(RESEARCH ARTICLE)

Check for updates

CP-MLR/PLS directed quantitative structure-activity relationship study on the CDK8 inhibitory activity: The derivatives of naphthyridine and isoquinoline

Raghuraj Parihar [1], Mukesh Meena [2], Dinesh Kumar Meena [2] and Brij Kishore Sharma [2, *]

[1] Department of Chemistry, Government College, Kota-324 001, Rajasthan, India.
[2] Department of Chemistry, Government College, Bundi-323 001, Rajasthan, India.

Abstract

The CDK8 and 7dF3 inhibition activity of naphthyridine and isoquinoline derivatives have been quantitatively analyzed in terms of Dragon descriptors. The statistically validated quantitative structure-activity relationship (QSAR) models provided rationales to explain the inhibition activities of these congeners. The descriptors identified through CP-MLR analysis for the CDK8 inhibitory activity have highlighted the role of sum of the topological distances between N..N (T(N..N)), distance/detour ring index of order 6 (D/Dr06), aromatic ratio (ARR), number of double bonds (nDB), number of 6-membered rings (nR06), number of sulphur atoms (nS), number of unsubstituted sp2 hybridized aromatic carbon atoms (nCaH), number of aliphatic N hydrazines (nN-N), number of aromatic primary amines (nNH2Ph) and certain atom centered fragments such as R--CH--R (C-024), X--CR..X (C-034), H attached to C1(sp3)/C0(sp2) (H-047) and RCO-N</>N-X=X (N-072) to quantify the inhibitory actions. The highest eigen value n.2 of Burden matrix/weighted by atomic polarizabilites (BEHp2) and atomic Sanderson electronegativities weighted Geary autocorrelation of lag 8 (descriptor GATS8e) have also shown prevalence to model the CDK8 inhibitory activity.

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. The models obtained from the descriptor pool that was chosen for the CDK8 inhibitory activity is able to explain nearly 74% variance in the observed 7dF3 activities of titled compounds.

Keywords: QSAR; CDK8 inhibitors; Combinatorial protocol in multiple linear regression (CP-MLR) analysis; Partial least square (PLS) analysis; Dragon descriptors; Naphthyridines; Isoquinolines

1. Introduction

The pivotal role played by cyclin-dependent kinases (CDKs) in the regulation of cell progression and gene transcription puts these as promising therapeutic targets for various diseases [1]. In the recent decades intensive efforts have been made in search of novel and potent CDK inhibitors. The unique member of the CDK family, CDK8 that is a transcriptional CDK, is a component of the mediator complex. By participation in a myriad of signaling pathways CDK8 regulates the transcription of related genes in a context-specific way [2]. The implications of CDK8 in colorectal and gastric cancers as an oncogene, is due to the activation of WNT signaling [3-6]. Because of the association of CDK8 in the development and progression of cancers its selective inhibition has been regarded as a promising approach for cancer therapy. A variety of small molecule modulators of CDK8 have been recently reported and these include Sorafenib [7], Senexin A [8], Cortistatin A [9], 6-azabenzothiophene derivatives [10]. The optimization of a 3,4,5-trisubstituted pyridine series led to potent, selective and orally bioavailable dual CDK8/19 ligandsCCT251545 and CCT251921 [11-13]. Mallinger et

*Corresponding author: Brij Kishore Sharma
Department of Chemistry, Government College Bundi, Rajasthan, India.

al has been reported a novel series of 2,8-disubstituted-1,6-naphthyridine- and 4,6-disubstituted-isoquinolines as dual CDK8/19 ligands by means of scaffold-hop approach [14].

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

2. Material and methods

2.1. Data-set

For present work the reported fifty naphthyridines and isoquinolines have been considered as the data set [14]. The general structure of these compounds is represented in Figure 1 and structural variations are mentioned in Table 1.
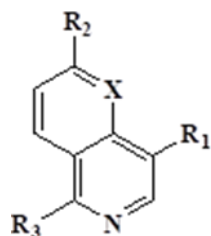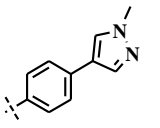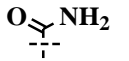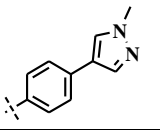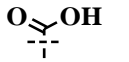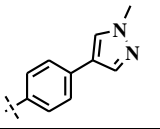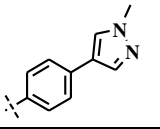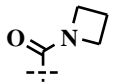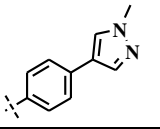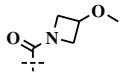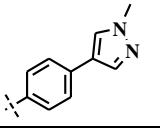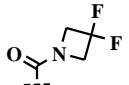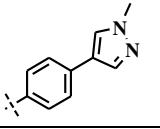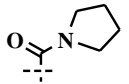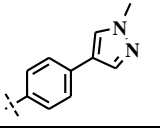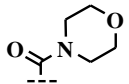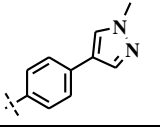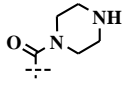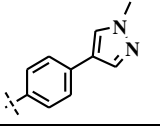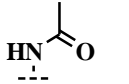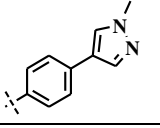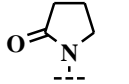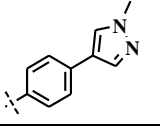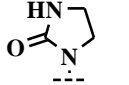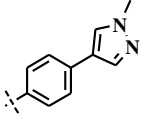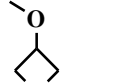

Figure 1 General structure of naphthyridine (X=N) and isoquinoline analogues
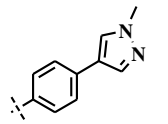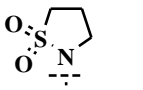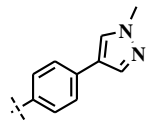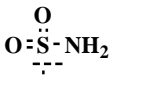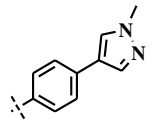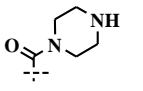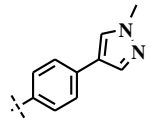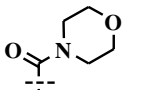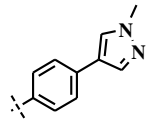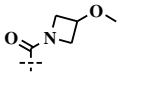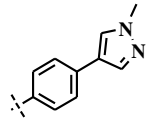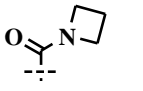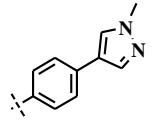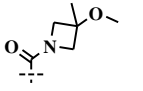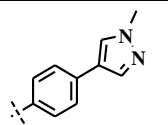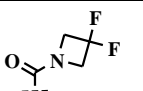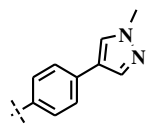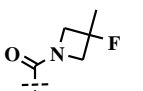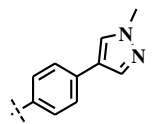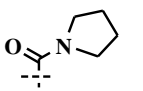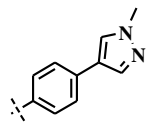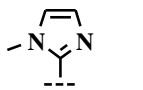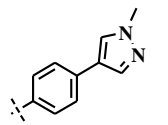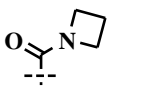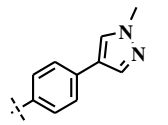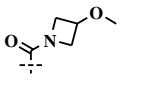
These derivatives were evaluated for their inhibition of CDK8 and WNT signaling in luciferase reporter assay in HEK293 7dF3 cells. Both the inhibition activities have also been reported in Table1 [14]. The inhibition activity, IC_{50}, represents the concentration of a compound to achieve 50% inhibition of CDK8 and 7dF3. The same is expressed as pIC_{50} on a molar basis and considered as the dependent variable for the present quantitative analysis. In the dataset, the initial assessment of activity with all descriptors has suggested the compound 23 as potential outlier. An outlier to a QSAR can indicate the limits of applicability of QSAR models. This outlier is not part of the data set. 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.

Table 1 Structural variations and observed CDK8 and 7dF3 inhibition activities of naphthyridine and isoquinoline derivatives

Cpd.	R₁	R₂	R₃	X	IC_{50} (nM)[a]	
					CDK8	7dF3
1			H	N	3.6	0.2
2[b]			H	N	5.4	3.4
3			H	N	6.9	4.6
4			H	N	0.9	1.7
5			H	N	5.3	11.2

6		O=C-NH₂	H	N	1.5	0.5
7		O=C-OH	H	N	4.7	590
8		C≡N	H	N	390	396
9[b]		(azetidine amide)	H	N	3.6	2
10[b]		(3-methoxyazetidine amide)	H	N	5.2	2.5
11[b]		(3,3-difluoroazetidine amide)	H	N	26	12.6
12		(pyrrolidine amide)	H	N	3.5	9.9
13		(morpholine amide)	H	N	13	27.8
14		(piperazine amide)	H	N	25	65
15[b]		HN-C(=O)CH₃	H	N	1.8	17.5
16[b]		(2-oxopyrrolidine)	H	N	2.4	30
17		(imidazolidinone)	H	N	1	4.2
18[b]		(3-methoxyazetidine)	H	N	816	1418

19			H	N	1195	2554
20			H	N	470	1946
21			H	N	1230	1232
22			H	N	16	45
23[c]			H	N	1.71	4.7
24[b]			H	C	4.5	0.6
25			H	C	2.1	1.1
26[b]			H	C	11	16
27			H	C	6.2	19.4
28[b]			H	C	4.9	34
29			H	C	34	16
30[b]			H	C	5.6	7.3
31			H	C	4.6	4.7

32			H	C	6.7	87
33			H	C	5.6	43
34			NH$_2$	N	100	559
35			NH$_2$	N	52	73
36[b]			NH$_2$	N	1.9	7.2
37			NH$_2$	N	21	10.7
38			NH$_2$	N	12	8.2
39			NH$_2$	N	17	5.8
40			NH$_2$	N	37	9.8
41[b]			NH$_2$	N	28	35.5
42			NH$_2$	N	160	87
43			NH$_2$	C	36	70
44			NH$_2$	C	12	21

45			NH_2	C	44	23.1
46			NH_2	C	12	17.9
47			NH_2	C	33	20.2
48			NH_2	C	29	76
49			NH_2	C	9.1	9
50[b]			NH_2	C	20	24

[a]Reference [14], IC_{50} represents the concentration of a compound to bring out 50% inhibition of CDK8 and 7dF3; [b]Compound included in test set; [c]"Outlier" compound not included in data set.

2.2. Molecular descriptors

The structures of the compounds (Table 1), under study, have been drawn in 2D ChemDraw [15] and were converted into 3D objects using the default conversion procedure implemented in the CS Chem3D Ultra. The generated 3D-structures of the compounds were subjected to energy minimization in the MOPAC module, using the AM1 procedure for closed shell systems, implemented in the CS Chem3D Ultra. This will ensure a well defined conformer relationship across the compounds of the study. All these energy minimized structures of respective compounds have been ported to DRAGON software [16] for computing the descriptors corresponding to 0D-, 1D-, and 2D-classes.

2.3. Development and validation of model

The combinatorial protocol in multiple linear regression (CP-MLR) [17-21] and partial least squares (PLS) [22-24] procedures have been used in the present work for developing QSAR models. The CP-MLR is a "filter"-based variable selection procedure, which employs a combinatorial strategy with MLR to result in selected subset regressions for the extraction of diverse structure–activity models, each having unique combination of descriptors from the generated dataset of the compounds under study. The embedded filters make the variable selection process efficient and lead to unique solution. Fear of "chance correlations" exists where large descriptor pools are used in multilinear QSAR/QSPR studies [25,26]. Furthermore, in order to discover any chance correlations associated with the models recognized in CP-MLR, each cross-validated model has been put to a randomization test [27,28] by repeated randomization of the activity to ascertain the chance correlations, if any, associated with them. For this, every model has been subjected to 100 simulation runs with scrambled activity. The scrambled activity models with regression statistics better than or equal to that of the original activity model have been counted, 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. For each model, derived by involving n data points, 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. 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.

2.4. Applicability Domain

The utility of a QSAR model is based on its accurate prediction ability for new compounds. 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 is assessed by the leverage values for each compound [29]. The Williams plot (the plot of standardized residuals versus leverage values, h) can then be used for an immediate and simple graphical detection of both the response outliers (Y outliers) and structurally influential chemicals (X outliers) in the model. In this plot, the applicability domain is established inside a squared area within ± x (s.d.) and a leverage threshold h*. The threshold h* 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. Prediction must be considered unreliable for compounds with a high leverage value (h > h*). 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.

3. Results and discussion

3.1. QSAR results

For the compounds in Table 1, a total number of 506 descriptors belonging to 0D- to 2D- classes of DRAGON have been computed. Prior to model development procedure, all those descriptors that are inter-correlated beyond 0.90 and showing a correlation of less than 0.1 with the biological endpoints (descriptor versus activity, $r < 0.1$) were excluded. This procedure has reduced the total descriptors from 506 to 107 as relevant ones to explain the biological actions of titled compounds and these were subjected to CP-MLR analysis with default "filters" set in it. The descriptors have been scaled between the intervals 0 to 1 [30] to ensure that a descriptor will not dominate simply because it has larger or smaller pre-scaled value compared to the other descriptors. In this way, the scaled descriptors would have equal potential to influence the QSAR models.

In multi-descriptor class environment, exploring for best model equation(s) along the descriptor class provides an opportunity to unravel the phenomenon under investigation. In other words, the concepts embedded in the descriptor classes relate the biological actions revealed by the compounds.

The 49 compounds were divided into training-set and test-set. Fourteen compounds (nearly 30% of total population) have been selected for test-set. The identified test-set was then used for external validation of models derived from remaining thirty five compounds in the training-set. The squared correlation coefficient between the observed and predicted values of compounds from test-set, r^2_{Test}, was calculated to explain the fraction of explained variance in the test-set which is not part of regression/model derivation. It is a measure of goodness of the derived model equation. A high r^2_{Test} value is always good. But considering the stringency of test-set procedures, often r^2_{Test} values in the range of 0.5 to 0.6 are regarded as logical models. Following the strategy to explore only predictive models, CP-MLR resulted into 08, 44 and 18 models in two, three and four descriptors, respectively. The generated models in two and three descriptors, all having $r^2_{Test}<0.5$, for the CDK8 inhibitory activity. The selected models are mentioned in Table 2.

Table 2 Highest significant models in two, three and four parameters derived for training set through CP-MLR for CDK8 inhibitory activity

Model	R	S	F	Q^2_{LOO}	r^2_{Test}	Eq.
pIC50 = 7.644 -1.799(0.359)T(N..N) +1.855(0.347)N-072	0.777	0.509	24.418	0.506	0.307	(1)
pIC50 = 5.522 +6.029(0.890)nDB -3.069(0.698)BEHp2 +2.903(0.681)nCaH	0.813	0.479	20.165	0.534	0.474	(2)
pIC50 = 4.486 +7.193(0.768)nDB -2.779(0.424)nS +4.554(0.750)nCaH -1.565(0.354)C-024	0.883	0.391	26.793	0.692	0.582	(3)

The signs of the regression coefficients have indicated the direction of influence of explanatory variables in above 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 above model Eqs., (1-3), the descriptor T(N..N) is topological class descriptor. The other participating descriptors are BEHp2 (BCUT class descriptor), nDB and nS (constitution class descriptors), N-072 and C-024 (atom centered fragments) and nCaH (functional group). The positive sign of regression coefficients of descriptors nDB (number of double bonds), nCaH (number of unsubstituted sp2 hybridized aromatic carbon atoms) and presence of RCO-N</>N-X=X type atom centered fragment (descriptor N-072) in a molecular structure suggested that a higher value of these descriptors would be beneficial to augment the CDK8 inhibitory activity. On the other hand, a lower value of descriptors T(N..N) (sum of the topological distances between N..N), nS (number of sulfur atoms), C-024 (R--CH--R) and BEHp2 (highest eigen value n.2 of Burden matrix/weighted by atomic polarizabilites) would be supportive to the CDK8 inhibition.

Considering the number of observation in the dataset, models with up to five descriptors were explored. It has resulted in 18 five-parameter models with test set $r^2 > 0.50$. These models (with 107 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 (=0.867) has been used as the new threshold of filter-3 for the next generation. These models have shared 16 descriptors among them. All these 16 descriptors along with their brief 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 physical meaning, average regression coefficient and incidence[b], in modeling the CDK8 inhibitory activities

Descriptor class	Descriptor (physical meaning), avg reg coeff (incidence)
Constitutional	nDB (number of double bonds), 4.681(17); nS (number of Sulfur atoms), -2.501(8); nR06 (number of 6-membered rings), -0.635(3)
Topological	IVDE (mean information content vertex degree equality), -1.454(1); SIC3 (structural information content of 3rd order neighborhood symmetry), 1.256(1); D/Dr06 (distance/detour ring index of order 6), -0.920(6)
BCUT	BEHp2 (atomic polarizabilities weighted lowest eigenvalue n.2 of Burden matrix), -2.547(2)
2D autocorrelations	GATS8e (atomic Sanderson electronegativity weighted Geary autocorrelation oflag 8), 0.888(3)
Functional groups	nCaH (number of unsubstituted sp2 hybridized aromatic carbon atoms), 3.265(17); nNH2Ph(number of aromatic primary amines), -0.838(1); nN-N (number of aliphatic N hydrazines), -1.016(8)
Atom centered fragments	C-024 (R--CH--R), -1.351(10); C-034 (X--CR..X), -1.150(1); H-047 (H attached to C1(sp3)/C0(sp2)), -0.660(2); N-072 (RCO-N</>N-X=X.)), 1.904(9)
Empirical	ARR(aromatic ratio), 1.223(1)

[a]The descriptors are identified from the four parameter models for activity emerged from CP-MLR protocol with filter-1 as 0.79, filter-2 as 2.0, filter-3 as 0.867 and filter-4 as $0.3 \leq q^2 \leq 1.0$ with a training set of 35 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.

Following are the selected five-descriptor models for the CDK8 inhibitory activities emerged through CP-MLR.

pIC_{50} = 4.803 +7.016(0.654)nDB −2.831(0.362)nS -0.595(0.170)nR06 +4.403(0.641)nCaH -1.698(0.304)C-024

n = 35, r = 0.919, s = 0.333, F = 31.911, Q^2_{L00} = 0.750, Q^2_{L50} = 0.757, r^2_{Test} = 0.636 .... (4)

pIC_{50} = 4.792 +2.757(0.346)nDB +1.769(0.401)nCaH -1.211(0.284)nN-N +2.079(0.327)N-072 +1.222(0.259)ARR

n = 35, r = 0.912, s = 0.348, F = 28.838, Q^2_{L00} = 0.746, Q^2_{L50} = 0.740, r^2_{Test} = 0.534 .... (5)

pIC_{50} = 4.664 +6.570(0.728)nDB −2.466(0.399)nS +4.464(0.677)nCaH -0.817(0.292)nN-N -1.447(0.322)C-024

n = 35, r = 0.909, s = 0.353, F = 27.869, Q^2_{LOO} = 0.749, Q^2_{L5O} = 0.762, r^2_{Test} = 0.606 .... (6)

pIC_{50} = 5.965 +5.372(0.693)nDB −0.730(0.184)nR06 -2.965(0.546)BEHp2 +2.692(0.512)nCaH -1.019(0.294)nN-N

n = 35, r = 0.907, s = 0.358, F = 26.969, Q^2_{LOO} = 0.714, Q^2_{L5O} = 0.731, r^2_{Test} = 0.515 .... (7)

These models have accounted for nearly 84% variance in the observed activities. In the randomization study (100 simulations per model), none of the identified models has shown any chance correlation. The values greater than 0.5 of Q^2 index is in accordance to a reasonable robust QSAR model. The pIC_{50} values of training set compounds calculated using Eqs. (4) to (7) have been included in Table 4. The models (4) to (7) are validated with an external test set of 14 compounds listed in Table 4. The predictions of the test set compounds based on external validation are found to be satisfactory as reflected in the test set r^2 (r^2_{Test}) values and the same is reported 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.

Table 4 Observed and calculated CDK8 and 7dF3 inhibition activities of naphthyridine and isoquinoline derivatives

Cpd.	CDK8 Inhibition pIC_{50}(M)[a]						7dF3 Inhibition pIC_{50}(M)[a]				
	Obs.	Calculated					Obs.	Calculated			
		Eq.(4)	Eq.(5)	Eq.(6)	Eq.(7)	PLS		Eq.(8)	Eq.(9)	Eq.(10)	Eq.(11)
1	8.44	8.38	8.16	8.38	8.13	8.43	9.70	8.55	8.32	8.47	8.27
2[b]	8.27	7.88	7.90	7.60	8.20	7.88	8.47	8.30	8.00	8.00	8.58
3	8.16	7.88	7.90	7.60	8.18	8.18	8.34	8.27	8.04	8.00	8.62
4	9.05	8.99	8.69	8.77	8.37	8.76	8.77	8.67	8.16	8.00	8.57
5	8.28	8.99	8.83	8.77	8.63	8.86	7.95	8.52	8.40	8.00	7.82
6	8.82	8.49	8.75	8.36	8.42	8.53	9.30	8.13	8.65	8.47	8.63
7	8.33	8.49	7.71	8.36	8.45	7.81	6.23	7.01	7.05	7.14	7.51
8	6.41	6.16	6.75	6.17	6.97	7.07	6.40	6.67	6.39	6.90	6.47
9[b]	8.44	8.49	8.13	8.36	8.30	8.29	8.70	8.27	8.26	8.47	7.89
10[b]	8.28	8.49	7.86	8.36	8.28	8.25	8.60	8.30	8.00	8.47	7.54
11[b]	7.59	8.49	7.86	8.36	8.30	8.29	7.90	8.27	8.10	8.47	7.92
12	8.46	8.49	7.99	8.36	8.30	8.26	8.00	8.27	7.98	8.47	7.77
13	7.89	7.90	7.86	8.36	7.60	7.65	7.56	7.27	7.67	7.38	7.54
14	7.60	7.90	7.86	8.36	7.60	7.65	7.19	7.27	7.27	7.38	6.87
15[b]	8.74	8.49	8.37	8.36	8.87	8.39	7.76	7.79	7.80	8.23	7.53
16[b]	8.62	8.49	7.94	8.36	8.76	8.34	7.52	7.93	7.76	8.23	7.76
17	9.00	8.49	8.98	8.36	8.79	9.15	8.38	8.98	9.22	9.56	8.83
18[b]	6.09	6.16	5.98	6.17	6.94	6.84	5.85	6.87	5.92	6.67	5.66
19	5.92	5.56	5.98	6.17	6.30	6.40	5.59	5.84	5.53	5.58	5.66
20	6.33	6.77	6.12	6.12	6.16	6.16	5.71	5.77	5.79	5.81	5.94
21	5.91	6.77	6.12	6.12	6.08	6.13	5.91	5.85	5.83	5.81	5.68
22	7.80	7.62	7.49	7.66	7.47	7.40	7.35	7.10	7.39	7.61	7.24
23[c]	-[c]	-[c]	-[c]	-[c]	-[c]	-[c]	-[c]	-[c]	-[c]	-[c]	-[c]

24[b]	8.35	8.38	8.87	8.38	8.18	8.56	9.22	8.41	8.68	8.47	8.92
25	8.68	8.38	8.42	8.38	8.15	8.53	8.96	8.50	8.53	8.47	8.17
26[b]	7.96	8.38	8.16	8.38	8.10	8.31	7.80	8.55	8.22	8.47	7.89
27	8.21	8.38	8.16	8.38	8.13	8.43	7.71	8.55	8.32	8.47	8.27
28[b]	8.31	8.38	8.66	8.38	8.67	8.56	7.47	8.04	8.02	8.23	7.83
29	7.47	7.51	7.79	7.68	7.26	7.58	7.80	7.36	7.62	7.61	7.58
30[b]	8.19	7.51	7.63	7.68	7.18	7.36	8.14	7.44	7.30	7.61	7.37
31	8.34	8.38	8.23	8.38	8.61	8.56	8.33	8.13	7.81	8.23	8.06
32	8.17	7.89	7.98	8.11	8.58	8.07	7.06	7.41	7.23	7.14	6.99
33	8.25	7.89	8.54	8.11	8.95	8.31	7.37	7.50	7.99	7.14	7.21
34	7.00	7.17	7.45	7.62	7.10	7.06	6.25	6.68	6.73	6.70	6.57
35	7.28	7.17	7.45	7.62	7.10	7.01	7.14	6.68	7.22	6.70	7.38
36[b]	8.72	7.76	7.45	7.62	7.77	7.59	8.14	7.68	7.54	7.79	7.38
37	7.68	7.76	7.70	7.62	7.77	7.73	7.97	7.68	7.78	7.79	7.70
38	7.92	7.76	7.34	7.62	7.74	7.58	8.09	7.71	7.58	7.79	7.82
39	7.77	7.76	7.45	7.62	7.77	7.53	8.24	7.68	7.65	7.79	7.76
40	7.43	7.76	7.45	7.62	7.77	7.54	8.01	7.71	7.70	7.79	7.76
41[b]	7.55	7.76	7.57	7.62	7.80	7.70	7.45	7.65	7.51	7.79	7.59
42	6.80	6.89	6.88	6.92	6.85	6.56	7.06	6.57	6.49	6.93	6.67
43	7.44	7.65	7.99	7.64	7.62	7.94	7.15	7.93	8.04	7.79	8.10
44	7.92	7.65	7.75	7.64	7.62	7.68	7.68	7.96	7.80	7.79	7.78
45	7.36	7.65	7.64	7.64	7.62	7.64	7.64	7.99	7.84	7.79	8.22
46	7.92	7.65	7.75	7.64	7.62	7.68	7.75	7.96	7.91	7.79	8.16
47	7.48	7.65	7.75	7.64	7.62	7.67	7.69	7.96	7.96	7.79	8.16
48	7.54	7.65	8.21	7.64	8.16	7.55	7.12	7.42	7.51	7.56	7.66
49	8.04	7.65	7.99	7.64	7.65	8.02	8.05	7.91	8.07	7.79	8.03
50[b]	7.70	7.65	7.81	7.64	8.10	7.56	7.62	7.54	7.37	7.56	7.95

[a]IC$_{50}$ on molar basis, taken from reference [14]; [b]Compound included in test set; [c]"Outlier" compound not included in data set.

.
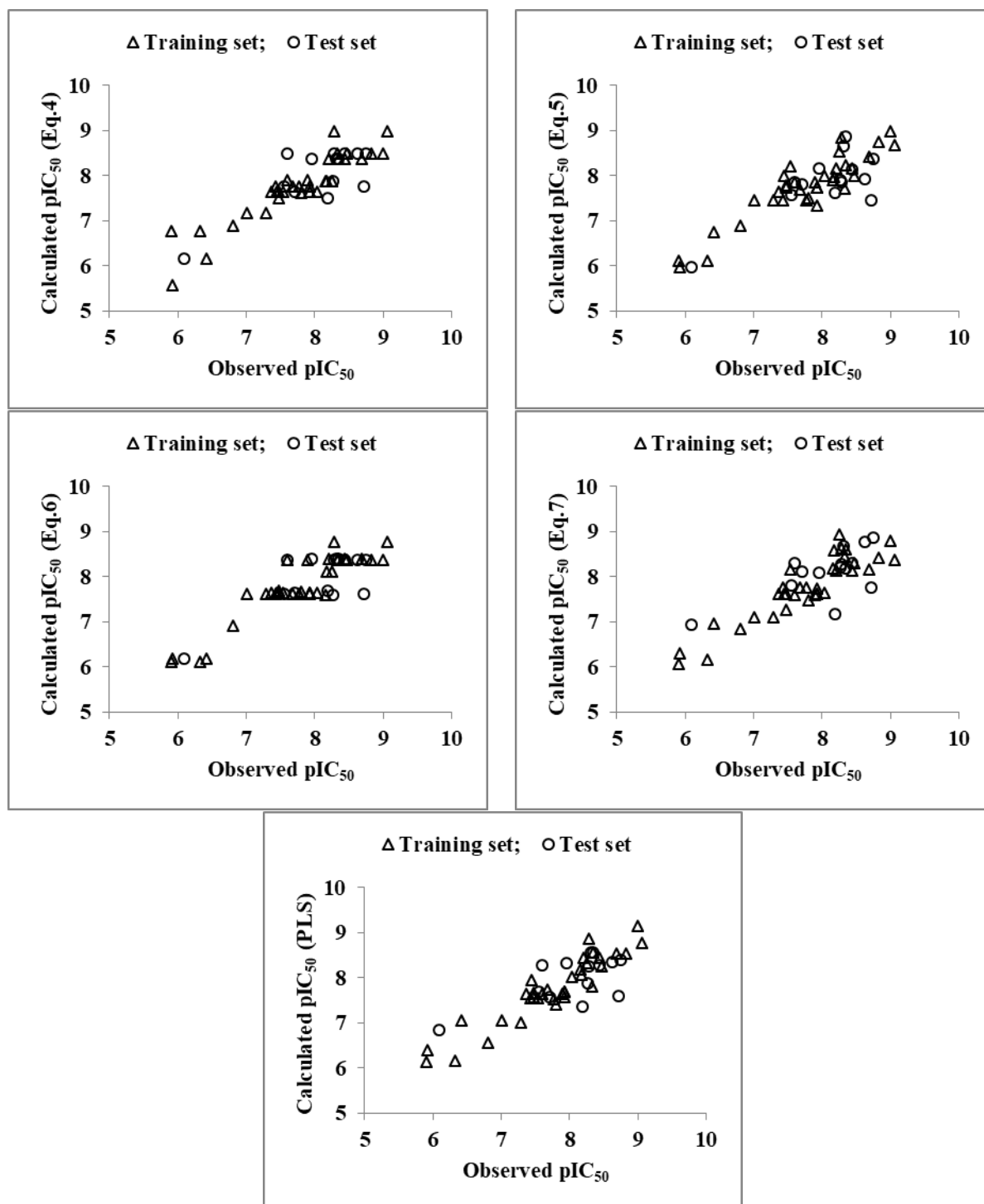
Figure 2 Plot of observed versus caculated pIC$_{50}$ values for training- and test-set compounds for CDK8 inhibition

The newly appeared descriptors in above models are nR06 (constitutional class), ARR (empirical class) and nN-N (functional group class). The descriptor ARR has shown positive correlation to the activity whereas descriptors nR06 and nN-N have correlated negatively to the activity. The signs of regression coefficients advocated that higher value of aromatic ratio (descriptor ARR) and lesser number of 6-membered rings in a molecular structure (descriptor nR06) and number of aliphatic N hydrazines (descriptor nN-N) would be helpful to augment the CDK8 inhibitory activity.

A partial least square (PLS) analysis has been carried out on these 16 CP-MLR identified descriptors, mentioned in Table 3, 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, four components are found to be the optimum for these 16 descriptors and they explained 86.86% variance in the activity. The MLR-like PLS coefficients of these 16 descriptors are given in Table 5.

Table 5 PLS and MLR-like PLS models from the 16 descriptors of five parameter CP-MLR models for CDK8 inhibitory activities

A: PLS equation								
PLS components				**PLS coefficient (s.e.)[a]**				
Component-1				0.281(0.024)				
Component-2				0.285(0.038)				
Component-3				0.101(0.029)				
Component-4				0.107(0.040)				
Constant				7.745				
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	nDB	0.341	0.136	3	9	nCaH	0.152	0.061	7
2	nS	-0.033	-0.013	12	10	nNH2Ph	-0.473	-0.189	2
3	nR06	-0.173	-0.069	6	11	nN-N	-0.211	-0.084	4
4	IVDE	0.003	0.001	16	12	C-024	0.030	0.012	14
5	SIC3	-0.027	-0.011	15	13	C-034	-0.126	-0.050	8
6	D/Dr06	-0.065	-0.026	10	14	H-047	-0.081	-0.032	9
7	BEHp2	0.032	0.013	13	15	N-072	0.532	0.213	1
8	GATS8e	0.186	0.074	5	16	ARR	-0.037	-0.015	11
Constant = 6.800								
C: PLS regression statistics				**Values**				
n				35				
r				0.932				
s				0.302				
F				49.877				
Q^2_{LOO}				0.800				
Q^2_{L5O}				0.799				
r^2_{Test}				0.513				

[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. Figure 3 shows a plot of the fraction contribution of normalized regression coefficients of these descriptors to the activity.
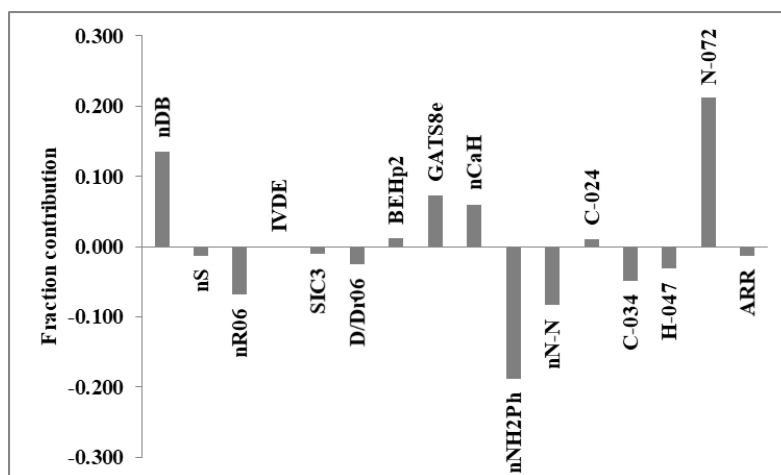
Figure 3 Plot of fraction contribution of MLR-like PLS coefficients (normalized) against 16 CP-MLR identified descriptors (Table 3) associated with CDK8 inhibitory activity of naphthyridine and isoquinoline derivatives

The PLS analysis has suggested N-072 as the most determining descriptor for modeling the activity of the compounds (descriptor S. No. 15 in Table 5; Figure 3). The other nine descriptors in decreasing order of significance are nNH2Ph, nDB, nN-N, GATS8e, nR06, nCaH, C-034, H-047 and D/Dr06. The descriptors nDB, nN-N, nR06 and nCaH are part of Eqs. (1) to (7) and convey same inference in the PLS model as well. The topological class descriptor D/Dr06, the distance/detour ring index of order 6, advocates that a lower value of it would be beneficiary to the activity. The negative influence of descriptors, nNH2Ph (number of aromatic primary amines), C-034 (X--CR..X) and H-047 (H attached to C1(sp3)/C0(sp2)) recommended absence of such functionality or fragments in a compound for improved activity. The positive regression coefficient of atomic Sanderson electronegativities weighted Geary autocorrelation of lag 8 (descriptor GATS8e), advocates that a higher positive value of it is incremental to the activity. It is also observed that PLS model from the dataset devoid of CP-MLR identified 16 descriptors (Table 3) is inferior in explaining the activity of the analogues.

CP-MLR analysis has also been carried out for another reported inhibition activity 7dF3 using same descriptor pool and test set. Following are the selected five-descriptor models for the 7dF3 inhibitory activities emerged through CP-MLR.

pIC_{50} = 6.673 -0.973(0.266)nR06 +0.993(0.429)BEHe1 -0.759(0.217)nNH2Ph -1.417(0.415)nN-N

+2.166(0.398)N-072

n = 35, r = 0.860, s = 0.526, F = 16.499, Q^2_{LOO} = 0.633, Q^2_{L50} = 0.580, r^2_{Test} = 0.587 ....(8)

pIC_{50} = 4.942 -1.125(0.396)T(N..N) +2.077(0.529)BEHm3 –1.252(0.459)nN-N +3.751(0.538)N-072

+2.009(0.494)ARR

n = 35, r = 0.859, s = 0.527, F = 16.386, Q^2_{LOO} = 0.615, Q^2_{L50} = 0.581, r^2_{Test} = 0.711 .... (9)

pIC_{50} = 6.669 +0.937(0.436)nBM -1.092(0.259)nR06 –0.673(0.206)nNH2Ph -1.797(0.460)nN-N

+2.659(0.443)N-072

n = 35, r = 0.856, s = 0.531, F = 16.025, Q^2_{LOO} = 0.615, Q^2_{L50} = 0.600, r^2_{Test} = 0.567 ....(10)

pIC_{50} = 5.023 -1.764(0.423)T(N..N) +2.151(0.596)GGI2 -0.996(0.461)nN-N +3.189(0.506)N-072

+2.203(0.558)ARR

n = 35, r = 0.850, s = 0.542, F = 15.200, Q^2_{LOO} = 0.619, Q^2_{L50} = 0.581, r^2_{Test} = 0.639 ....(11)

Except the descriptors BEHe1, BEHm3, nBM and GGI2, all the descriptors participated in models (8) to (11) are part of earlier discussed models (1) to (7) and convey same inference. It is evinced from the models mentioned above that the descriptors BEHe1, BEHm3, nBM and GGI2 contributed positively to the activity. Thus a higher value of descriptors BEHe1 (atomic Sanderson electronegativities weighted highest eigenvalue n.1 of Burden matrix), BEHm3 (atomic masses weighted highest eigenvalue n.3 of Burden matrix), nBM (number of multiple bonds) and GGI2 (Galvez topological charge index of 2nd order) will be supportive to enhance the 7dF3 inhibition activity.

These models have accounted for nearly 74% variance in the observed activities. The values greater than 0.5 of Q^2 index is in accordance to a reasonable robust QSAR model. The pIC_{50} values of training set compounds calculated using Eqs. (8) to (11) have been included in Table 4. The models (8) to (11) are validated with an external test set of 5 compounds listed in Table 4. The predictions of the test set compounds based on external validation are found to be satisfactory as reflected in the test set r^2 (r^2_{Test}) values and the same is reported 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 4.
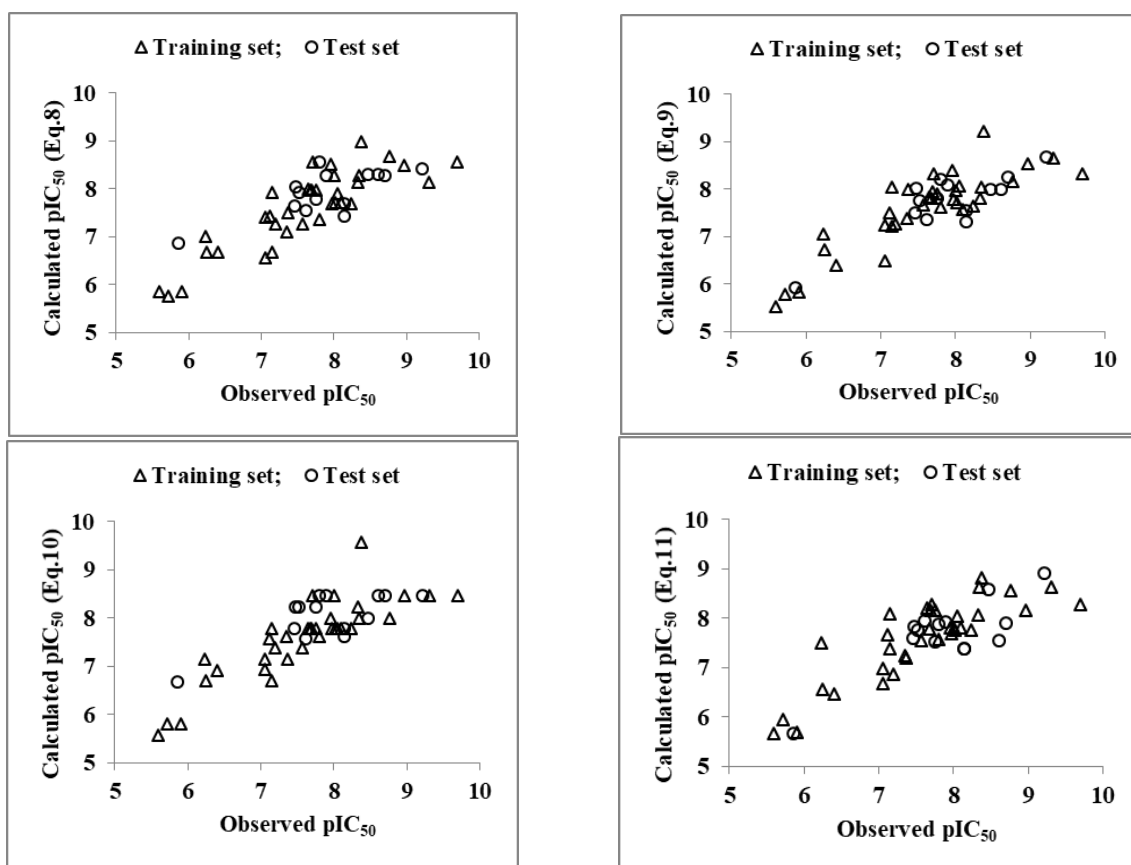
Figure 4 Plot of observed versus caculated pIC_{50} values for training- and test-set compounds for 7dF3 inhibition

3.2. Applicability domain

On analyzing the applicability domain (AD) for the CDK8 inhibitory actions in the Williams plot (Figure 5) of the model based on the whole data set (Table 5), No any compound has been identified as an obvious 'outlier' for the CDK8 inhibitory activity if the limit of normal values for the Y outliers (response outliers) was set as 2.5×(standard deviation) units. None of the compound was found to have leverage (h) values greater than the threshold leverage (h*). 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 5 Models derived for the whole data set (n = 49) in descriptors identified through CP-MLR for CDK8 inhibitory actions

Model	r	s	F	Q^2_{LOO}	Eq.
pIC50 = 5.009 +6.652(0.591)nDB -2.675(0.341)nS -0.626(0.180)nR06 +4.212(0.582)nCaH-1.688(0.289)C-024	0.889	0.369	32.519	0.719	(4a)
pIC50 = 5.149 +2.490(0.359)nDB +1.483(0.376)nCaH -1.264(0.313)nN-N +1.969(0.318)N-072 +1.148(0.241)ARR	0.872	0.395	27.415	0.690	(5a)
pIC50 = 4.905 +6.316(0.644)nDB -2.404(0.369)nS +4.296(0.607)nCaH -0.846(0.308)nN-N-1.518(0.304)C-024	0.878	0.386	29.111	0.708	(6a)
pIC50 = 5.914 +5.055(0.630)nDB -0.682(0.198)nR06 -2.569(0.474)BEHp2 +2.678(0.484)nCaH -1.099(0.321)nN-N	0.863	0.407	25.206	0.648	(7a)
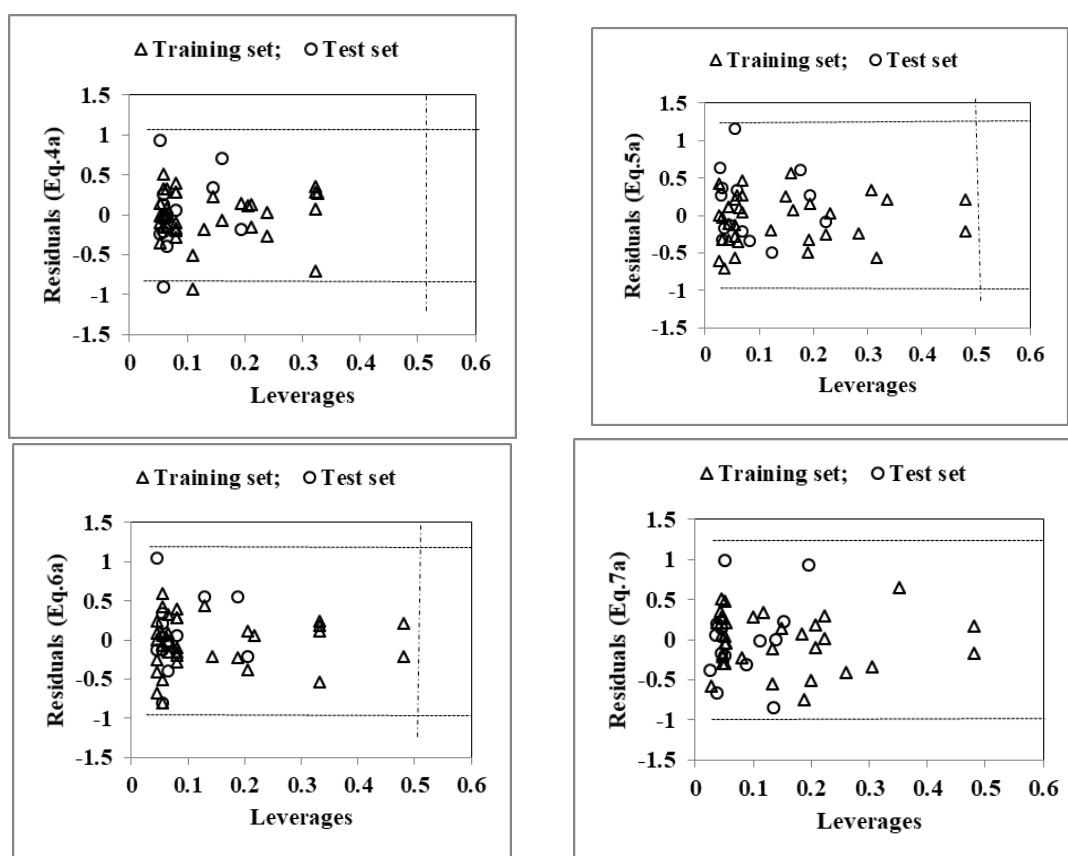
Figure 5 Williams plot for the training-set and test-set for CDK8 inhibition activity of compounds in Table 1. The horizontal dotted line refers to the residual limit (±3×standard deviation) and the vertical dotted line represents threshold leverage h* (=0.514)

4. Conclusion

The CDK8 and 7dF3 inhibition activity of naphthyridine and isoquinoline derivatives have been quantitatively analyzed in terms of Dragon descriptors. The statistically validated quantitative structure-activity relationship (QSAR) models provided rationales to explain the inhibition activities of these congeners. The descriptors identified through CP-MLR analysis for the CDK8 inhibitory activity have highlighted the role of sum of the topological distances between N..N (T(N..N)), distance/detour ring index of order 6 (D/Dr06), aromatic ratio (ARR), number of double bonds (nDB), number of 6-membered rings (nR06), number of sulphur atoms (nS), number of unsubstituted sp2 hybridized aromatic carbon atoms (nCaH), number of aliphatic N hydrazines (nN-N), number of aromatic primary amines(nNH2Ph) and

certain atom centered fragments such as R--CH--R (C-024),X--CR..X (C-034), H attached to C1(sp3)/C0(sp2) (H-047) andRCO-N</>N-X=X (N-072) to quantify the inhibitory actions. The highest eigen value n.2 of Burden matrix/weighted by atomic polarizabilites (BEHp2) and atomic Sanderson electronegativities weighted Geary autocorrelation of lag 8 (descriptor GATS8e) have also shown prevalence to model the CDK8 inhibitory activity.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. The models obtained from the descriptor pool that was chosen for the CDK8 inhibitory activity is able to explain nearly 74% variance in the observed 7dF3 activities of titled compounds.

Compliance with ethical standards

Disclosure of conflict of interest

The authors report no conflicts of interest.

References

[1]   Heptinstall AB, Adiyasa IWS, Cano C, Hardcastle IR. Recent advances in CDK inhibitors for cancer therapy. Future Med Chem. 2018; 10:1369-1388.

[2]   Philip S, Kumarasiri M, Teo T, Yu M, Wang S. Cyclin-dependent kinase 8: a new hope in targeted cancer therapy? J Med Chem. 2018; 61:5073-5092.

[3]   Morris EJ, Ji JY, Yang F, Di Stefano L, Herr A, Moon NS, Kwon EJ, Haigis KM, Naar AM, Dyson NJ. E2F1 represses beta-catenin transcription and is antagonized by both pRB and CDK8. Nature. 2008; 455:552-556.

[4]   Firestein R, Bass AJ, Kim SY, Dunn IF, Silver SJ, Guney I, Freed E, Ligon AH, Vena N, Ogino S, Chheda MG, Tamayo P, Finn S, Shrestha Y, Boehm JS, Jain S, Bojarski E, Mermel C, Barretina J, Chan JA, Baselga J, Tabernero J, Root DE, Fuchs CS, Loda M, Shivdasani RA, Meyerson M, Hahn WC. CDK8 is a colorectal cancer oncogene that regulates betacatenin activity. Nature. 2008; 455:547-551.

[5]   Firestein R, Shima K, Nosho K, Irahara N, Baba Y, Bojarski E, Giovannucci EL, Hahn WC, Fuchs CS, Ogino S. CDK8 expression in 470 colorectal cancers in relation to beta-catenin activation, other molecular alterations and patient survival. Int J Cancer. 2010; 126:2863-2873.

[6]   Lim SC, Kim MY, Han SI. Roles of cyclin-dependent kinase 8 and b-catenin in the oncogenesis and progression of gastric adenocarcinoma. Int J Oncol. 2011;   38:1375-1383.

[7]   Schneider EV, Böttcher J, Blaesse M, Neumann L, Huber R, Maskos K. The Structure of CDK8/CycC   Implicates Specificity in the CDK/Cyclin Family and Reveals Interaction with a Deep Pocket Binder. J Mol Biol. 2011; 412:251-266.

[8]   Porter DC, Farmaki E, Altilia S, Schools GP, West DK, Chen M, Chang BD, PuzyrevAT, Lim C, Rokow-Kittell R, et al. Cyclin-Dependent Kinase 8 Mediates Chemotherapy Induced Tumor-Promoting Paracrine Activities. Proc Natl Acad Sci USA. 2012; 109:13799-13804.

[9]   Pelish HE, Liau BB, Nitulescu II, Tangpeerachaikul A, Poss ZC, Da Silva DH, CarusoBT, Arefolov A, Fadeyi O, Christie AL, Du K, Banka D, Schneider EV, Jestel A, Zou G, Si C, Ebmeier CC, Bronson RT, Krivtsov AV, Myers AG, Kohl NE, Kung AL, Armstrong SA, Lemieux ME, Taatjes DJ, Shair MD. Mediator kinase inhibition further activates superenhancer-associated genes in AML. Nature. 2015; 526:273-276.

[10]   Koehler MFT, Bergeron P, Blackwood EM, Bowman K, Clark KR, Firestein R, Kiefer   JR, Maskos K, McCleland ML, Orren L, Salphati L, Schmidt S, Schneider EV, Wu J, Beresini MH. Development of a Potent, Specific CDK8 Kinase Inhibitor Which Phenocopies CDK8/19 Knockout Cells. ACS Med Chem Lett. 2016; 7:223.

[11]   Mallinger A, Crumpler S, Pichowicz M, Waalboer D, Stubbs M, Adeniji-Popoola O, Wood B, Smith E, Thai C, Henley AT, Georgi K, Court W, Hobbs S, Box G, Ortiz-Ruiz MJ, Valenti M, De Haven Brandon A, TePoele R, Leuthner B, Workman P, Aherne W, Poeschke O, Dale T, Wienke D, Esdar C, Rohdich F, Raynaud F, Clarke PA, Eccles SA, Stieber

F, Schiemann K, Blagg J. Discovery of Potent, Orally Bioavailable, Small- Molecule Inhibitors of WNT Signaling from a Cell-Based Pathway Screen. J Med Chem. 2015; 58:1717-1735.

[12] Dale T, Clarke PA, Esdar C, Waalboer D, Adeniji-Popoola O, Ortiz-Ruiz M, Mallinger A, Samant RS, Czodrowski P, Musil D, Schwarz D, Schneider K, Stubbs M, Ewan K, Fraser E, TePoele R, Court W, Box G, Valenti M, de Haven Brandon A, Gowan S, Rohdich F, Raynaud F, Schneider R, Poeschke O, Blaukat A, Workman P, Schiemann K, Eccles SA, Wienke D, Blagg J. Identification of a potent and selective chemical probe for exploring the role of Mediator complex-associated protein kinases CDK8 and CDK19 in human disease. Nat Chem Biol. 2015; 11:973-980.

[13] Mallinger A, Schiemann K, Rink C, Stieber F, Calderini M, Crumpler S, Stubbs M, Adeniji-Popoola O, Poeschke O, Busch M, Czodrowski P, Musil D, Schwarz D, Ortiz- Ruiz M-J, Schneider R, Thai C, Valenti M, De Haven Brandon A, Burke R, Workman P, Dale T, Wienke D, Clarke PA, Esdar C, Raynaud FI, Eccles SA, Rohdich F, Blagg J. J Med Chem. 2016; 59:1078-1101.

[14] Mallinger A, Schiemann K, Rink C, Sejberg J, Honey M A, Czodrowski P, Stubbs M, Poeschke O, Busch M, Schneider R, Schwarz D, Musil D, Burke R, Urbahns K, Workman P, Wienke D, Clarke PA, Raynaud FI, Eccles SA, Esdar C, Rohdich F, Blagg J. 2,8-Disubstituted-1,6-Naphthyridines and 4,6-Disubstituted-Isoquinolines with Potent, Selective Affinity for CDK8/19. ACS Med Chem Lett. 2016; 7:573-578.

[15] Chemdraw ultra 6.0 and Chem3D ultra, Cambridge SoftCorporation, Cambridge, USA. http://www.cambridgesoft.com

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

[17] Prabhakar YS. A combinatorial approach to the variable selection in multiple linear regression: analysis of Selwood et al. Data Set-a case study. QSAR Comb Sci. 2003; 22:583-595.

[18] Sharma S, Prabhakar YS, Singh P, Sharma BK. QSAR study about ATP-sensitive potassium channel activation of cromakalim analogues using CP-MLR approach, Eur J Med Chem. 2008; 43:2354-2360.

[19] Sharma S, Sharma BK, Sharma SK, Singh P, Prabhakar YS. 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, Eur J Med Chem. 2009; 44:1377-1382.

[20] Sharma BK, Pilania P, Singh P, Prabhakar YS. 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 QSAR Environ Res. 2010; 21: 169-185.

[21] Sharma BK, Singh P, Sarbhai K, Prabhakar YS. A quantitative structure-activity relationship study on serotonin 5-HT6 receptor ligands: Indolyl and piperidinyl sulphonamides. SAR QSAR Environ Res. 2010; 21:369-388.

[22] Wold S. Cross-validatory estimation of the number of components in factor and principal components models. Technometrics. 1978; 20:397-405.

[23] Kettaneh N, Berglund A, Wold S. PCA and PLS with very large data sets. Comput Stat Data Anal. 2005; 48:69-85.

[24] Stahle L, Wold S. Multivariate data analysis and experimental design. In: Ellis GP, West WB. Eds., Biomedical research. Progress in medicinal chemistry. Elsevier Science Publishers, BV, Amsterdam.1988; 25:291-338.

[25] Topliss JG, Edwards RP. Chance factors in studies of quantitative structure–activity relationships. J Med Chem. 1979; 22:1238–1244.

[26] Katritzky AR, Dobchev DA, Slavov S, Karelson M. Legitimate utilization of large descriptor pools for QSAR/QSPR models. J Chem Inf Model. 2008; 48:2207-2213.

[27] So S-S, Karplus M. Three-dimensional quantitative structure–activity relationship from molecular similarity matrices and genetic neural networks. 1. Method and validation. J Med Chem. 1997; 40:4347-4359.

[28] Prabhakar YS, Solomon VR, Rawal RK, Gupta MK, Katti SB. CP-MLR/PLS directed structure–activity modeling of the HIV-1 RT inhibitory activity of 2,3-diaryl-1,3-thiazolidin-4-ones. QSAR Comb Sci. 2004; 23:234–244.

[29] Gramatica, P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007; 26:694-701.

[30] Golbraikh A, Tropsha A. Be aware of q^2! J Mol Graph Model. 2002; 20:269-76.