriptor will augment the activity profile of a compound while the negative coefficient will cause detrimental effect to it. In the randomization study (100 simulations per model), none of the identified models has shown any chance correlation.

The participated descriptors, ATS7e, MATS7m and GATS1e, in above models belong to 2D-AUTO class. It is apparent from the above mentioned equations that a lower values of Broto-Moreau autocorrelation of a topological structure of lag-7 weighted by atomic Sanderson electronegativities (ATS7e), Moran autocorrelation of lag-6 weighted by atomic masses and Geary autocorrelation of lag-1 weighted by atomic Sanderson electro negativities would be helpful to elevate the agonistic activity. The topological class descriptor LP1 (Lovasz-Pelikan index) and Galvez class descriptor (8th order Galvez topological charge index, GGI8) shown positive correlation to the activity suggesting higher values of these as beneficial to the activity. The negative sign of correlation coefficient of modified Burden eigenvalue (BCUT) class descriptor BELp8 (lowest eigenvalue n.8 of Burden matrix/weighted by atomic polarizabilities and PROP class descriptor Hy (hydrophilic factor) advocated that a lower value of descriptor BELp8 and less hydrophilic factor or nature of molecule would be advantageous to the activity.

The three descriptor models could estimate nearly 73% variance in observed activity of the compounds. Considering the number of observation in the dataset, models with up to four

descriptors were explored through CP-MLR and the result was 27 four-parameter models with test set $r^2 > 0.50$ sharing 40 descriptors among them. The selected models, in four parameters are given below as Eqs. (4)-(7). The shared descriptors along with their brief physical meaning, average regression coefficients, and total incidence are listed in Table 3, which will serve as a measure of their estimate across these models.

Table 3. Identified descriptors[a] along with their class, average regression coefficient and incidence[b], in modeling the GPR119 agonistic activities of indole-based compounds.

Descriptor class, average regression coefficient and (incidence)	
Constitutional descriptors (**CONST**):	AMW, 2.560(1); Mv, 2.958(1), -1.937(1); nAT, 3.015(1); nDB, 0.634(1); nO,-1.343(1)
Topological descriptors (**TOPO**):	ZM2V, 2.406(1); MSD, -3.249(2); MAXDP, 1.889(1); X1A, -1.964(1); PW4, -1.570(1); AECC, -1.989(1); IC1, 3.103(1); LP1, 2.517(1)
Molecular walk counts (**MWC**):	MWC10, -0.860(1)
Modified Burden Eigen values (**BCUT**):	BEHm2, 2.038(1); BELm5, 2.459(7); BEHv5, 2.113(1); BELv1, -1.229(1); BELv2, -1.528(1); BELe8,-1.596(5); BELp8, -1.616(2)
Galvez topological charge indices (**GALVEZ**):	GGI5, -1.342(1); GGI8, 3.396(18)
2D autocorrelations (**2D-AUTO**):	ATS6v, -3.354(1); ATS6e, -2.152(1); ATS7e, -2.466(9); ATS7p, -3.955(5); MATS6m, 1.759(2); MATS7m, -2.651(1); MATS5v, -2.126(2); MATS6v, 1.977(2); MATS1e, 1.529(1); MATS3p, -2.721(4); GATS1e, -2.382(3); GATS4e, 2.088(1)
Atom centered fragments (**ACF**)	C-008, -1.072(1); C-029, 1.136(2); H-046, 2.491(4); H-050, -1.647(10)
Molecular properties (**PROP**)	Hy, -3.595(6)

[a]The descriptors are identified from the four parameter models for PPARγ binding activity transactivation activity emerged from CP-MLR protocol with filter-1 as 0.79, filter-2 as 2.0, filter-3 as 0.822 and filter-4 as $0.3 \leq q^2 \leq 1.0$ with a training set of 20 compounds. [b] The average regression coefficient of the descriptor corresponding to all models and the total number of its incidence. The arithmetic sign of the coefficient represents the actual sign of the regression coefficient in the models. **CONST**: AMW, average molecular weight; Mv, mean atomic van der Waals volume (scaled on Carbon atom); nAT, number of atoms; nDB, number of double bonds; nO, number of Oxygen atoms; **TOPO:** ZM2V, second Zagreb index by valence vertex degrees; MSD, mean square distance index (Balaban); MAXDP, maximal electrotopological positive variation; X1A, average connectivity index chi-1; PW4, path/walk 4-Randic shape index; AECC, average eccentricity; IC1, information content index (neighborhood symmetry of 1-order) ; LP1, Lovasz-Pelikan index (leading eigenvalue); **MWC**: MWC10, molecular walk count of order 10;

BCUT: BEHm2, highest eigenvalue n.2 of Burden matrix/weighted by atomic masses; BELm5, lowest eigenvalue n.5 of Burden matrix/weighted by atomic masses; BEHv5, highest eigenvalue n.5 of Burden matrix/weighted by van der Waals volumes; BELv1, lowest eigenvalue n.1 of Burden matrix/weighted by van der Waals volumes BELv2, lowest eigenvalue n.2 of Burden matrix/weighted by van der Waals volumes; BELe8, lowest eigenvalue n.8 of Burden matrix/weighted by atomic Sanderson electronegativities, BELp8, lowest eigenvalue n.8 of Burden matrix/weighted by atomic polarizabilities; **2D-AUTO:** ATS6v, Broto-Moreau autocorrelation of a topological structure - lag 6 / weighted by atomic van der Waals volumes; ATS6e, Broto-Moreau autocorrelation of a topological structure - lag 6 / weighted by atomic Sanderson electronegativities; ATS7e, Broto-Moreau autocorrelation of a topological structure - lag 7 / weighted by atomic Sanderson electronegativities; ATS7p,Broto-Moreau autocorrelation of a topological structure - lag 7 / weighted by atomic polarizabilities; MATS6m, Moran autocorrelation - lag 6 / weighted by atomic masses; MATS7m, Moran autocorrelation - lag 7 / weighted by atomic masses; MATS5v, Moran autocorrelation - lag 5 / weighted by atomic van der Waals volumes; MATS6v, Moran autocorrelation - lag 6 / weighted by atomic van der Waals volumes; MATS1e, Moran autocorrelation of lag-1/ weighted by atomic Sanderson electronegativities; MATS3p; Moran autocorrelation of lag-3/ weighted by atomic polarizabilities; GATS1e, Geary autocorrelation of lag-1/ weighted by atomic Sanderson electronegativities; GATS4e, Geary autocorrelation of lag-4/ weighted by atomic Sanderson electronegativities; GALVEZ: GGI5, topological charge index of order 5; GGI8, topological charge index of order 8; **ACF:**C-008, CHR2X; C-029, R--CX—X; H-046, H attached to C0(sp3) no X attached to next C; H-050, H attached to heteroatom; **PROP:** Hy, hydrophilic factor.

$$pEC_{50} = 6.142 + 2.359(0.639)BELm5 + 4.496(0.825)GGI8 - 3.661(0.828)ATS7p$$

$$- 1.185(0.410)H\text{-}050$$

$n = 20$, $r = 0.888$, $s = 0.509$, $F = 14.010$, $Q^2_{LOO} = 0.651$, $Q^2_{L5O} = 0.607$

$r^2_{Test} = 0.550$, $FIT = 1.556$, $LOF = 0.541$, $AIC = 0.432$ (4)

$$pEC_{50} = 7.812 + 2.891(0.833)GGI8 - 2.094(0.502)ATS7e - 1.072(0.372)C\text{-}008$$

$$- 1.453(0.409)H\text{-}050$$

$n = 20$, $r = 0.888$, $s = 0.509$, $F = 13.994$, $Q^2_{LOO} = 0.624$, $Q^2_{L5O} = 0.703$

$r^2_{Test} = 0.745$, $FIT = 1.554$, $LOF = 0.541$, $AIC = 0.433$ (5)

$$pEC_{50} = 6.316 + 2.163(0.922) GGI8 - 2.820(0.490) GATS1e + 2.088(0.589) GATS4e$$

$$+ 1.307(0.355)C\text{-}029$$

$n = 20$, $r = 0.887$, $s = 0.511$, $F = 13.904$, $Q^2_{LOO} = 0.604$, $Q^2_{L5O} = 0.514$

$r^2_{Test} = 0.697$, FIT = 1.544, LOF = 0.544, AIC = 0.435 (6)

$pEC_{50} = 8.553 + 2.516(0.760)\ LP1 - 2.066(0.380)BELp8 - 2.651(0.481)MATS7m$

$- 2.459(0.927)He$

n = 20, r = 0.885, s = 0.515, F = 13.610, $Q^2_{LOO} = 0.699$, $Q^2_{L5O} = 0.637$

$r^2_{Test} = 0.540$, FIT = 1.512, LOF = 0.553, AIC = 0.442 (7)

Nearly 79% variance in the observed activity has been accounted by these models. None of the CP-MLR identified model has shown any chance correlation in the randomization study (100 simulations per model). The values of Q^2 index, greater than a specified cutoff (0.5), hint that derived models are reasonable robust QSAR models. The pEC_{50} values of training set compounds calculated using Eqs. (4) to (7) and predicted from LOO procedure have been included in Table 4.

Table 4: Observed and modeled GPR119 activity of indole-based agonists.

S. No.	pEC50(M)[a]										
	Obsd[b]	Eq. (4)		Eq. (5)		Eq. (6)		Eq. (7)		PLS	
		Calc	Pred[c]	Calc	Pred[c]	Calc	Pred[c]	Calc	Pred[c]	Calc	Pred[c]
1	5.85	5.87	5.88	6.16	6.28	5.95	6.06	6.10	6.26	6.31	6.52
2	6.92	6.47	6.37	6.54	6.43	7.23	7.34	6.49	6.41	6.82	6.79
3	7.17	6.43	6.35	7.08	7.07	6.51	6.46	6.65	6.58	6.80	6.77
4	5.60	5.38	4.97	5.36	4.97	6.43	6.52	6.52	6.66	5.96	6.04
5[d]	5.59	6.42	_[d]	5.70	_[d]	6.46	_[d]	6.41	_[d]	6.26	_[d]
6	8.08	8.21	8.31	7.73	7.47	7.45	7.18	7.31	7.17	7.66	7.55
7	5.62	5.84	6.25	5.86	6.25	6.04	6.39	5.63	5.87	5.38	5.31
8	7.60	7.06	6.91	6.89	6.72	7.24	7.11	7.43	7.38	7.15	7.12
9[d]	6.28	6.89	_[d]	6.90	_[d]	6.79	_[d]	7.23	_[d]	6.94	_[d]
10	6.11	6.74	6.85	6.62	6.71	6.73	6.89	6.26	6.33	6.54	6.61
11	6.80	6.98	7.03	6.89	6.92	6.39	6.32	7.11	7.15	6.88	6.88
12	5.32	5.69	5.80	6.01	6.15	5.09	4.98	5.38	5.42	5.43	5.46
13	7.60	7.89	7.96	7.42	7.40	7.40	7.37	7.46	7.43	7.40	7.39
14	7.68	7.56	7.54	7.44	7.42	7.34	7.29	7.46	7.42	7.35	7.33
15	7.36	7.39	7.39	7.05	7.03	7.37	7.37	7.56	7.60	7.22	7.21
16	5.77	6.48	6.59	6.07	6.20	6.08	6.13	5.69	5.66	5.67	5.65
17[d]	8.17	9.12	_[d]	7.78	_[d]	8.01	_[d]	8.18	_[d]	8.08	_[d]
18	8.41	8.71	8.93	8.74	8.93	8.29	8.22	8.22	8.00	8.61	8.67
19	6.31	6.57	6.63	6.23	6.19	6.82	6.86	6.76	6.82	6.63	6.65
20[d]	7.06	7.15	_[d]	6.40	_[d]	6.53	_[d]	6.66	_[d]	6.65	0.00
21	7.77	7.02	6.95	6.72	6.44	7.40	7.17	6.75	6.66	7.11	7.05
22[e]	5.06	_[e]	_[e]	_[e]	_[e]	_[e]	_[e]	_[e]	_[e]	_[e]	_[e]
23	6.28	6.80	6.98	6.99	7.18	6.12	6.07	6.79	6.83	6.59	6.64
24[d]	7.80	8.04	_[d]	8.21	_[d]	7.56	_[d]	7.21	_[d]	6.78	_[d]
25	7.96	7.90	7.88	8.35	8.46	7.58	7.33	8.34	8.46	8.31	8.39
26	8.11	7.35	7.14	8.16	8.17	8.86	9.39	8.41	8.51	8.49	8.60

[a] On molar basis; [b] Taken from ref. [23]; [c] Leave-one-out (LOO) procedure; [d] Compound included in test set; [e] Oulier compound.

The models (4) to (7) are validated externally with test set of 5 compounds mentioned in Table 1. The test set r^2 (r^2_{Test}) values greater than 0.5 of these models reflect that these models have satisfactory external validation capability. The predicted activity values of test set compounds are in tune to the observed ones and the same is mentioned in Table 4. The plot showing goodness of fit between observed and calculated activities for the training and test set compounds is given in Figure 2.
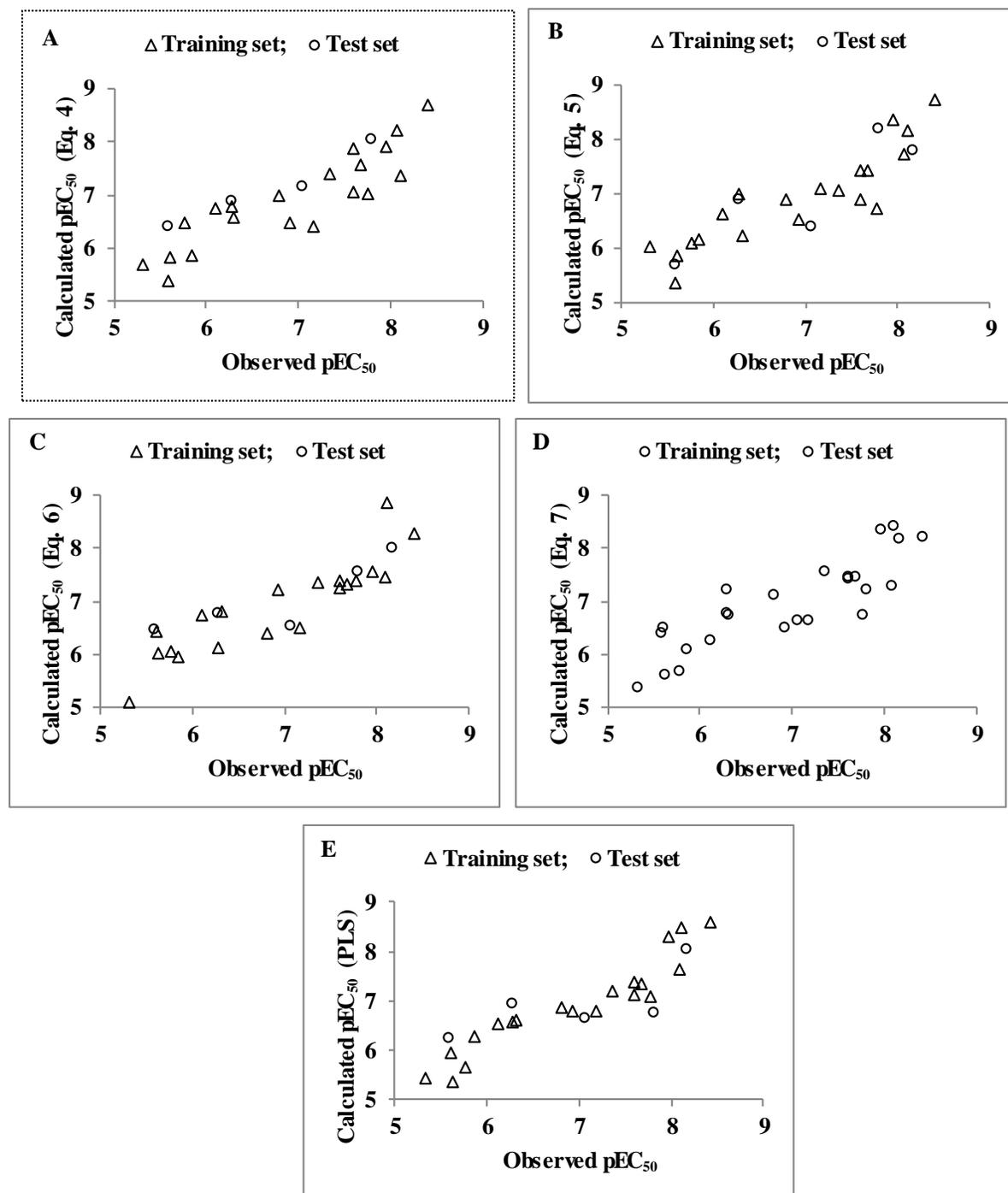

Figure 2: Plot of observed and calculated pEC$_{50}$ values of training- and test-set compounds

for indole-based GPR119 agonists.

The newly appeared descriptors in above models C-008, C-029 and H-050 are from the atom centered fragment (ACF) class of descriptors. Descriptor BELm5 belong to BCUT class and the remaining two ATS7p and GATS4e are 2D-autocorrelations. The signs of regression coefficients of ACF descriptors suggested absence of CHR2X type fragment (descriptor C-008) and H attached to heteroatom (descriptor H-050) and presence of R--CX--X type structural fragment (descriptor C-029) beneficial to the activity. Additionally, higher values of descriptors BELm5 (lowest eigenvalue n.5 of Burden matrix/weighted by atomic masses) and GATS4e (Geary autocorrelation of lag-4/weighted by atomic Sanderson electro negativities), and a lower value of descriptor ATS7p (Broto-Moreau autocorrelation of a topological structure - lag 7/weighted by atomic polarizabilities) would be advantageous the agonistic activity.

A partial least square (PLS) analysis has been carried out on these 13 descriptors, emerged in above mentioned models (4) to (7), to facilitate the development of a "single window" structure–activity model. For the purpose of PLS, the descriptors have been autoscaled (zero mean and unit SD) to give each one of them equal weight in the analysis. In the PLS cross-validation, two components are found to be the optimum for these 13 descriptors and they explained 87.79% variance in the activity. The MLR-like PLS coefficients of these 13 descriptors are given in Table 5.

Table 5: PLS and MLR-like PLS models from the 13 descriptors of four parameter CP-MLR models for GPR119 agonistic activities.

A: PLS equation								
PLS components				PLS coefficient (s.e.)[a]				
Component-1				-0.561(0.052)				
Component-2				0.141(0.045)				
Constant				6.916				

B: MLR-like PLS equation									
S. No.	Descriptor	MLR-like coefficient[b]	(f.c.)[c]	Order	S. No.	Descriptor	MLR-like coefficient[b]	(f.c.)[c]	Order
1	LP1	0.030	0.013	13	8	GATS1e	-0.217	-0.097	6
2	BELm5	0.163	0.073	8	9	GATS4e	0.133	0.059	9
3	BELp8	-0.230	-0.103	3	10	C-008	-0.100	-0.045	11
4	GGI8	0.235	0.105	2	11	C-029	0.199	0.088	7
5	ATS7e	-0.272	-0.121	1	12	H-050	-0.228	-0.102	4
6	ATS7p	-0.089	-0.040	12	13	Hy	-0.126	-0.056	10
7	MATS7m	-0.222	-0.099	5		Constant = 7.774			

C: PLS regression statistics	Values
n	20
r	0.937
s	0.363

F	61.307
FIT	5.108
LOF	0.175
AIC	0.178
Q^2_{LOO}	0.828
Q^2_{L5O}	0.808
r^2_{Test}	0.538

[a] Regression coefficient of PLS factor and its standard error. [b]Coefficients of MLR-like PLS equation in terms of descriptors for their original values; [c]f.c. is fraction contribution of regression coefficient, computed from the normalized regression coefficients obtained from the autoscaled (zero mean and unit s.d.) data.

For the sake of comparison, the plot showing goodness of fit between observed and calculated activities (through PLS analysis) for the training and test set compounds is also given in Figure 2. The fraction contribution of normalized regression coefficients of these descriptors to the activity is shown in Figure 3.
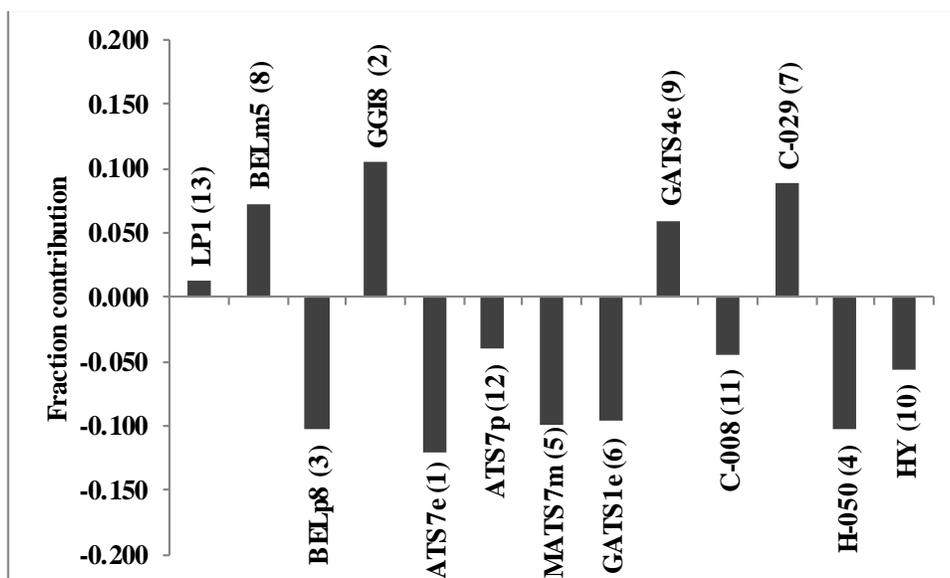
Figure 3: **Plot of fraction contribution of MLR-like PLS coefficients (normalized) against 13 CP-MLR identified descriptors (Table 5) associated with GPR119 agonistic activity of indole-based derivatives.**

The PLS analysis has suggested ATS7e as the most determining descriptor for modeling the agonistic activity of the compounds (descriptor S. No. 5 in Table 5; Figure 3). The other descriptors in decreasing order of significance are GGI8, BELp8, H-050, MATS7m, GATS1e, C-029, BELm5, GATS4e, Hy, C-008, ATS7p and LP1 and convey same inference in the PLS model as well. It is also observed that PLS model from the dataset devoid of these 13 descriptors is inferior in explaining the activity of the analogues.

Applicability domain (AD)

On analyzing the model AD in the Williams plot, shown in Figure 4, of the model based on the whole dataset (Table 6), it has appeared that one compound (S. No. **22**, Table 1) was identified as an obvious outlier for the GPR119 agonistic activity if the limit of normal values for the *Y* outliers (response outliers) was set as 2 (standard deviation) units. An outlier to a QSAR is identified normally by having a large standard residual activity and can indicate the limits of applicability of QSAR models. Two compounds, listed in Table 1 at S. No. **7** and 24 found to have leverage (h) values greater than the threshold leverage (h*) suggesting these compounds as chemically influential compound. For both the training-set and test-set, the suggested model matches the high quality parameters with good fitting power and the capability of assessing external data. Furthermore, all of the compounds were within the applicability domain of the proposed model and were evaluated correctly.

Table 6: **Models derived for the whole data set (n = 26) for the GPR119 agonistic activity in descriptors identified through CP-MLR.**

Model	r	s	F	Q^2_{LOO}	Eq.
$pEC_{50} = 6.237 + 1.680(0.679)BELm5 + 4.123(0.813)GGI8 - 3.263(0.841)ATS7p - 1.088(0.527)H-050$	0.786	0.692	8.484	0.497	(4a)
$pEC_{50} = 7.855 + 2.836(0.567)GGI8 - 2.118(0.446)ATS7e - 1.255(0.377)C-008 - 1.421(0.452)H-050$	0.847	0.595	13.345	0.592	(5a)
$pEC_{50} = 6.398 + 2.663(0.694)GGI8 - 2.844(0.660)GATS1e + 1.794(0.699)GATS4e + 0.628(0.425)C-029$	0.765	0.721	7.414	0.292	(6a)
$pEC_{50} = 8.523 + 2.745(0.698)LP1 - 1.938(0.432)BELp8 - 2.882(0.536)MATS7m - 2.472(1.116)Hy$	0.829	0.625	11.600	0.365	(7a)
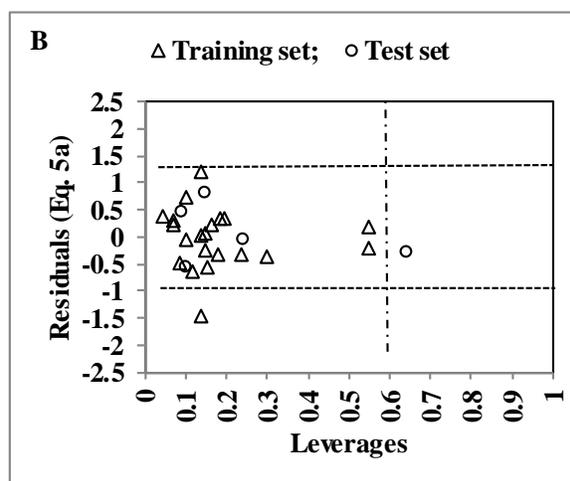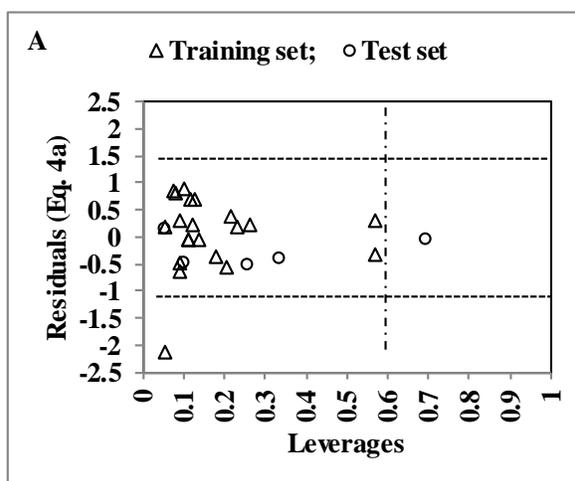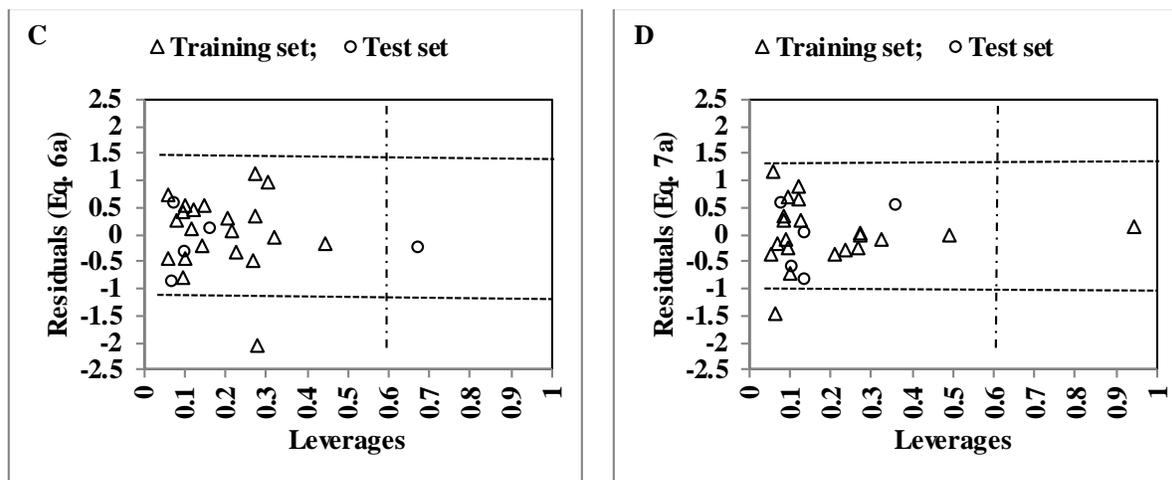
Figure 4: Williams plot for the training-set and test- set compounds for GPR119 agonistic activity. The horizontal dotted line refers to the residual limit (±2×standard deviation) and the vertical dotted line represents threshold leverage h* (= 0.580).

CONCLUSION

QSAR study has been carried out on the GPR119 agonistic activity of indole-based derivatives in 0D- to 2D-Dragon descriptors. The derived QSAR models have revealed that the atomic Sandersons electronegativities weighted and charge accounting descriptors ATS7e, GATS1e, GATS4e and GGI8, molecular mass weighted descriptors, MATS7m and BELm5, and atomic polarizabilities weighted descriptors ATS7p and BELp8, and molecular topology accounting feature Lovasz-Pelikan index (LP1) played a pivotal role in rationalization of GPR119 agonistic activity of titled compounds. Hydrophilic factor (Hy) and certain structural fragments, such as CHR2X (C-008), R--CX--X (C-008) and H attached to heteroatom (H-050) are also predominant to explain GPR119 agonistic actions of indole-based derivatives. PLS analysis has also corroborated the dominance of CP-MLR identified descriptors. Applicability domain analysis revealed that the suggested model matches the high quality parameters with good fitting power and the capability of assessing external data and all of the compounds was within the applicability domain of the proposed model and were evaluated correctly.

COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgements

Authors are thankful to their institution for providing necessary facilities to complete this study.

Disclosure of conflict of interest

The authors declare no conflict of interest.

REFERENCES

1. Millatt LJ, Hanf R and Hum DW. (2011). Present and future challenges in Type 2 diabetes.

Drug Discovery World Summer 2011, 9-15.

2. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, Lin JK, Farzadfar F, Khang Y-H, Stevens GA, Rao M, Ali MK, Riley LM, Robinson CA and Ezzati M. (2011). National, regional, and global trends in fasting plasma glucose and diabetesprevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. Lancet, 378,31-40.

3. Koro CE, Bowlin SJ, Bourgeois N and Fedder DO. (2004). Glycemic control from 1988to 2000 among U.S. adults diagnosed with type 2 diabetes: a preliminary report. DiabetesCare 27, 17-20.

4. Lovshin JA and Drucker DJ. (2009). Incretin-based therapies for type 2 diabetes mellitus. Nature Reviews Endocrinology. 5, 262-269.

5. Verspohl EJ. (2009). Novel therapeutics for type 2 diabetes: incretin hormone mimetics (glucagon-like peptide-1 receptor agonists) and dipeptidyl peptidase-4 inhibitors. Pharmacology and Therapeutics, 124, 113-138.

6. Mikhail N. (2008). Incretin mimetics and dipeptidyl peptidase 4 inhibitors in clinical trials for the treatment of type 2 diabetes. Expert Opinion on Investigative Drugs 17, 845-853.

7. Shah U and Kowalski TJ. (2010). GPR119 agonists for the potential treatment of type 2 diabetes and related metabolic disorders Vitamins and Hormones, 84, 415-448

8. Shah U. (2009). GPR119 agonists: a promising new approach for the treatment of type 2 diabetes and related metabolic disorders. Current Opinion in Drug Discovery and Development, 12, 519-532.

9. Jones RM, Leonard JN, Buzard DJ and Lehmann, J. (2009). GPR119 agonists for the treatment of type 2 diabetes. Expert Opinion on Therapeutic Patents, 19, 1339-1359.

10. Fyfe MC, McCormack JG, Overton HA, Procter MJ and Reynet C. (2008). GPR119 agonists as potential new oral agents for the treatment of type 2 diabetes and obesity Expert Opinion on Drug Discovery, 3, 403-413.

11. Overton HA, Fyfe MC and Reynet C. (2008). GPR119, a novel G protein-coupled receptor target for the treatment of type 2 diabetes and obesity. British Journal of Pharmacology, 153 (Suppl 1), S76-S81.

12. Ohishi T and Yoshida S. (2012). The therapeutic potential of GPR119 agonists for type 2 diabetes. Expert Opinion on Investigative Drugs, 21, 321-328.

13. Baggio LL and Drucker DJ. (2006). Therapeutic approaches to preserve islet mass in type 2 diabetes. Annual Review of Medicine, 57, 265-281.

14. Aaboe K, Krarup T, Madsbad S and Holst JJ. (2008). GLP-1: physiological effects and potential therapeutic applications. Diabetes, Obesity and Metabolism, 10, 994-1003.

15. Overton HA, Babbs AJ, Doel SM, Fyfe MC, Gardner LS, Griffin G, Jackson HC, Procter MJ, Rasamison CM, Tang-Christensen M, Widdowson PS, Williams GM and Reynet C. (2006). Deorphanization of a G protein-coupled receptor for oleoylethanolamide and its use in the discovery of small-molecule hypophagic agents. Cell Metabolism, 3, 167-175.

16. Fyfe, M, Mccormack J, Overton H, Procter M and Reynet C. (2008). PSN821: A novel oral GPR119 agonist for the treatment of type 2 diabetes producing substantial glucose lowering and weight loss in rats. Abstract of papers. American Diabetes Association 68th Annual Scientific Sessions, San Francisco, CA, USA, Abstract 297-OR.

17. Flock G, Holland D, Seino Y and Drucker DJ. (2011). GPR119 regulates murine glucose homeostasis through incretin receptor-dependent and independent mechanisms. Endocrinology, 152, 374-383.

18. Goodman ML, Dow J, Van Vliet AA, Pleszko A , Lockton JA. (2011). Orally administered GPR119 agonist PSN821 shows clinically significant glucose lowering and other potential cardiometabolic benefits in patients with type 2 diabetes. Diabetologia, 54, S85.

19. Roberts B, Gregoire FM, Karpf DB, Clemens E, Lavan B, Johnson J, Mcwherter CA, Martin R and Wilson M. (2009). MBX-2982, a Novel oral GPR119 agonist for the treatment of type 2 diabetes: results of single & multiple dose studies. Abstract of papers. American Diabetes Association 69th Annual Scientific Sessions, New Orleans, LA, USA, Abstract 164-OR.

20. Roberts B, Karpf DB, Martin R, Lavan B, Wilson M and Mcwherter CA. (2010). MBX-2982, a novel GPR119 agonist, shows greater efficacy in patients with the most glucose intolerance: results of a phase 1 study with an improved formulation. Abstract of papers. American Diabetes Association 70th Annual Scientific Sessions, Orlando, FL, USA, Abstract 603-P.

21. Goodman, ML, Dow J, Van Vliet AA, Hadi S, Karbiche D, Lockton JA. (2011). The Novel GPR119-receptor agonist PSN821 shows glucose lowering and decreased energy Intake in patients with T2DM after 14 days treatment. Diabetes, 60, A273.

22. Nunez DJ, Lewis EW, Swan S, Bush MA, Cannon C, Mcmullen SL, Collins DA and Feldman PL. (2010). A study in healthy volunteers to assess the safety, tolerability, pharmacokinetics (PK) and pharmacodynamics (PD) of single and multiple doses of

GSK1292263, a novel GPR119 agonist. Abstract of papers. American Diabetes Association 70th Annual Scientific Sessions, Orlando, FL, USA, 2010; Abstract 80-OR.

23. Sato K, Sugimoto H, Rikimaru K, Imoto H, Kamaura H, Negoro N, Tsujihata Y, Miyashita H, Odani T and Murata T. (2014). Discovery of a novel series of indoline carbamate and indolinylpyrimidine derivatives as potent GPR119 agonists. Biooganic and Medicinal Chemistry, 22, 1649–1666.

24. Chemdraw ultra 6.0 and Chem3D ultra, Cambridge Soft Corporation, Cambridge, USA. http://www.cambridgesoft.com

25. Dragon software (version 1.11-2001) by Todeschini R.; Consonni V. Milano, Italy. http://www.talete.mi.it/dragon.htm

26. Prabhakar YS. (2003). A combinatorial approach to the variable selection in multiple linear regression: analysis of Selwood et al. Data Set-a case study. QSAR and Combinatorial Science, 22, 583-595.

27. Sharma S, Prabhakar YS, Singh P and Sharma BK. (2008). QSAR study about ATP sensitive potassium channel activation of cromakalim analogues using CP-MLR. approach. European Journal of Medicinal Chemistry, 43, 2354-2360.

28. Sharma S, Sharma BK, Sharma SK, Singh P and Prabhakar YS. (2009). Topological descriptors in modeling the agonistic activity of human A3 adenosine receptor ligands: The derivatives of 2-Chloro-N[6]-substituted-4'-thioadenosine-5'-uronamide. European Journal of Medicinal Chemistry, 44, 1377-1382.

29. Sharma BK, Pilania P, Singh P and Prabhakar YS. (2010). Combinatorial protocol in multiple linear regression/partial least-squares directed rationale for the caspase-3 inhibition activity of isoquinoline-1,3,4-trione derivatives. SAR and QSAR in Environmental Research, 21, 169-185.

30. Sharma BK, Singh P, Sarbhai K and Prabhakar YS. (2010). A quantitative structure-activity relationship study on serotonin 5-HT6 receptor ligands: Indolyl and piperidinyl sulphonamides. SAR and QSAR in Environmental Research, 21, 369-388.

31. Wold S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. Technometrics, 20, 397–405.

32. Kettaneh N, Berglund A and Wold S. (2005). PCA and PLS with very large data sets. Computational Statistics and Data Analysis, 48, 69-85.

33. Stahle L and Wold S. (1988). Multivariate data analysis and experimental design. In: Ellis GP and West WB. (Eds.), Biomedical research. Progress in medicinal chemistry. Elsevier

Science Publishers, BV, Amsterdam.25, 291–338.

34. Topliss JG and Edwards RP. (1979). Chance factors in studies of quantitative structure–activity relationships. Journal of Medicinal Chemistry, 22, 1238–1244.

35. Katritzky AR, Dobchev DA, Slavov S and Karelson M. (2008). Legitimate utilization of large descriptor pools for QSAR/QSPR models. Journal of Chemical Information and Modeling, 48, 2207–2213.

36. So S-S and Karplus M. (1997). Three-dimensional quantitative structure–activity relationship from molecular similarity matrices and genetic neural networks. 1. Method and validation. Journal of Medicinal Chemistry, 40, 4347–4359.

37. Prabhakar YS, Solomon VR, Rawal RK, Gupta MK and Katti SB. (2004). CP-MLR/PLS directed structure–activity modeling of the HIV-1 RT inhibitory activity of 2,3-diaryl-1,3-thiazolidin-4-ones. QSAR and Combinatorial Science, 23, 234-244.

38. Akaike H. (1973). Information theory and an extension of the minimum likelihood principle. In: Petrov BN and Csaki F. (Eds.). Second international symposium on information theory. Budapest: Akademiai Kiado; p 267–281.

39. Akaike H. (1974). A new look at the statistical identification model. IEEE Transactions on Automatic Control, AC-19, 716–723.

40. Kubinyi H. (1994). Variable selection in QSAR studies. I. An evolutionary algorithm. Quantitative Structure-Activity Relationship, 13, 285–294.

41. Kubinyi H. (1994). Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution. Quantitative Structure-Activity Relationship,13,393–401.

42. Friedman J. (1990). In Technical Report No. 102. Laboratory for Computational Statistics, Stanford University, Stanford.

43. Gramatica P. (2007). Principles of QSAR models validation: internal and external. QSAR and Combinatorial Science, 26, 694-701.