RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

SCAFFOLD AND CELL LINE BASED APPROACHES FOR QSAR STUDIES ON ANTICANCER AGENTS

Shruti Satbhaiya*, O.P. Chourasia

Heterocyclic Research Laboratory, Department of Chemistry

Dr. Hari Singh Gour Vishwavidyalaya, Sagar M.P. 470003 (India)

*** TO WHOM ALL CORRESPONDENCES MAY BE ADDRESSED**

Prof. O.P. CHOURASIA ***(opcchem@gmail.com)***

Heterocyclic Research Laboratory

Department of Chemistry

Dr. Hari Singh Gour Vishwavidyalaya

Sagar-470003 M.P. (India)

Phone: +91-9009509089, +91-9098999717

E-mail:opcchem@gmail.com; satbhaiyashruti@gmail.com.

Keywords: Analogue Based Approach, Anticancer Agent, QSAR and Descriptor etc.

Highlights:

- Importance of 2D QSAR in drug discovery.
- Lower number of descriptors containing models shows best statistical parameters.
- Number of involved scaffolds in models affecting the statistical values.

Abstract:

Based on the linear heuristic method, Quantitative Structure Activity Relationship was developed for the prediction of available *in vitro* anticancer activity, based on the linear heuristic method. Each type of compound was represented by several calculated structural descriptors. Most of the computational studies are carried out targeting insufficient number of cell lines. The predictive models were built for 482 compouds with experimental data against 30 different cancer cell lines. Strong statistical analysis shows a high correlation, cross validation coefficirnt values and provides a range of QSAR equation. Quantum chemical descriptors were found in 42 out of 46 models, electrostatic in 16, topological in 12, geometrical in 7, thermodynamical in 5 and constitutional in 7. It is intresing to note that in most cases three descriptor based models are relevant. Pancreatic cancer cell lines show best statistical values (average R^2= 0.87) followed by leukaemia cancer (average R^2= 0.86).

1. Introduction:

Cancer is a multifactor disease of striking significance in the world today. It has become second leading cause of death among human population [1-2] after cardiovascular diseases. It ranks high among human diseases and has become the second reason of mortality in the world. Therefore the development of potent and precise anticancer agents is urgently in need [3] and still a major challenge to medicinal chemistry research. Researchers have given attention towards the discovery of novel anticancer agents due to lack of extensive range of anticancer drugs to take advantage on new discoveries concerning tumour genesis, tied with the exclusive growth pattern of various repertoires of cancer [4] but due to acquisition by cancer cells

of multiple-drug resistance, current anticancer chemotherapy still suffers. A vast increase in the number of feasible molecular targets, the focus has shifted from target identification to target validation [5].

The main sources of lead compounds for drug development are natural products because of their intrinsic biorelevance presence of small hetero-aromatic compounds they have shown unexpected biological properties and became basic for the whole number of innovative medicinal agents [6]. The collection of these compounds is dramatically higher than those resulting from high throughput screens of combinatorial libraries [7-9]. Preparation of libraries based on natural products requires sophisticated and laborious synthetic sequences. In addition, therapeutic development of promising leads resulting from these libraries is significantly impeded by the problem of large-scale compound supply. Because of the improved interest in natural products by the failure of alternative methods to provide many therapeutic lead compounds and by the pharmaceutical industry these challenges are becoming increasingly more pertinent. [10]

A pharmaceutical industry has to make sure the safety, quality, and efficacy of a marketed drug by subjecting the drug to a range of analysis [11]. To acquire the complete object of drug discovery it takes long time approximately 12 years [12] and was projected to high cost for marketed drug [13]. Due to this expensive and lengthy process may cause failure of drug development. Thus, it will be useful to predict these failures prior to the clinical stage in order to reduce drug development costs [14]. In the drug development stage to filter out potential failures various methods such as *in vitro, in vivo* or *in silico* methods are being used. Quantitative structure–activity relationship (QSAR) model is an example of an *in silico* method, which can be used to understand drug action, designing of new compounds, and screen chemical libraries [15-18]. Combinatorial approaches is a influential tool in selection to speed up drug discovery and with different mechanism of action this method is being adopted to cure the cancer [19-20]. QSAR has become crucial into the molecular interpretation of biological properties [21-26]. This technique is the most important tool used in analogue-based drug design and has been broadly used for calculation of

assorted properties like carcinogenicity [27], ADME [28], stability [29], toxicity [30-31], retention time [32] and other physicochemical properties apart from the biological activity [33-36]. QSAR method make possible the theoretical prediction of structures with desired property values by combining the QSAR method with pattern recognition techniques. In lead optimization, development of QSAR using various physicochemical descriptors has been a vital task. [37]. The use of such multiple QSAR to derive mechanistic approach can be illustrated by a comparison of the experimental data available on the anticancer agents. Computational methods aid also the rapid generation of new hypotheses moreover the design and interpretation of hypothesis-driven experiments in the field of cancer research.

A number of quantum chemical descriptors (such as molecular orbital, charge and dipole moment, etc.), electrostatic descriptors (such as charge based descriptors etc.), geometrical descriptors (such as moment of inertia etc.) and thermo dynamical descriptors (such as entropy and vibrational frequency etc.) have been effectively applied to set up QSAR models for predicting activities of compounds [38-40]. There are a large number of cell lines available for a cancer type, on which *in vitro* biological activity can be executed, but the results of this prediction differ based on the cell line used for assay. As a result it becomes complicated for computational chemists to select experimental data from a pool of existing biological activity for a single scaffold type. *In vitro* experimental data for anti-cancer activity is available against many different cell lines. In the literature, QSAR studies are carried out mainly for any one particular cell line, which may not be a good approach. The study considering all the available experimental data for many different cell lines to build predictive models, will suggest medicinal chemists to more reliably design new and potent compounds. Analyses of the obtained descriptors for models against all the cell lines, may suggest the significance of a particular type of descriptor in modeling anti-cancer activity against a cancer type.

2. Computational methods:

2.1. Data Set for Analysis:

Reported *in vitro* assay for 16 different scaffolds against 30 various cell lines for total 482 compounds were considered for the present investigation (Table S1-S16 in supplementary file). The inhibitor activities (IC_{50}) against different cell lines were converted pIC_{50} according to the formula $pIC_{50} = - \log (IC_{50})$. The parent structure of all the scaffolds with a number of compounds and name of cell lines are reported in figure-1. Table-1 represents the name of scaffolds considered, different cell lines and number of molecules corresponding to cell lines [20, 41-54].

2.2. Optimization:

A total of 482 compounds are collected along with their anti-cancer activity against 30 cancer cell lines which belong to 16 different chemical scaffolds (Figure-1). All the structures were initially optimized and their vibrational frequencies calculated using semi-empirical AM1 procedure using AMPAC 5.0 and obtained a Gaussian output files for each structure, which act as a input file for CODESSA program for calculating descriptors as well regression analysis.

2.3. Calculating 2D Descriptors and regression analysis:

CODESSA (COmprehensive DEscriptors for Structural and Statistical Analysis) version 2.0was used for calculating 2D descriptors as well as for regression analysis [55]. Figure-2 provides a schematic illustration of work flow accepted for current study to developing and validating various QSAR models. Initially approximately 540 default descriptors were calculated and these descriptors were further classified into following groups viz. constitutional, topological, geometrical, electrostatic, quantum-chemical and thermo dynamical descriptors [37].

Two different schemes were opted to develop statistically significant QSAR models. In the first scheme, 16 QSAR models were developed for the 16 scaffolds used in this investigation (i.e. scaffold-based QSAR models), whereas in the second scheme 30 different QSAR models were

developed based on the availability of IC50 values against 30 cancer cell lines by combining all the scaffolds (i.e. cell lines-based QSAR models). For all models inter-correlation of the descriptors was also tested. Then, models containing highly inter-correlated descriptors were replaced and refined so that the descriptors, which employed in a given models are practically orthogonal to each other.

Large number of descriptors will create confusions and reduce the predictive ability and statistical robustness of the model. So we scrupulously developed 3, 4 and 5 descriptor-based models for all sets of compounds to find out the minimum number of descriptor defining activity with the help of heuristic method which belong to multilinear regression method. This method is better than other methods due to its high speed. This method usually produce correlation 2-5 times faster than other methods with comparable quality and it has no restriction on the size of data set. On the comparison with four and five descriptors-based models, three descriptor-based models were found satisfactory for all sets of compounds. For assessing of statistical quality of the models various parameters like R^2, R_{cv}^2, AE, s^2, F and t-test are essential, which are obtained from the correlation of approximately 540 descriptors (constitutional, geometrical, topological, electrostatic, thermo dynamical and quantum chemical etc.) in different combination [56]. Where R^2 value is relative measure of quality of fit, F represents F-ratio between the variance of calculated and experimental activity and t-test reflects significance of the parameter within the model. The effect of the number of descriptors on the correlation coefficient was examined on the set of molecules using heuristic method at 1-10 descriptors.

3. Results & Discussion:

By using IC50 values as dependent variables and deliberated properties as independent variables, regression was executed for QSAR analysis of various developed models. It would be suitable to obtain insight into the physical meaning of the correlation obtained as an output of the regression analysis. To improve the anticancer activity of molecules, magnitude of a descriptor could be used as guidelines.

Among the developed models sixteen and thirty models were selected on the basis of several statistical and other parameters such as R^2, R_{cv}^2, S2, AE values, Fischer's value (F test) and t-test. The relation between number of descriptors and correlation values for all models were experienced by correlating 1-10 descriptors individually and presented in Figure 4(a) and 4(b) for cell lines based models and scaffold based models respectively. Among all models, three descriptors models were acceptable for getting a best correlation because higher than six descriptors models may give high correlation values, which may be phony and may not be constructive for the further prediction of biological activities.

All the models were separated into training set and test set. Developed models, which were construct using training set compounds, were used to determine the activity of test set compounds. Lower average residual values obtained from both the training and test set is indicate that which models have high potential to establish the correlation between the structure and activity.

Most of the scaffold based QSAR models along with regression equation, cancer types and the name of the cell lines are given in Table 2(a). We obtained superior statistically quality for most of the scaffolds based QSAR models with higher correlation coefficient values than cell lines based models. There is an important reason for high correlation coefficient of these models is contribution of lower number of compounds. The range of activity of compounds in three (S2, S10 and S12) models is poor. On comparison, models containing broad activity range compounds show high correlation coefficient while narrow activity range shows lower correlation coefficient values. Besides these models all the scaffold based models with high correlation coefficients values seen rational and can be used for further prediction.

All QSAR models were cross validated by these high R_{cv}^2 values, obtained by leave one out method for validation of model R_{cv}^2 should be greater than 0.5 [57]. Regression summary for cell lines based QSAR models (M4, M6, M9, M11, M17, M16, M18, M19, M23 and M24) show high statistically quality (avg. R^2=0.93, R_{cv}^2= 0.89) and appear precious for the existing class of compounds. The statistical quality of few other cell lines based models (M1, M7, M12, M14, M25, M26, M28, M29, M5, M3

and M30) is also showing moderate statistically quality (avg. R^2= 0.71 and Rcv^2= 0.69), and these models can also used for the prediction. However some models (M27, M22, M21, M20, M15, M13, M10, M8 and M2) cannot be used for the further prediction because of the narrow statistical quality of these models (avg. R^2= 0.58, R_{cv}^2= 0.45). The reason for irrerelevant results obtained from these models are probably due to the contribution of higher number of compounds and 3 to 5 different scaffolds in these models. The increase in the number of descriptors in narrow range activity models is not much effective to improve the statistical quality of models and shows that the currently used descriptors are not agreeable for developing the structure activity relationship for these models, and one needs to try or develop additional descriptors. However the involvements of single scaffold in these models provide a good statistical quality. All details for cell lines based models are illustrated in Table. 2 (b)

The calculated and experimental biological activity with residuals and descriptor values for all models are given in Additional file A (Table S18 to S63). Figure. 3(a) & 3(b) are showing the plots between experimental and calculated activity values for 15 cell lines and 8 scaffold- based QSAR models. Enduring plots are given in Additional file A (Figure.S1(a) & S1(b)). According to plots, the average residual for test and training set compounds clearly represent that compounds of test set are closer to the line compared with the compounds of training set.

Total 109 descriptors were used in different combinations for development of all QSAR models. Figure.5 illustrates the percentage of all types of descriptors involved in models. and this figure shows the importance of quantum chemical descriptors (Approx. 63%) followed by electrostatic (13.6%), topological (9.6%), geometrical (5.5%), thermo dynamical and constitutional (both in 4.5%). The inter-correlation of the descriptors for all the developed models has been done and inter-correlation of all descriptors explains that descriptors are rationally orthogonal. In quantum chemical descriptors, charged based descriptors such as Max n-n repulsion for a C-N bond, Max e-n attraction for a C-C bond and ESP-RPCG Relative positive charge etc. present in approximately 40 (approx.37%) models. This was followed

by valency-based descriptors and bond order based descriptors presents in approx.23 % and 3% respectively. This represents the importance of charged-based, valency-based and bond-order based descriptors.

Cell lines of different cancer types, considered in the current study presented in Additional file A (Table. S66). Among them 7 cancer types have experimental data for more than one cell line. Thus, comparative statistical significance of various types of cancer has been done and presented in Additional file A (Table S66). Pancreatic (R^2= 0.87, R_{cv}^2= 0.73), leukemia (R^2= 0.86, R_{cv}^2= 0.80), renal (R^2= 0.85, R_{cv}^2= 0.76), cervical (R^2= 0.77, R_{cv}^2= 0.71), brain (R^2= 0.77, R_{cv}^2= 0.60), lung (R^2= 0.76, R_{cv}^2= 0.67) and CNS (R^2= 0.75, R_{cv}^2= 0.63) types of cancer have better statistical values compared with other types of cancer such as colon, breast, ovarian, skin, prostate, neuronal, melanoma and heptocellular etc (Avg. R^2=0.60 , Avg. R_{cv}^2=0.51).

4. Conclusion:

Our motto in this investigation was to biologically evaluate a series for anticancer agents by modifying methodically the molecule, in order to explore the SAR of these scaffolds. A total of 46 QSAR models, 16 and 30 for different scaffolds and different cell lines respectively, were built to assess the predictive power of QSAR models where the number of descriptors is improved from 1 to 10.This study reveals that three descriptors - based models are found satisfactory for further prediction and also show that quantum chemical descriptors are the most important type of descriptors followed by electrostatic, topological, geometrical, thermo dynamical and constitutional descriptors. An analogue-based designing approach is important for modeling anti-cancer compounds. Developed models for all experimentally tested compounds contain higher correlation coefficient (R^2), higher cross-validation coefficient (R_{cv}^2) values and lower average residuals (AE) values. Cell lines in pancreatic cancer average R^2= 0.87 followed by cell lines in leukaemia cancer with average R^2= 0.86 provided the best statistical values. Although the derived

equation is of restricted validity due to the limited size of the training set, this result may prove fruitful in predicting new anticancer agents with desired activity.

Acknowledgement:

SS and OPC thank to UGC, New Delhi for financial assistant. The support from Department of Chemistry, Dr. H. S. Gour Central University Sagar M.P. India is also acknowledged.

Conflict Of Intrest

The authors have declared no conflict of interest.

Tables & Figures

Table.1: Details of scaffolds considered in the study and the cell lines against which their anticancer activity was reported along with the number of molecules in each cell lines.

S.No.	Scaffold Name	Cell Lines	Cancer Type	No. of Compound	Ref.
S1	Acridine	P388	Leukemia	41	[41]
		LLc	Lung	41	
		JLc	Leukemia	41	
S2	Cantharidine	HT-29	Colon	35	[42]
		SW480	Colon	35	
		MCF-7	Breast	35	
		A2780	Ovarian	35	
		H460	Lung	35	
		A431	Skin	35	
		DU145	Prostate	35	
		BE2-C	Neuronal	35	
		SJ-G2	Brain	35	
S3	Chalcone	ACHN	Renal	19	[43]
		Pancc1	Pancreatic	19	
		Calu1	Lung	19	
		H460	Lung	19	
		HCT116	Colon	19	
S4	Tetrahropyrimidine	MCF-7	Breast	23	[44]
S5	Isatin	HCT116	Colon	32	[45]
		MCF-7	Breast	32	
S6	Isoflavne	HCT116	Colon	23	[46]
S7	Nitroalkene	HeLa	Cervical	22	[47]
S8	Phenazine	H69	Lung	18	[48]
S9	Podophyllotoxin	HeLa	Cervical	30	[49]
		MCF7	Breast	30	
S10	Pyrazole	HeLa	Cervical	17	[49]

		MCF-7	Breast	17	
S11	Pyrazoline	MCF-7	Breast	20	[50]
		B16-F10	Melanoma	20	
S12	Pyrimidine	BEL-7402	Heptocellular	37	[20]
S13	Quinazoline	MCF-7	Breast	36	[51]
		U251	CNS	36	
		SW480	Colon	36	
		H522	Lung	36	
		M14	Melanoma	36	
		SKOV3	Ovarian	36	
		DU145	Prostate	36	
		A498	Renal	36	
S14	Quinoxaline	MCF-7	Breast	22	[52]
		H460	Lung	22	
		SF-268	CNS	22	
S15	Semicarbazide	L120	Leukemia	30	[53]
S16	Stillbene	A549	Lung	69	[54]
		MCF-7	Breast	69	
		HT-29	Colon	69	
		SKMEL-5	Melanoma	69	
		MLM	Melanoma	69	

Table 2(a): Cell line with type of cancer in parenthesis, scaffolds involved, regression summary (regression equation, correlation coefficient R^2, cross validation coefficient R_{cv}^2 and average residual) and number of compounds (training set TR, and test set TS) in various scaffolds based QSAR models.

No.	Cell lines (Type)	Regression equation	R^2	R_{CV}^2	AE	F	S^2	# Comp	
								TR	TE
S1	P388 (Leukemia)	=-6.2155*VE/T + 2.3164* WPSA3Q + 3.3250*LNMVF + 1.5252	0.75	0.67	0.35	26.74	0.145	31	9
S2	HT29 (Colon)	=1.8444* RNB – 3.3083* MaenAC + .13180* PMIA + 6.1097	0.69	0.55	0.12	15.36	0.033	27	8
S3	ACHN (Renal)	=7.4622 * FPSA3z – 7.8674*WNSA2z – 1.0224* BI + 1.8722	0.98	0.95	0.05	105.92	0.001	15	4
S4	MCF7 (Breast)	= -2.0431 * PPSA3z + 2.7466 * ZXS/ZXR- 2.0527 * RNCGQ + 3.5395	0.89	0.73	0.09	29.10	0.009	17	5
S5	HCT116 (Colon)	=4.1330* RNCl – 2.1896 *RNCSz – 1.2796 * FNSA2Q -2.7626	0.77	0.66	0.21	21.31	0.028	23	8
S6	HCT116 (Colon)	=4.0785* EMiNACC -4.1004 * PNSA2Q- 1.4298 EHBCAQ + 2.0450	0.88	0.82	0.14	31.98	0.018	18	5
S7	HeLa (Cervical)	=-9.3376*PMIB – 2.0744* EHDSAQ + 1.7527 * EMaNACC + 3.8717	0.85	0.75	0.14	19.85	0.036	15	4
S8	H69 (Lung)	= 4.7221* MaPCHz – 2.5135 * TE/#A-T + 2.6179 * MiERIC	0.96	0.93	0.12	76.93	0.019	14	4
S9	HeLa (cervical)	=2.1519 *MiERIN + 3.4050* MaREHN - 1.0293 * ABOC – 3.4442	0.93	0.90	0.12	84.32	0.074	22	6

S10	HeLa (Cervical)	=9.0910* MaenACH -6.2862 * HNMVF + 1.2038 * MaPPBO -1.5070	0.96	0.90	0.15	50.38	0.051	14	3
S11	B16-F10 (Melanoma)	=-2.7430 *1XGP + 5.4109 * DPSA2z- 6.9408* EHDSAQ + 3.9304	0.93	0.88	0.11	41.90	0.012	15	5
S12	BEL-7402 (Melanoma)	= 2.98993 * HLEG + 1.0598* MaeeRN + 2.9834 * KSI3- 4.0099	0.70	0.60	0.35	20.86	0.106	32	11
S13	M14 (Melanoma)	=4.4562* MiERIN – 3.6455 * MiAOEP + 1.0863 * MiBON (0.1) + 1.3577	0.65	0.56	0.31	15.05	0.215	28	8
S14	SF-268 (CNS)	=-1.3737 * MaTICC + 1.2771* EPNSA3Q-1.2524*MiTICN + 4.8510	0.74	0.59	0.46	11.22	0.240	17	5
S15	L120 (Leukemia)	= 1.712 *RNN – 4.0400* MiERIC + 9.1240* MIA – 4.5014	0.92	0.89	0.24	71.94	0.055	22	8
S16	A549 (Lung)	=-1.2148* MaenACC + 5.4537 * FNSA1Q – 6.7738* AIC1	0.48	0.38	0.24	12.46	0.087	49	17

***R2 is the square of the correlation coefficient and represents the statistical significance of the model. Rcv2 is the cross-validated R2, a measure of the quality of the QSAR model. AE is the average of absolute difference between experimental and calculated IC50 values. F is the Fischer statistics, the ratio between explained and unexplained variance for a given number of degrees of freedom, thereby indicating a factual correlation or the significance level for QSAR models. S2 is the standard deviation. TR is number of molecules in training set and TE is test set molecules.**

Table 2(b): Cell line with type of cancer in parenthesis, scaffolds involved, regression summary (regression equation, correlation coefficient R^2, cross validation coefficient R_{cv}^2, average residual AE) and number of compounds (training set TR, test set TS) in various cell lines based QSAR models.

No	Cell line (Type)	Scf.	Regression equation	R^2	R_{cv}^2	AE	F	S^2	# of comp.	
									TR	TS
M1	A498 (renal)	**S13**	=-3.3738* MannRCN+-1.2453* ASIC1 -1.0807* RNCSz+4.6355	0.71	0.56	0.62	14.87	0.239	31	5
M2	A549 (Lung)	**S16**	=-1.3066* MaenACC +1.1275* PNSA1Q -7.9011* PNSA2z +3.3541	0.56	0.49	0.27	18.41	0.094	56	10
M3	A2780 (Ovarian)	**S2**	=5.3016* MiNRIO +6.1285* HACA1Q-1.222* RPCSz -3.6341	0.68	0.54	0.22	13.89	0.043	30	5
M4	ACHN (Renal)	**S3**	=7.4622*FPSA3z -7.8674* WNSA2z-1.0224* BI+1.8722	0.96	0.95	0.05	105.93	0.001	16	3
M5	A431 (Skin)	**S2**	=-7.2276* YZS +1.0111*RI0* +7.4112.* MiERIC +2.5462	0.69	0.59	0.14	13.43	0.036	30	5
M6	B16-F10 (Melanoma)	**S11**	=-1.0251* WPSA1Q -1.7686* MiTICS +1.6039* MiNACN +2.7567	0.94	0.89	0.11	62.56	0.015	17	3
M7	BE2-C (Neuronal)	**S2**	=-5.0255* XYS/XYR +1.4971* MiERIC -6.7892* RNO +5.2368	0.72	0.63	0.16	16.67	0.030	29	7
M8	BEL-7402 (Heptocellular)	**S12**	=2.6386* HLEG -6.0993* WNSA2Q-4.8693* MiERIN-2.2423	0.58	0.44	0.31	12.34	0.177	35	10
M9	Calu1 (Lung)	**S3**	=2.2933* A1ERIC -2.3063*SIC1 -3.0798*YZS +5.2315	0.93	0.80	0.12	41.25	0.023	16	3
M10	DU145	**S3,**	=7.0109* MiTICN-7.8249* PNSA1z	0.43	0.32	0.45	11.60	0.307	57	15

	(Prostate)	**S13**	-2.7692* CHaSz -6.9257							
M11	H69 (Lung)	**S8**	=-1.1403* FNSA2z -5.8440* ERPCGQ -5.2628* MienACH +3.4678	0.93	0.81	0.13	44.62	0.033	15	3
M12	H522 (Lung)	**S1**3	=1.4711*MiPCCz* +2.3775* MiBOC(0.1) +8.7939* THCMD-2.1601	0.73	0.63	0.24	18.26	0.201	28	8
M13	HCT116 (Colon)	**S3, S5, S6**	=2.7531* RE/T+3.3081* LNMVF+64.6887* ACIC1 -1.0055	0.59	0.49	0.29	24.10	0.119	60	14
M14	Hela (Cervical)	**S7, S9, S10**	=-6.4020*MaBON - 9.5518*MienACN +7.3732NN +3.7878	0.76	0.71	0.40	43.55	0.377	56	11
M15	HT29 (Colon)	**S2, S16**	=-6.4490*FNSA2Q+ 1.3955*PP/SD +9.1032*MaenAC-1. 6224	0.30	0.22	0.27	9.54	0.098	81	20
M16	JLc (Leukemia)	**S1**	=-7.0103*EMaNACH + 1.5761*HDSA2Q -1.0051* MaeeRC + 1.0507	0.86	0.82	0.33	44.61	0.069	30	11
M17	L120 (Leukemia)	**S15**	=-4.8986*HDCA2Q + 1.7065 *WNSA2z – 1.4993 *MienANN + 6.9159	0.90	0.84	0.17	43.50	0.066	25	6
M18	LLc (Lung)	**S1**	=- 1.5310 * ZXS/ZXR + 2.7870 *ERNCSQ- 6.4016* MiBOC(0.1)+ 7.9845	0.83	0.78	0.38	36.82	0.155	32	9
M19	M14 (Melanoma)	**S13**	=-8.1796*NN -5.0035*RNBr +1.1723* MiERIC+4.9692	0.81	0.70	0.25	28.09	0.156	30	6

M20	MCF7 (Breast)	**S2, S4, S5, S9, S10, S11, S13, S14, S16**	=6.4410*MaeeRC-3.4532* ERPCSQ -1.7867*ASIC1 -2.7106	0.46	0.44	0.55	52.87	0.663	231	45
M21	MLM (glioblastoma)	**S16**	=-8.2245*EFPSA1Q + 1.1671* ANRIO -4.4003*EHDSAQ	0.48	0.40	0.28	14.37	0.124	53	13
M22	H460 (Lung)	**S2, S3, S14**	=-1.3004* MiPC +6.3227* Ma1ERIC+ 2.4755*MaenACO	0.59	0.49	0.41	19.82	0.152	60	16
M23	P388 (Leukemia)	**S1**	=-3.4460*WPSA1z + 6.8634*MiTICN + 8.1021*HDCA1Q – 9.9755	0.81	0.73	0.31	31.74	0.0751	32	9
M24	Panc1 (Pancreatic)	**S3**	=1.8296*SIC2+ 3.3629*LNMVF - 1.7681* FPSA3Q-3.0118	0.87	0.73	0.11	21.63	0.016	16	3
M25	SF-468 (CNS)	**S14**	=2.4189*ABOC-3.9606* ERNCSQ +2.4054*EMaNAC -2.3761	0.74	0.56	0.27	9.38	0.171	18	4
M26	SJ-G2 (Brain)	**S2**	=5.1327*CIC2 +1.5429*AVN +2.0716* MiRECN -7.3156	0.77	0.60	0.13	18.67	0.034	26	9
M27	SKMEL-5 (Melanoma)	**S16**	=-1.2423*HOMO1+4.6277*MaTICH – 2.3144*EFHDSA -6.8356	0.51	0.45	0.22	15.52	0.111	51	15
M28	SKOV3 (Ovarian)	**S13**	=2.7606* MiPCNz - 3.5240*EE+eeRCC +1.3856*EFHDCAQ +5.0366	0.76	0.66	0.28	20.69	0.141	29	7

M29	SW480 (Colon)	**S2**	=7.0573*MaASEN - 605367* RNCl - 1.1217*HDSAQ+1.3202	0.69	0.	0.27	32.20	0.130	50	15
M30	U251 (CNS)	**S13**	=-1.1620*IOKSE + 5.3498*EFHBSAQ – 1.5086*RNN+7.2762	0.76	0.69	0.31	24.27	0.154	29	7

*** Same as footnote given in Table 2(a) for definition of the statistical parameters as well as other abbreviations.**

Figure.1: 482 compounds which have IC50 values represented into different scaffolds (S1-S16), the number of compounds in each scaffold in parenthesis and different cell lines against which the cytotoxicity values were reported (please see Tables S1-S16 in Additional file A for structure of all the compounds with their in vitro IC50 values against various cell lines).

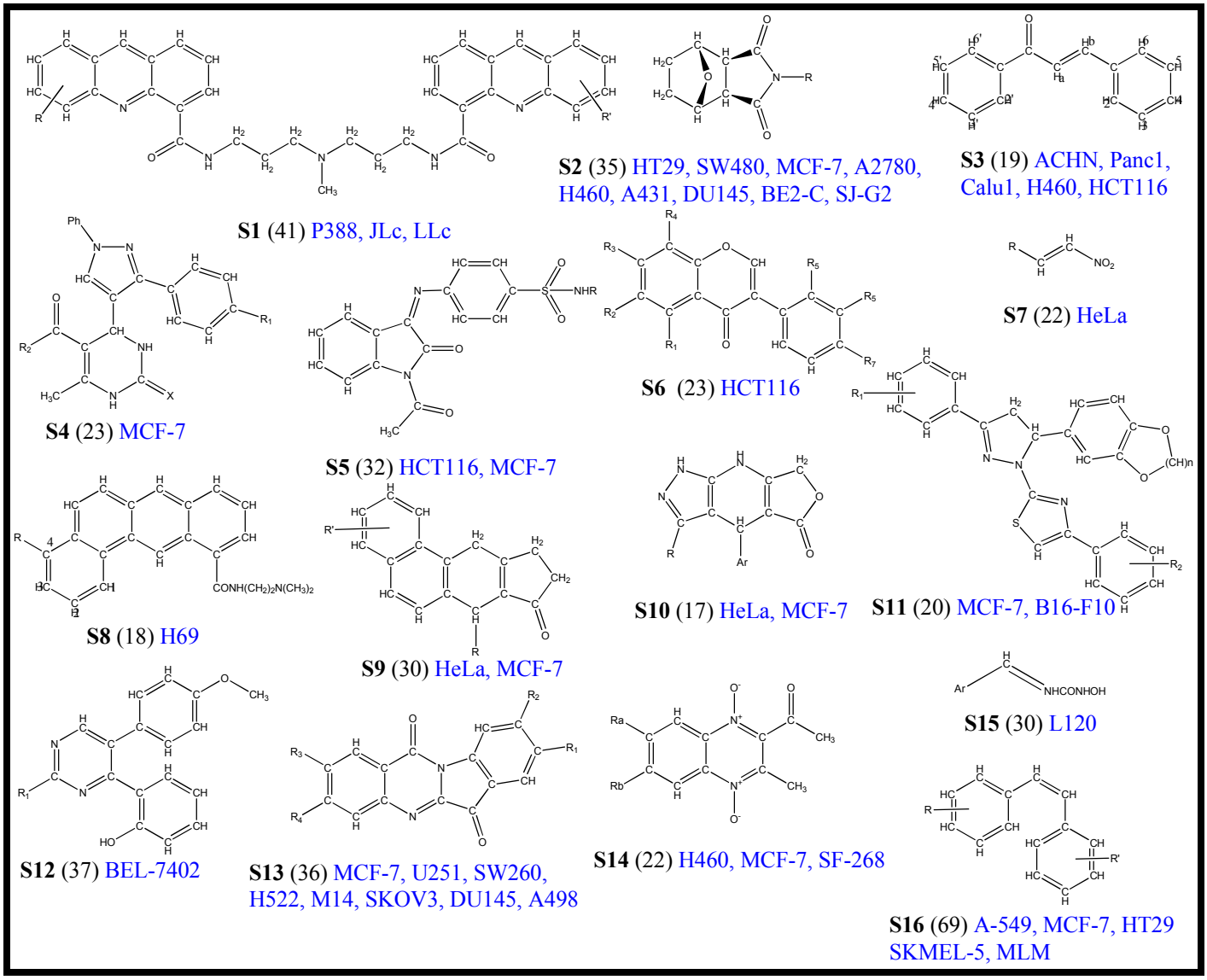


Figure.2: Flowchart for methodology accepted for developing and validating QSAR models.

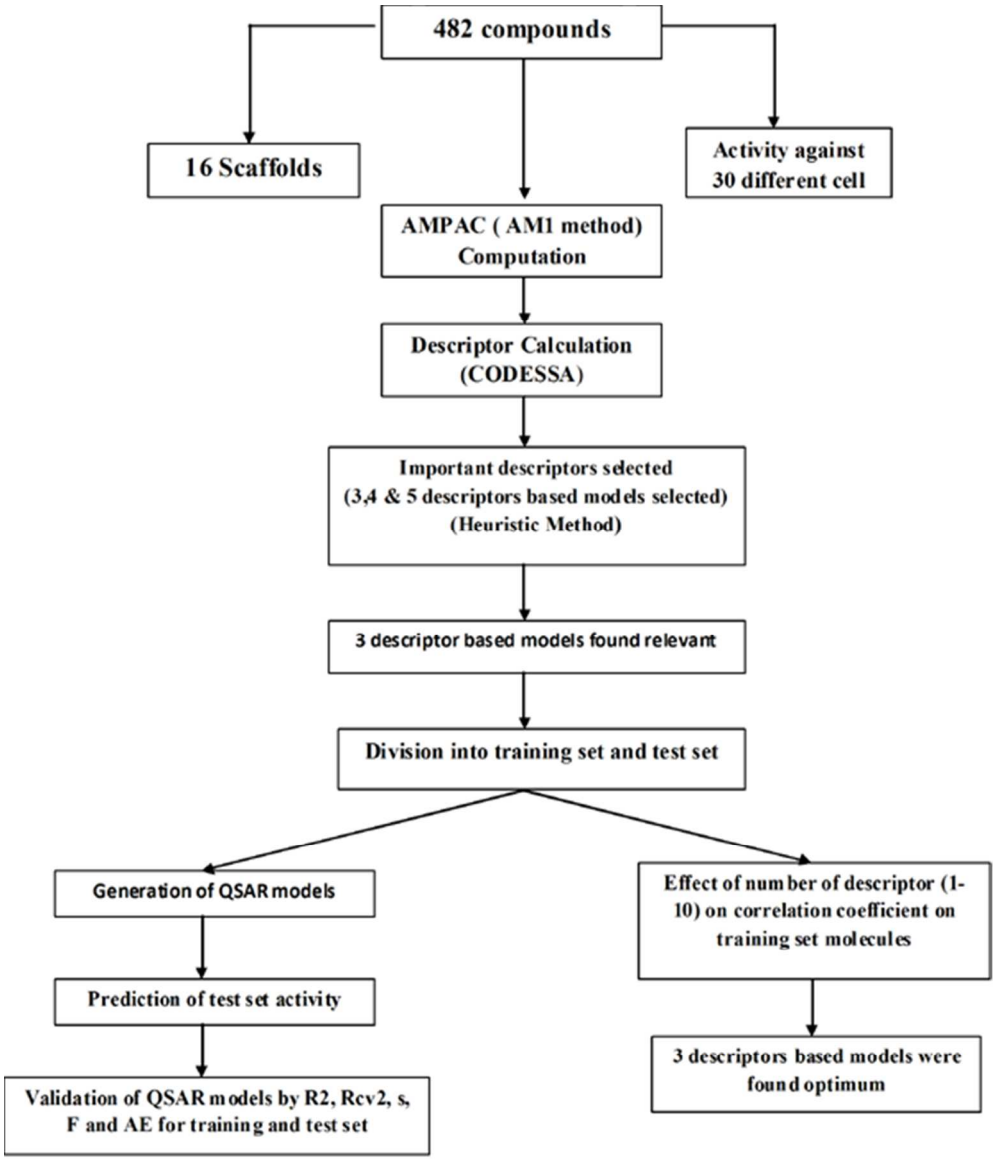
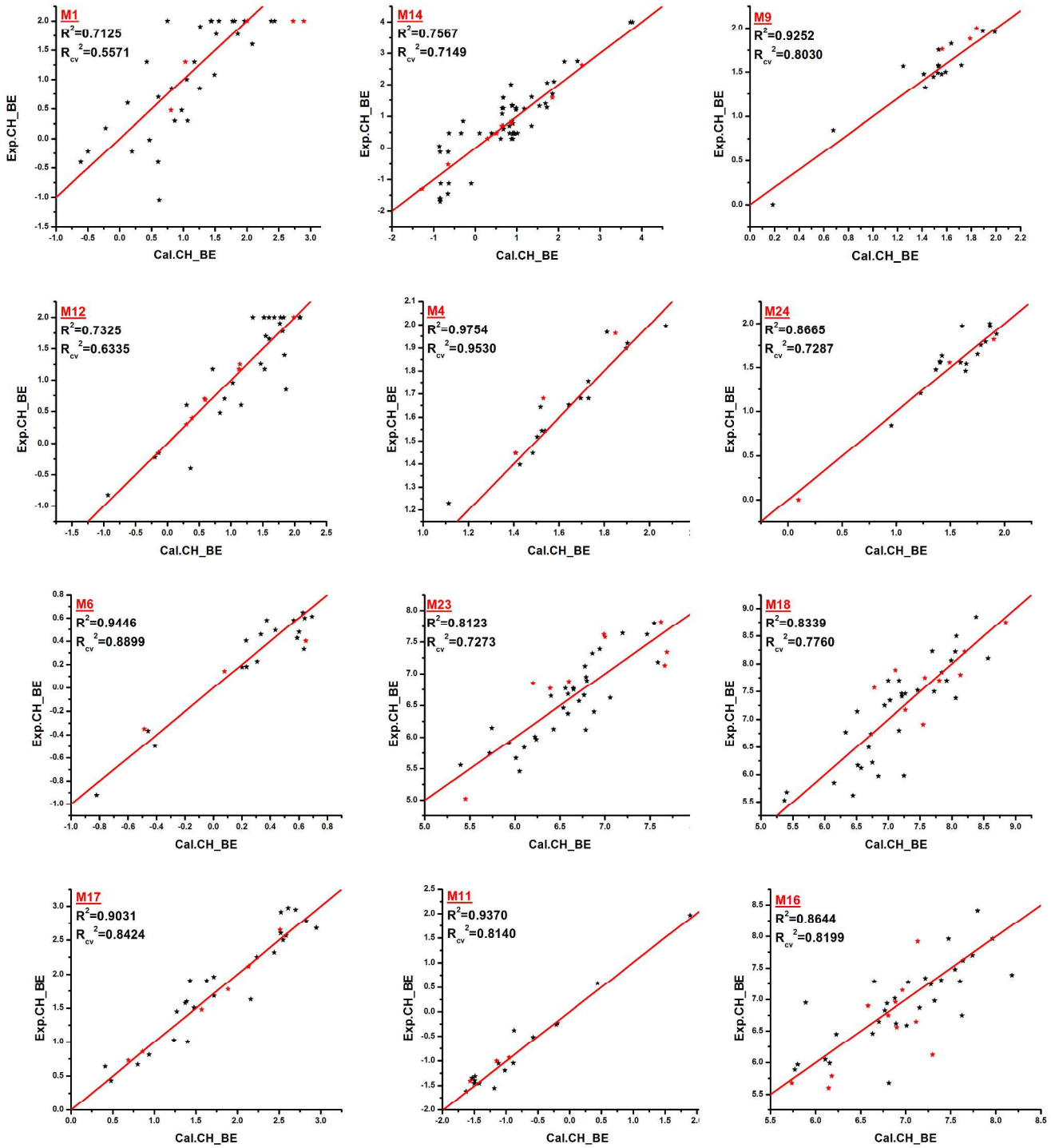
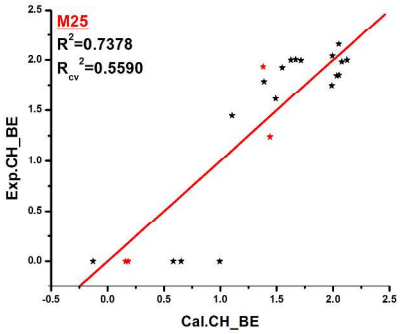
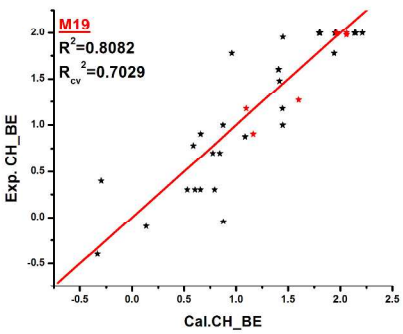


Figure.3(a): Plot between experimental and predicted IC_{50} values for cell lines based QSAR models with correlation coefficient and cross validation coefficient for high quality statistical 15 models.







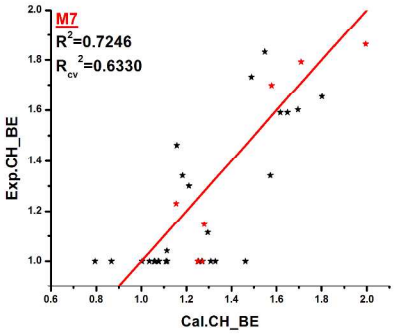


Figure.3(b): Plot between experimental and predicted IC_{50} values for scaffold based QSAR models with correlation coefficient and cross validation coefficient for high quality statistical 8 models.

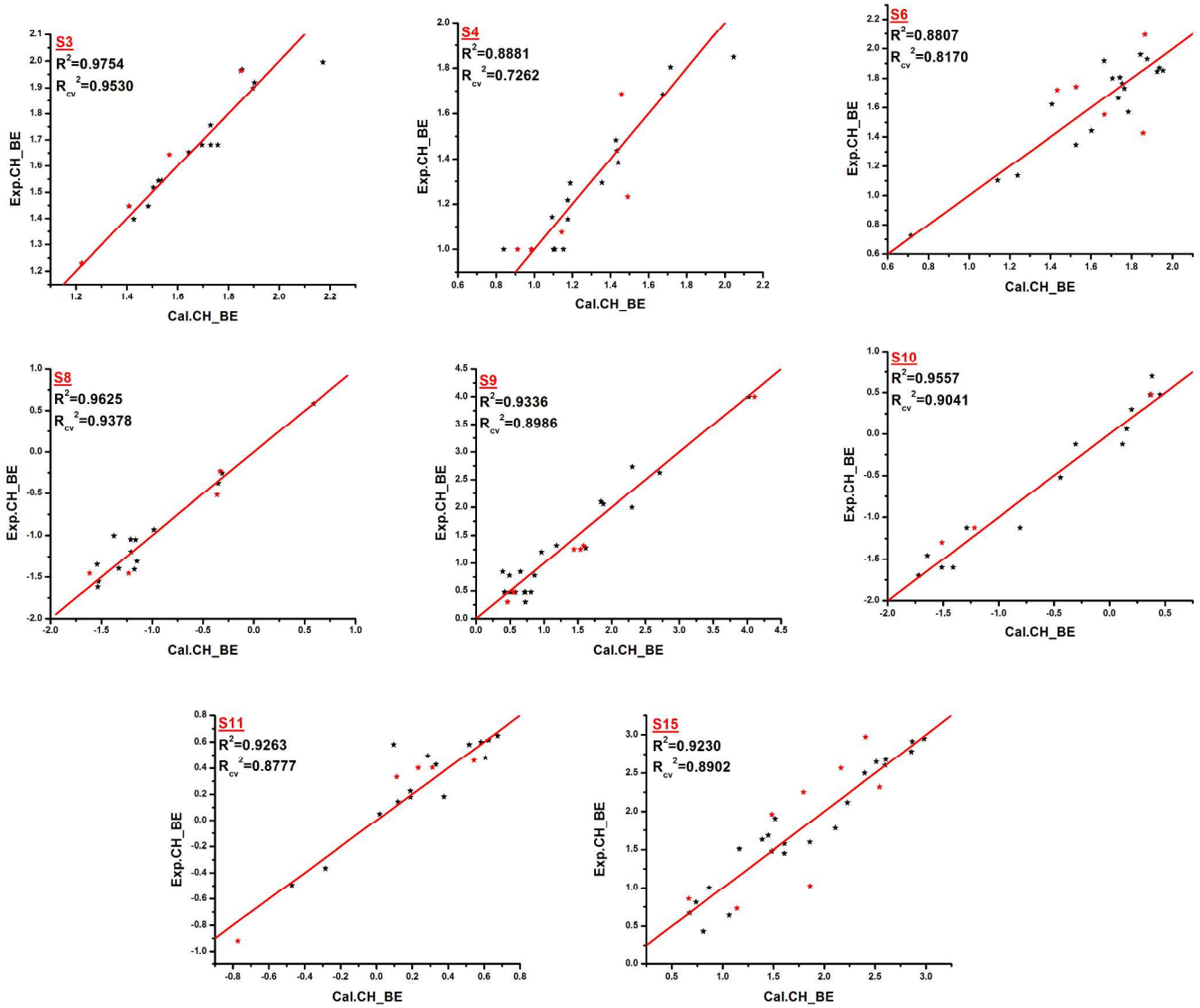


Figure.4(a): Effect of descriptor's number on the correlation coefficient on the basis of cell line-based QSAR models.

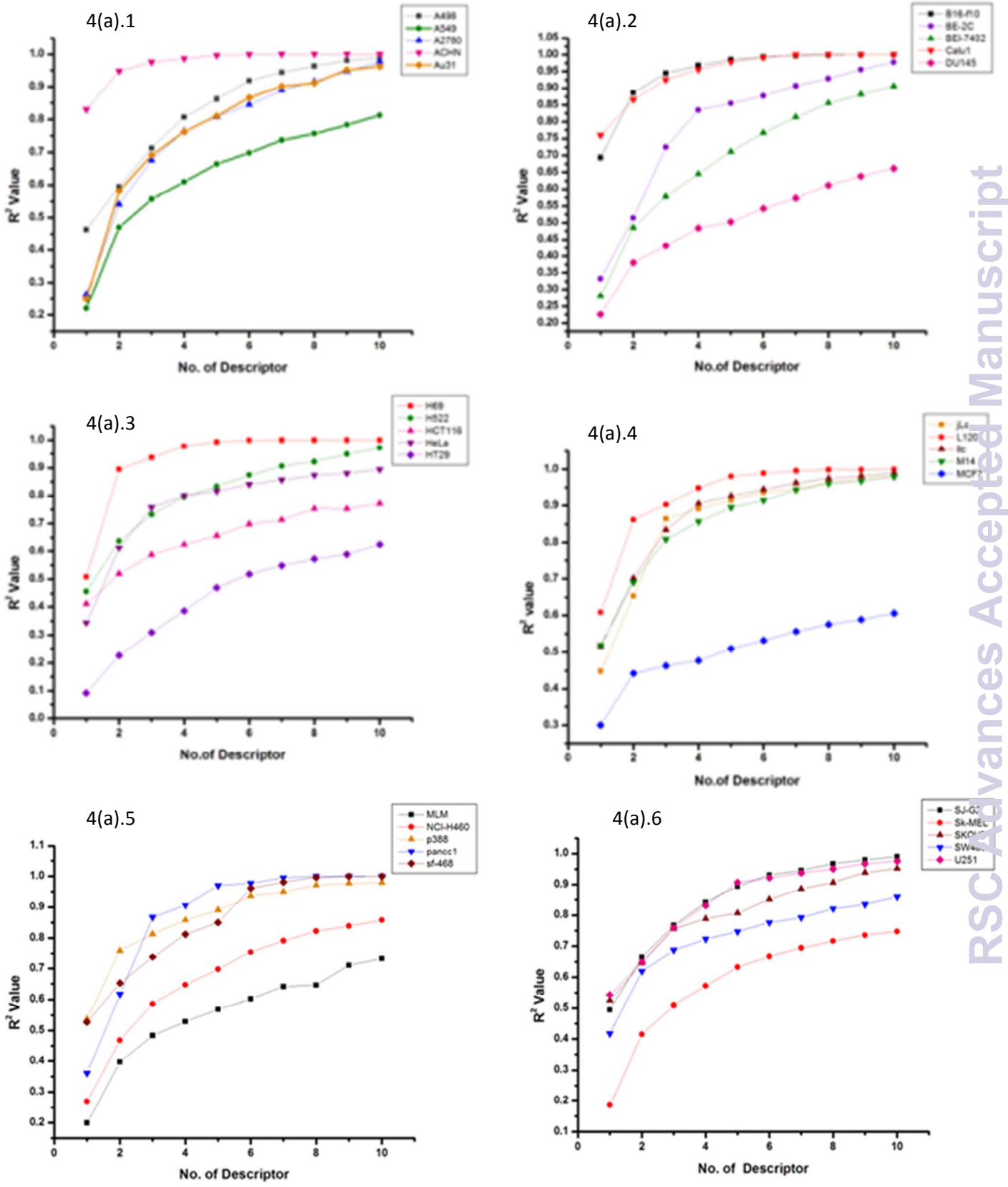


Figure .4(b): Effect of descriptor's number on the correlation coefficient on the basis of scaffold-based QSAR models.

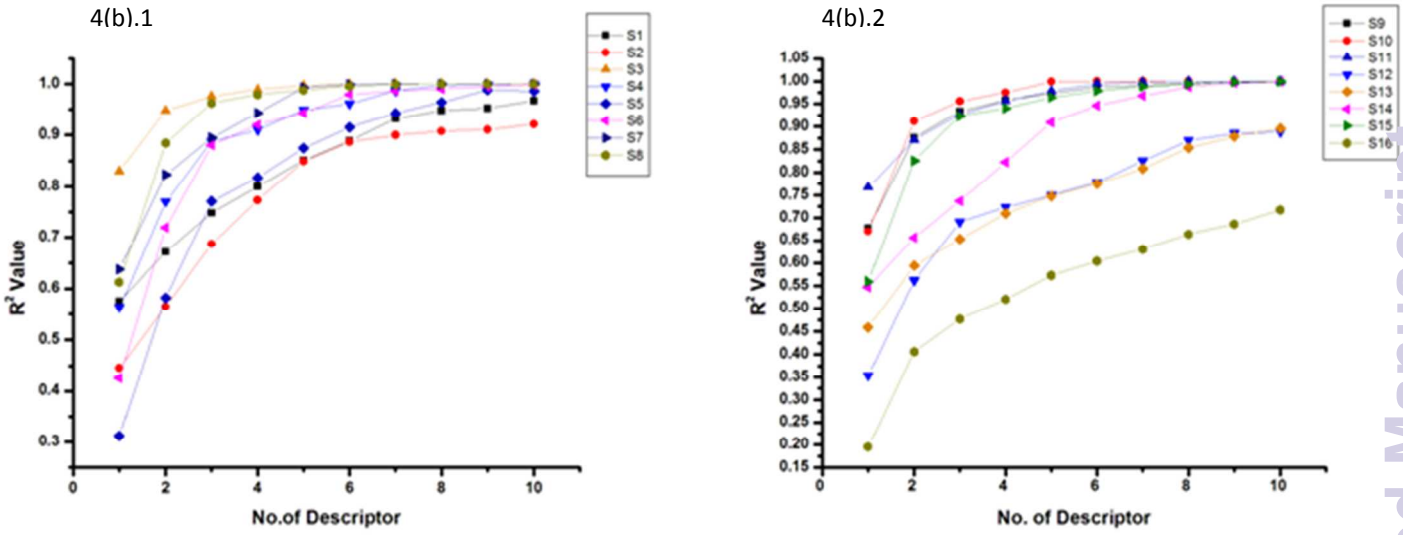
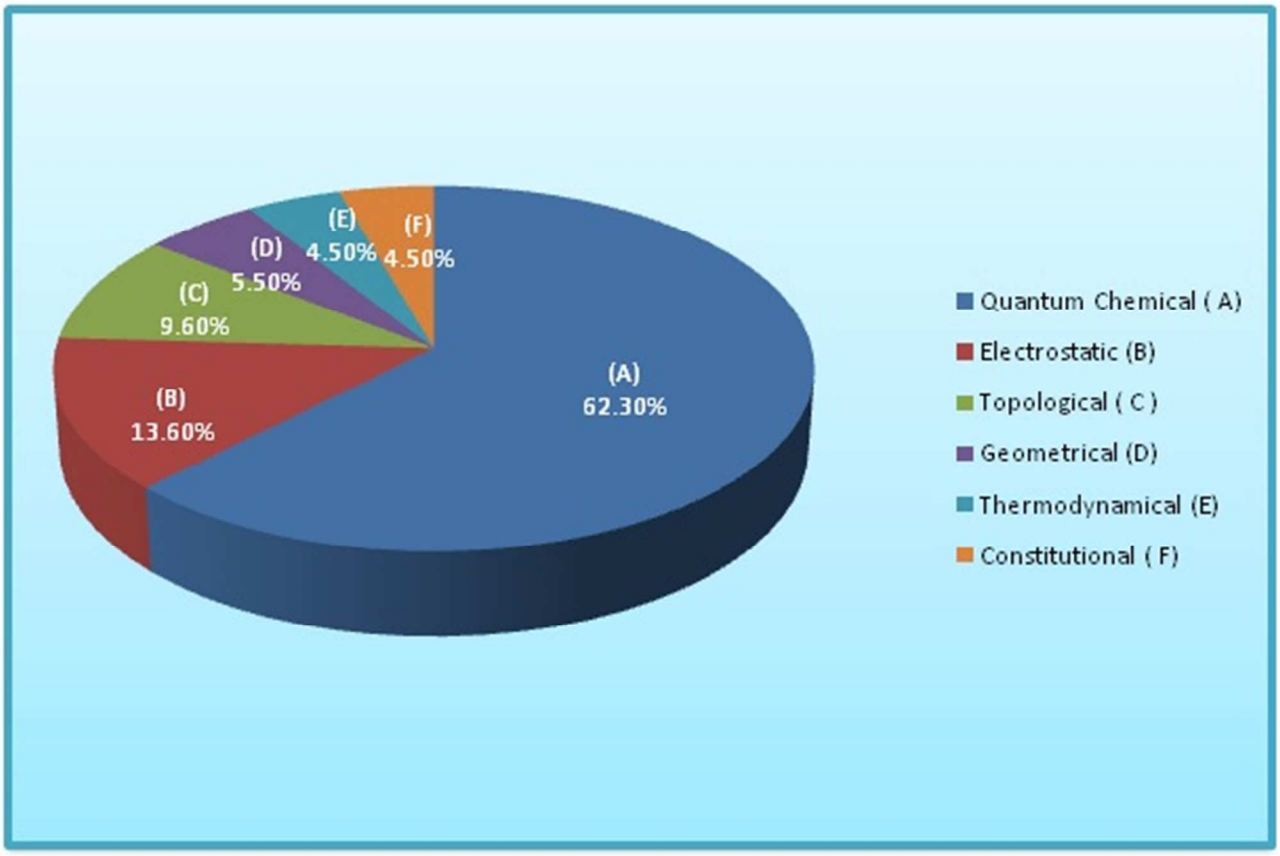


Figure.5: Percentage of various descriptors involved in QSAR models (See in Additional file A Table S17 for the details of all descriptors).



References:

1. A Frace, C Loge, S Gallet, N Lebegue, P Carato, P Chavatte, P Berthelot and D Lesieur, *J. Enzyme Inhib. Med. Chem.*, 2004, **19,** 541.
2. JB Gibbs, *Cancer research. Science,* 2000, **287,** 1969.
3. A Kamal, Y V V Srikanth, M N A Khan, T B Shaik and M Ashraf, *Bioorg. Med. Chem. Lett.,* 2010, **20,** 5229.
4. RJ Abdel-Jalil, EQE Momani, M Hamad, W Voelter, M S Mubarak, BH Smith and DG Peters, *Monatsh Chem,* 2010, **141,** 251.
5. P. Hanumantharao and S. V. Sambasivarao, *Bioorg. Med. Chem. Lett, 2*005, **15,** 3167.
6. T K Olszewski and B Boduszek, *Tetrahedron,* 2010, **66,** 8661.
7. DJ Newman, GM Cragg and KM Snader, *J Nat Prod.,* 2003, **66,** 1022.
8. R Breinbauer, IR Vetter and H Waldmann, *Angew Chem Int Ed,* 2002, **41,** 2879.
9. GM Cragg, PG Grothaus and DJ Newman, *Chem Rev,* 2009, **109**, 3012.
10. A L Harvey, *Drug Discovery Today,* 2008, **13,** 894.
11. DJ Snodin, *Toxicology Letters,* 2002, **127,** 161.
12. S Kraljevic, P J Stambrook, and K Pavelic, *EMBO Reports,* 2004, **5,** 837.
13. CP Adams and VV Brantner, *Health Aff (Millwood),* 2006, **25,** 420.
14. Critical Path Opportunities Reports Challenges and Opportunities Report – March, 2004.
15. CW Yap, Y Xue, ZR Li and YZ Chen, 2006, **6,** 1593.
16. RV Guido, G Oliva and AD Andricopulo, *Current Medicinal Chemistry,* 2008, **15,** 37.
17. A Schwaighofer, T Schroeter, S Mika and G Blanchard, *Comb. Chem. High Throuh. Scree*, 2009, **12,** 453.
18. LG Valerio, *Toxico App. Pharm.*, 2009, **241,** 356.
19. E Ahmed Kamal, M Vijaya Bharathi, D Janaki Ramaiah, J Dastagiri, A Surendranadha Reddy,

Viswanath, SNCVL Farheen Sultana, Pushpavalli, Manika Pal-Bhadra, H K Srivastava G. Narahari Sastry, A Juvekar, S Sen and S Zingde; *Bioorg. Med. Chem.,* 2010, **18,** 526.

20. F Xie, H Zhao, L Zhao, L Lou and Y Hu; *Bioorg. Med. Chem. Lett.,* 2009, **19,** 275.
21. YC Martin, *Perspect. Drug DiscoV,* 1998, **12,** 3.
22. U Norinder, *Perspect. Drug DiscoV,* 1998, **12,** 25.
23. DJ Maddalena, *Expert Opin. Ther. Pat,* 1998, **8,** 249.
24. H Kubinyi, *Drug DiscoV. Today,* 1997, **2**, 538.
25. C Hansch and T Fujita, *Am. Chem. Soc.*, Washington, DC, 1995, **606,** 1.
26. C Hansch and A Leo, *Am. Chem. Soc.*,Washington, DC, 1995.
27. R Benigni and A Giuliani, *Bioinformatics,* 2003, **19,** 1194.
28. C Hansch, A Leo, SB Mekapati and A Kurup, *Bioorg Med Chem* , 2004, **12,** 3391.
29. HK Srivastava, M Chourasia, D Kumar and GN Sastry, *J. Chem. Inf. Model,* 2011, **51**, 558.
30. MT Cronin and JC Dearden, *Quant. Struct. Act. Relat,* 1995, **14,** 117.
31. M Mwense, XZ Wang, FV Buontempo, N Horan, A Young and D Osborn, *SAR QSAR Environ Res,* 2006, **17,** 53.
32. M Zhao, Z Li, Y Wu, YR Tang, C Wang, Z Zhang and S Peng, *Eur. J. Med. Chem.,* 2007, **42,** 955.
33. AS Reddy, SP Pati, PP Kumar, HN Pradeep and GN Sastry, *Curr Protein Pept. Sci.,* 2007, **8,** 329.
34. FA Pasha, M Muddassar and SJ Cho, *Chem. Biol. Drug Des.,* 2009, **73,** 292.
35. HK Srivastava, FA Pasha and PP Singh, *Int. J. Quant. Chem.,* 2005, **103,** 237.
36. P Srivani, and G Narahar Sastry, *J Mol Graph Mod,* 2009, **27**, 676.
37. S. Janardhan, P. Srivani and G Narahari Sastry, *QSAR Comb. Sci.,* 2006, **25,** 860.
38. P. Sivaprakasam, A. Xie and RJ Doerksen, *Bio. Med. Chem.,* 2006, **14,** 8210.
39. JC Chen, Y Shen, SY Liao, LM Chen and KC Zheng, *Int J Quant Chem,* 2007, **107,** 1468.
40. S Zhang, L Wei, K Bastow, W Zheng, A Brossi, KH Lee and A Tropsha, *J Comput Aided Mol Des.,*

2007, **21,** 97.

41. R Garg, WA. Denny and C Hansch, *Bioorg Med. Chem.*, 2000, **8,** 1835.
42. TA. Hill, SG Stewart, SP Ackland, J Gilbert, B Sauer, JA Sakoff and A McCluskey; *Bioorg. Med. Chem.,* 2007, **15,** 6126.
43. BP Bandgar, SS Gawande, RG Bodade, JV Totre and CN Khobragade., *Bioorg. Med. Chem.,* 2010, **18,** 1364.
44. RK Yadlapalli, OP Chourasia, K Vemuri, M Sritharan and RS Perali, *Bioorg. Med. Chem. Lett.*, 2012, **22,** 2708.
45. M Kumar, K Ramasamy, V Mani, RK Mishra, AB Abdul Majeed, ED Clercq and B Narasimhan, *Arab. J. Chem.,* 2014, **7,** 396.
46. J Hyun, SY Shin, KM So, YH Lee and Y Lim, *Bioorg. Med. Chem. Lett.,* 2012, **22**, 2664.
47. R Mohan, N Rastogi, N Irishi, N Namboothiri, SM. Mobinc and D Pandaa, *Bioorg. Med. Chem.,* 2006, **14,** 8073.
48. JC Chen, L Qian, WJ Wu, LM Chen and KC Zheng,, *J. Mole. Struc: THEOCHEM,* 2005, **756**, 167.
49. IV Magedov, L Frolova, M Manpadi, UD Bhoga, H Tang, NM Evdokimo, O George, KH Georgiou, S Renner, M Getlic, TL Kinnibrugh, MA Fernandes, SV slambrouck, WFA Steelant, CB Shuster, S Rogelj, WAL van Otterlo, and A Kornienko; *J Med Chem.,* 2011, **23,** 4234.
50. HH Wang, KM Qiu, HE Cui, YS Yang, Y Luo, M Xing, XY Qui, LF Bai and HL Zhu, *Bioorg. Med. Chem.,* 2013, **21**, 448.
51. VM. Sharma, P Prasanna, KV Adi Seshu, B Renuka, CV Laxman Rao, G Sunil Kumar, C Prasad Narasimhulu, P Aravind Babu, RC Puranik, D Subramanyam, A Venkateswarlu, S Rajagopal, KB Sunil Kumar, C Seshagiri Rao, NVS Rao Mamidi, DS. Deevi, R Ajaykumarb and R Rajagopalanb, *Bioorg. Med. Chem. Lett.,* 2002, **12**, 2303.
52. B Zarranz, A Jaso, I Aldana and A Monge, *Bioorg. Med. Chem*., 2004, **12**, 3711.

53. S Ren, R Wang, K Komatsu, PB Krause, Y Zyrianov, CE McKenna, C Csipke, ZA. Tokes and EJ. Lien., *J. Med. Chem.,* 2002, **45**, 410.

54. M Cushman, D Nagarathnam D Gopal, AK Chakraborti, CM Lin and E Hamel, *J. Med. Chem.,* 1991, **34,** 2579.

55. (a) CODESSA version 2.0Semichem, 7204 Mullen, Shawnee, KS 66216 USA

(b) M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev*. 1996, **96**, 1027.

56. MH Bohari, HK Srivastava and GN Sastry, *Organic and Medicinal Chemistry Letters*, 2011, **1**, 3.

57**.** A. Golbraikh and A. Tropsha, *J. Mol. Graphics Model,* 2002, **20**, 269.