Original Research Article

Quantitative Structure Activity Relationship of new Isatin analogues as anti-breast cancer agent by different chemometrics tools

Abstract:

Breast cancer is the most common diagnosed cancer and the leading cause of related death in woman across the world. Nowadays, there are many effective chemotherapeutic agents used in the treatment of breast cancer, however due to the high side effects of these drugs, there is still an urgent need to develop new drugs for battle the disease. Computational chemistry is unique method in drug discovery which reduce cost. In this study 105 molecules were subjected to quantitative structure-activity relationship analysis to find the structure requirements for ligand binding. Then their structures were drowning in Hyperchem and also optimized, the structural invariants used in this study were those obtained from whole molecular structures: by both hyperchem and dragon. Four chemometrics method including MLR, FA-MLR, PCR and GA-PLS were employed to make connection between structural parameters and cytotoxic effects. GA-PLS showed Chemical ،Topological،Randic molecular ، Charge ،3D-Morse ،Functional ،Atom-centeredindices to be the most significant parameters on cytotoxic activity. The result of FA-MLR analysis revealed the effects of Chemical, Atom-centered, Galvez and Functional on the cytotoxic activity too. A comparison between the different statistical methods employed indicated that GA-PLS represented superior results and it could explain and predict 72% and 80% variances in the PIC_{50} data, respectively.

Keywords: Breast cancer, Isatin, MCF-7, QSAR, GA-PLS

Introduction:

Cancer is a diseases involving abnormal cell growth with the possible to invade or extent to other parts of the body. A report from the American Cancer Society showed in the United States, 15.5 million people with a history of cancer were living.

Each year, more than 40,000 people in different country receive a diagnosis of one of the types of cancer: for example breast, bladder, colon and rectal, endometrial, kidney and so on [1-3]. One of the best promising new of heterocyclic molecules having many interesting activity profiles are isatin and isatin derivatives. The isatin (1*H*-indole-2,3-dione) moiety is responsible for a wide spectrum of biological property such as anticancer, antibacterial, antifungal and antiviral in many synthetically versatile molecules [4–8]. Among these properties antineoplastic activities (breast cancer against MCF7) of this moiety was of our interest to study the quantitative structure-activity relationships of a series of 105 isatin derivatives reported in literature. Today the application of computational methods for designing newly biologically active compounds has opened a new window to modern drug discovery research. Computational methods can accelerate the procedure of discovering new drugs by designing new compounds and predicting potency or activity of them. Quantitative structure activity relationship (QSAR) studies provide pharmaceutical chemists valuable information that is useful for drug design and prediction of drug activity [9-13]. QSAR studies, as one of the most important areas in chemometrics, give information that is useful for molecular design and medicinal chemistry [4-8]. QSAR models are mathematical equations constructing a relationship between chemical scaffold and biological property. These models have another ability, which is providing a deeper knowledge about molecule design.

Linear QSAR models are mathematical equations that present us enough information about the mechanism of biological activity of compounds by constructing a relationship between chemical structures and biological activities. The most important step in building QSAR models is the appropriate representation of the structural and physicochemical features of chemical structures [14-17]. These features named molecular descriptors have high impact on the biological activity of compound [18-21]. Molecular descriptors have been classified into different categories such as physiochemical, constitutional, geometrical, topological, and quantum chemical descriptors. Dragon and hyperchem are two well-known computational softwares provide us more than 4000 of these descriptors [22,23].

Different QSAR methods including multiple linear regression (MLR), partial least squares combined with genetic algorithm for variable selection (GA-PLS), factor analysis–MLR (FA-MLR), principal component regression analysis (PCR) were used to make connections between structural descriptors and anti-cancer activity of compounds [24-27]. An important approach of the researchers in modification of the isatin moiety has been to establish a comprehensive structure–activity relationship (SAR), for this class of anti-cancer agents.

It has been shown that the introduction of electron-withdrawing halogens to the benzene ring of the isatin molecule showed with increased biological activity [28].The in vitro cytotoxic property of isatin bromo-derivatives were determined against the human monocyte-like, histiocytic lymphoma cell line (U937), appeared that the introduction of electronwithdrawing groups at positions C5, C6, and C7 significantly increased the cytotoxic activity when compared with isatin molecules, but the substitution at the 5-position being the best [29]. Substitution such as an aromatic ring with one or three carbon atom linker at N_1 enhances the activity too [30]. In this paper, it was of interest for us to investigate the QSAR of isatin derivatives that have been reported to exhibit anti-cancer activity against MCF7 in recent reports. Our QSAR analysis establishes mathematical relationship between biological activities and computable parameters such as chemical, topological, physicochemical, stereochemical or geometrical and so on indices.

2.Methods

2.1. Data set

The biological data used in this study were anti-cancer activity against MCF7, (in terms of -log IC_{50}), of a set of 105 isatin derivatives [31-36]. The data set was classified into calibration and prediction set by kenardston algorithm of the 25 prediction molecules from the spaces of the calculated descriptors. The structural features and biological activity of these compounds are listed in Table 1. Calculated descriptors for each molecule are summarized in Table 2.

[Table 1. near here], [Table 2. near here]

2.2. Descriptor generation

The structural features of the studied compounds are listed in Table 1. The two-dimensional structures of molecules were drawn by Hyperchem 8.0 software (Hypercube Inc.) to calculate whole molecular structure-based descriptors. The final geometries were obtained with semi-empirical AM1 calculations in Hyperchem program. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal mol^{-1} [22]. Some physicochemical parameters including molecular volume (V), molecular surface area (SA), hydrophobicity (Log P), hydration energy (HE) and molecular polarizability (MP) were calculated using Hyperchem Software. In order to calculate some molecular descriptors including topological, constitutional and functional group descriptors the optimized molecules were transferred into the Dragon package, developed by the Milano chemometrics and QSAR Group [23]. The calculated descriptors from whole molecular structures are briefly described in Table 2.

2.3. Data screening & model building

The selected descriptors from each class and the experimental data were analyzed by the stepwise regression SPSS (version 22.0) software. The calculated descriptors were collected in a data matrix whose number of rows and columns were the number of molecules and descriptors, respectively. Multiple linear regressions (MLR) and partial least squares (PLS) were used to derive the QSAR equations and feature selection was performed by the use of genetic algorithm (GA). MLR with factor analysis as the data pre-processing step for variable selection (FA-MLR) and principal component regression analysis (PCRA) methods were also used to derive the QSAR equations.

The resulted models were validated by leave-one out cross-validation procedure (using MATLAB software) to check their predictability and robustness.

A key step in QSAR modeling is evaluating model's stability and prediction ability. We used cross-validation and external test set for these proposes. Cross-validation has different variants such as leave-one-out (LOO), leave-group-out (LGO) and ν-fold. It was shown previously that LOO can leads to chance and overfitted models whereas LGO is more sensitive to chance variables [37]. Therefore, we used

LGO for model-validation utilizing correlation coefficient and root mean square error of cross-validation (*q2* and *RMSECV*, respectively) as scoring function. In addition, an external test set composed of 6 molecules was also used. The molecules in this set did not have contribution in the model step and thus their predicted values can give a final prediction power of the models as measured by correlation coefficient, root mean square errors of prediction, relative error of prediction (R^2_P, $RMSE_P$ and *REP*, respectively).

The PLS regression method used in this study was the NIPALS-based algorithm existed in the chemometrics toolbox of MATLAB software (version 12 Math work Inc.). Leave-one-out cross-validation procedure was used to obtain the optimum number of factors based on the Haaland and Thomas F-ratio criterion [38].

3. Results and discussion

3.1. MLR analysis

In the first step, separate stepwise selection-based MLR analyses were performed using different types of descriptors, and then, an MLR equation was obtained utilizing the pool of all calculated descriptors. The resulted QSAR models from different types of descriptors for the compounds (80 molecules as calibration and 25 molecules as prediction sets) are listed in Table 3.

[Table 3. near here]

The equation E1 of Table 3 shows among chemical descriptors, the positive effect of surface area of the molecules on cytotoxicity effect and it shows the positive effect of log p of the molecules on the activity. This equation shows the hydrophilic molecules shows better cytotoxic effect. The second equation of Table 3 demonstrated the effect of constitutional descriptors on the anti-cancer activity of these compounds. It explain the positive effect of nR05 (number of 5-membered rings), and nR09 (number of 9-membered rings). It also explain the negative effect of ns(number of sulfur) on the activity.

The effect of topological group counts parameter on anti-cancer activity of the studied compounds has been described by equation E_3 of Table 3. It shows that among topological DDr05, PJI2 (2D Petitjean shape index) and X3A (average connectivity index of order 3) have the positive effects on cytotoxic activity of the compounds.

The equation E_4 of Table 3 was found by using geometrical descriptors (E_4), which studied explains the positive effect of DDI index and negative effect of G1 compounds on the anti-cancer activity.

The equation E_5 of Table 3 shows the effect of functional group on anti-cancer activity. It explains the positive effect of nN-N (number of hydrazine derivatives) and nCar (number of aromatic C(sp2) on the activity.

The equation E_6-E_{17} of Table 3 demonstrated the effect of positive and negative effects of BCUT, Galvz topological Charge indices, 2D autocorrelations, Charge, Burden eigenvalues, RDF, 3D MoRSE, WHIM, GETAWAY and charge descriptors on the anti-cancer activity of these compounds.

The statistical parameters of prediction, listed in Table 4, indicate the suitability of the proposed QSAR model based on MLR analysis of molecular descriptors. The correlation coefficient of prediction is 0.65, which means that the resulted QSAR model could predict 59% of variances in the anti-cancer activity data. It has root mean square error of 0.23.

3.2. GA-PLS model

Multicolinearity is a real problem in MLR analysis. This problem in the descriptors is omitted by PLS analysis. In fact, in PLS analysis, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. This modeling method coincides with noisy data better than MLR, because a minimal number of latent variables are used for modeling in PLS. In GA-PLS analysis a variable selection method is used to find the more convenient set of descriptors because redundant variables degrade the performance of PLS analysis, similar to other regression methods. In the present study, GA was used as variable selection method. The data set (n = 105) was divided into two groups: calibration set (n = 80) and prediction set (n = 25). Given 80 calibration samples; cross-validation procedure was used to find the optimum number of latent variables for each PLS model. In this work, in each run of GA-PLS method a large number of acceptable models were created. GA produces a population of acceptable models in each run. In this work, many different GA-PLS runs were conducted using different initial set of populations (50-250) and therefore a large number of acceptable models were created. The most convenient GA-PLS model that resulted in the best fitness contained 10 descriptors including, Chemical, Topological, Randic molecular, Charge, 3D-Morse, Functional, Atom-centered. All of them being those obtained by different MLR-based QSAR models. The PLS estimate of the regression coefficients are shown in Figure 1.

This model not only has a high cross-validation statistics, but also represents a high ability for modeling external test samples. It could explain and predict about 72% of variances in the anti-cancer activity of the studied molecules. There is a close agreement between the experimental and predicted values of anti-cancer activity data.

To measure the significance of the 10 selected PLS descriptors in anti-cancer activity; In order to investigate the relative importance of the variable appeared in the final model obtained by GA-PLS method, variable important in projection (VIP) was employed [39]. VIP values reflect the importance of terms in PLS model. According to Erikson *et al.* X-variables (predictor variables) could be classified according to their relevance in explaining y (predicted variable), so that VIP > 1.0 and VIP < 0.8 mean highly or less influential, respectively, and 0.8 < VIP< 1.0 means moderately influential. The VIP analysis of PLS equation is shown in Figure 2. As it is observed, PJI2, JGI3 and PHP2 indices represent the most significant contribution in the resulted QSAR model. In addition, functional group parameters such as mor17v and mor18u have been found to be moderately influential parameters.

[Figure 1. Near here], [Figure 2. Near here]

3.3. FA-MLR and PCRA

FA-MLR was performed on the dataset. Factor analysis (FA) was used to reduce the number of variables and to detect structure in the relationships between them. This data-processing step is applied to identify the important predictor variables and to avoid collinearities among them [40]. Principle component regression analysis, PCRA, was tried for the dataset along with FA-MLR. With PCRA collinearities among **X** variables are not a disturbing factor and the number of variables included in the analysis may exceed the number of observations [41]. In this method, factor scores, as obtained from FA, are used as the predictor variables [40]. In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors.

Table 5 shows the four factor loadings of the variables (after VARIMAX rotation) for the compounds tested for cytotoxic activity. As it is observed, about 73% of variances in the original data matrix could be explained by the selected seven factors.

Based on the procedure explained in the experimental section, the following three-parametric equation was derived (Table 6).

$Y=4.229\ (\pm0.636) + 0.006(\pm0.001)\ SA+0.061(\pm0.018)\ H047-0.607\ (\pm0.10)\ C028-18.706(\pm5.057)\ JGI3-0.731(\pm0.275)\ nNN$

$R^2= 0.66$, $Q^2= 0.61$, $F=21.73$, $SE= 0.19$

This equation could explain about 66% of the variance and predict 61% of the variance in pIC_{50} data. It has a root mean square error of 0.19. This equation describes the effect of SA, H047, C028, JGI3 and nNN on cytotoxic activity of the studied molecules.

When factor scores were used as the predictor parameters in a multiple regression equation using forward selection method (PCRA), the following equation was obtained (Table 7):

$Y=5.026\ (\pm0.96) + 0.508\ (\pm.097)\ F1+ 0.250\ (\pm0.097)\ F2-0.211(\pm0.097)\ F3$

$R^2=0.77$, $Q^2=0.72$, $F=12.97$, $SE=0.23$

This equation could explain and predict 77% and 72% of the variances in pIC_{50} data, respectively. The root mean square error of PCRA analysis was 0.23. Since factor scores are used instead of selected descriptors, and any factor-score contains information from different descriptors, loss of information is thus avoided and the quality of PCRA equation is better than those derived from FA-MLR. Whilst the data of this analysis show acceptable prediction, we see that the predicted values of some molecules are near to each other.

[Table 5 near here], [Table 6 near here], [Table 7 near here]

As it is observed from Table 5, in the case of each factor, the loading values for some descriptors are much higher than those of the others. These high values for each factor indicate that this factor contains higher information about which descriptors. It should be noted that all factors have information from all descriptors but the contribution of descriptor in different factors are not equal. For example, factors 1 and 2 have higher loadings for the chemical, constitutional, Functional, Atom-center, BCUT Information, geometrical, Walk and path counts and 2D autocorrelations indices whereas information about the Connectivity indices, 3D WHIM, MoRSE descriptors and Functional descriptors are highly incorporated in factor 3 descriptors.

3.5. Robustness and applicability domain of the models

Leverage is one of standard methods for this purpose. Warning leverage (h^*) is another criterion for interpretation of the results. The warning leverage is, generally, fixed at $3k/n$, where n is the number of training compounds and k is the number of model parameters. A leverage greater than warning leverage h^* means that the predicted response is the result of substantial extrapolation of the model and therefore may not be reliable [42]. The calculated leverage values of the test set samples for different models and the warning leverage, as the threshold value for accepted prediction, are listed in Table 8. As seen, the leverages of all test samples are lower than h^* for all models. This means that all predicted values are acceptable.

[Table.8 near here]

4.Conclusions

Quantitative relationships between molecular structure and anti-cancer activity of isatin derivatives were discovered by four chemometrics methods: MLR, GA-PLS, PCR and FA-MLR. MLR analysis show positive effect of the log p, nR05 (number of 5-membered rings), nR09 (number of 9-membered rings) DDr05, (2D Petitjean shape index) and X3A (average connectivity index of order 3), DDI index, nN-N (number of hydrazine derivatives) and nCar (number of aromatic C(sp2) of the molecules on the activity. It also explain the negative effect of ns (number of sulfur) and G1on the activity. The FA-MLR describes the effect of SA, H047, C028, JGI3 and nNN on cytotoxic activity of the studied molecules. The quality of PCRA equation is better than those derived from FA-MLR. factors 1 and 2 have higher loadings for the chemical, constitutional, Functional, Atom-center, BCUT Information, geometrical, Walk and path counts and 2D autocorrelations indices whereas information about the Connectivity indices, 3D WHIM, MoRSE descriptors and Functional descriptors are highly incorporated in factor 3 descriptors.

A comparison between the different statistical methods employed revealed that GA-PLS represented superior results and it could explain and predict 80% and 72% of variances in the pIC_{50} data, respectively.

ACKNOWLEDGEMENTS

Financial support of Shiraz University of Medical Sciences (Shiraz, I.R. Iran) through the Grant No. 94-01-36-11242 is highly appreciated.

References:

1. Erfani N, Nazemosadat Z, Moein M. Cytotoxic activity of ten algae from the Persian Gulf and Oman Sea on human breast cancer cell lines; MDA-MB-231, MCF-7, and T-47D. Pharmacognosy research. 2015;7(2):133.
2. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, et al. International classification of diseases for oncology :World Health Organization; 2000.
3. Mousavi SM, Gouya MM, Ramazani R, Davanlou M, Hajsadeghi N, Seddighi Z. Cancer incidence and mortality in Iran. Annals of Oncology. 2009;20(3):556-63.
4. S.N. Pandeya, S. Smitha, M. Jyoti, S.K. Sridhar, Acta Pharm. 55, (2005) 27–46.
5. V.M. Sharma, P.Prasanna, V.A. Seshu, B. Renuka, V.L. Rao, G.S. Kumar, C.P. Narasimhulu, P.A. Babu, R.C. Puranik, D. Subramanyam, A. Venkateswarlu, S. Rajagopal, K.B.S. Kumar, C.S. Rao, N.V.S. R. Mamidi, D.S. Deevi, R. Ajaykumar, R. Rajagopalan, Bioorg. Med. Chem. Lett. 12, (2002) 2303–2307.
6. M.J. Moon, S.K. Lee, J.-W. Lee, W.K. Song, S.W. Kim, J.I. Kim, C. Cho, S.J. Choi, Y.-C. Kim, Bioorg. Med. Chem. 14, (2006) 237–246.
7. A.H. Abadi, S.M. Abou-Seri,; D.E. Abdel-Rahman, C. Klein, O. Lozach, L. Meijer, Eur. J. Med. Chem. 41, (2006) 296–305.
8. A. Gursoy, N. Karali, Eur. J. Med. Chem. 38, (2003) 633–643.
9. H. Schmidi, Multivariate prediction for QSAR, Chemom. Intell. Lab. Syst. 37 (1997) 125-134.
10. C. Hansch, A. Kurup, R. Garg, H. Gao, Chem-bioinformatics and QSAR: A review of QSAR lacking positive hydrophobic terms, Chem. Rev. 101(2001) 619-672.
11. S. Wold, J. Trygg, A. Berglund, H. Antii, Some recent developments in PLS modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131-150.
12. Sabet R.; Fassihi A.; Hemmateenejad B.; Saghaie L.; Miri R.; Gholami M.; Computer-aided drug design of novel antibacterial 3-hydroxypyridine-4-ones: application of QSAR methods based on the MOLMAP approch. Journal of Computer-Aided Molecular Design. 2012, 26,349-361.

13. Sabet, R.; Fassihi, A.; Moeinifard, B., QSAR study of PETT Derivatives as Potent HIV-Reverse Transcriptase Inhibitors. *J. Mol. Graph & Model***.** 2009, 28, 146-155.

14. C. Hansch, T. Fujita, ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc. 86 (1964) 1616-1626.

15. J. Wang, L. Zhang, G. Yang, C.G. Zhan, Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: A study of quantum chemical descriptors from density functional theory, J. Chem. Inf. Comput. Sci. 44 (2004) 2099-2105.

16. C. Hansch, D. Hoekman, H. Gao, Comparative QSAR: Toward a deeper understanding of chemicobiological interactions, Chem. Rev. 96 (1996) 1045-1075.

17. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, 2000.

18. D. Horvath, B. Mao, Neighborhood behavior. Fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity, QSAR Comb. Sci. 22 (2003) 498-509.

19. S. Putta, J. Eksterowicz, C. Lemmen, R. Stanton, A novel subshape molecular descriptor, J. Chem. Inf. Comput. Sci. 43 (2003) 1623-1635.

20. S. Gupta, M. Singh, A.K. Madan, Superpendentic index: A novel topological descriptor for predicting biological activity. J. Chem. Inf. Comput. Sci. 39 (1999) 272-277.

21. V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies, J. Chem. Inf. Comput. Sci. 42(2002) 693-705.

22. HyperChem, Release 8.0 for Windows, Molecular Modeling System: HyperCube.

23. Todeschini, R. Milano Chemometrics and QSAR Group. http://michem.disat.unimib.it/.

24. Fassihi, A.; Sabet, R., QSAR Study of p56^{lck} Protein Tyrosine Kinase Inhibitory Activity of Flavonoid Derivatives Using MLR and GA-PLS. *Int. J. Mol. Sci*. 2008, 9, 1876-1892.

25. Sabet, R.; Fassihi, A., QSAR Study of Antimicrobial 3-Hydroxypyridin-4-one and 3-Hydroxypyran-4-one Derivatives Using Different Chemometric Tools. *Int. J. Mol. Sci*. 2008, 9, 2407-2423.
26. Fassihi, A.; Abedi, D.; Saghaie, L.; Sabet, R.; Fazeli, H.; Bostaki, Gh.; Deilami, O.; Sadinpour, H., Synthesis, Antimicrobial Evaluation and QSAR Study of Some 3-hydroxypyridine-4- one and 3-hydroxypyran-4-one Derivatives. *Eur. J. Med. Chem.* 2009, 44, 2145-2157.
27. V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. 42 (2002) 693-705.
28. K.L. Vine, J.M. Locke, M. Ranson, K. Benkendorff, S.G. Pyne, J.B. Bremner, Bioorg. Med. Chem. 15, (2007) 931–938.
29. K.L. Vine, J.M. Locke, M. Ranson, S.G. Pyne, J.B. Bremner, J. Med. Chem. 50, (2007) 5109–5117.
30. K. Kumar, S. Sagar, L. Esau, M. Kaur, V. Kumar, Eur. J. Med. Chem. 58 (2012) 153.
31. Solomon VR, Hu C, Lee H. Hybrid pharmacophore design and synthesis of isatin–benzothiazole analogs for their anti-breast cancer activity. Bioorganic & medicinal chemistry. 2009;17(21):7585-92.
32. Solomon VR, Hu C, Lee H. Design and synthesisof anti-breast cancer agents from 4-piperazinylquinoline: a hybrid pharmacophore approach. Bioorganic & medicinal chemistry. 2010;18(4):1563-72.
33. Karthikeyan C, Solomon VR, Lee H, Trivedi P. Design, synthesis and biological evaluation of some isatin-linked chalcones as novel anti-breast cancer agents: A molecular hybridization approach. Biomedicine & Preventive Nutrition. 2013;3(4):325-30.
34. Taher AT, Khalil NA, Ahmed EM. Synthesis of novel isatin-thiazoline and isatin-benzimidazole conjugates as anti-breast cancer agents. Archives of pharmacal research. 2011;34(10):1615-21.
35. Vine KL, Locke JM, Ranson M, Pyne SG, Bremner JB. In vitro cytotoxicity evaluation of some substituted isatin derivatives. Bioorganic & Medicinal Chemistry. 2007;15(2):931-8.
36. Li P-K, Xiao Z, Hu Z, Pandit B, Sun Y, Sackett DL, et al. Conformationally restricted analogs of Combretastatin A-4 derived from SU5416. Bioorganic & medicinal chemistry letters. 2005;15(24):5382-5.

37. Leardi, R. Genetic Algorithms in Chemometrics and Chemistry: A Review. *J. Chemometrics.* 2001, *15*, 559-569.

38. Sabet R.; Fassihi A.; Saghaie L., Octanol-water partition coefficients determination and QSPR study of some 3-hydroxy pyridine-4-one derivatives, *Journal of Pharmaceutical Research International.* 2018 .22(4), 1-15.

39. Olah, M.; Bologa, C.; Oprea, T.I. An Automated PLS Search for Biologically Relevant QSAR Descriptors. *J. Comput. Aided Mol. Des.* 2004, *18*, 437-449.

40. R. Franke, A. Gruska, Chemometrics Methods in molecular design, in: H. van Waterbeemd, (Ed.), Methods and Principles in Medicinal Chemistry, VCH, Weinheim, 1995, Vol. 2, pp. 113–119.

41. H. Kubinyi, The quantitative analysis of structure-activity relationships, in: M.E. Wolff, (Ed.), Burger's Medicinal Chemistry and Drug Discovery, 5th Ed.; Wiley, New York, 1995, Vol. 1, pp. 506-509.

42. Brereton R. Chemometrics Data Analysis for the Laboratory and Chemical Plant. Wiley. 2004:47–54.

Figure 1. PLS regression coefficients for the variables used in GA-PLS model.

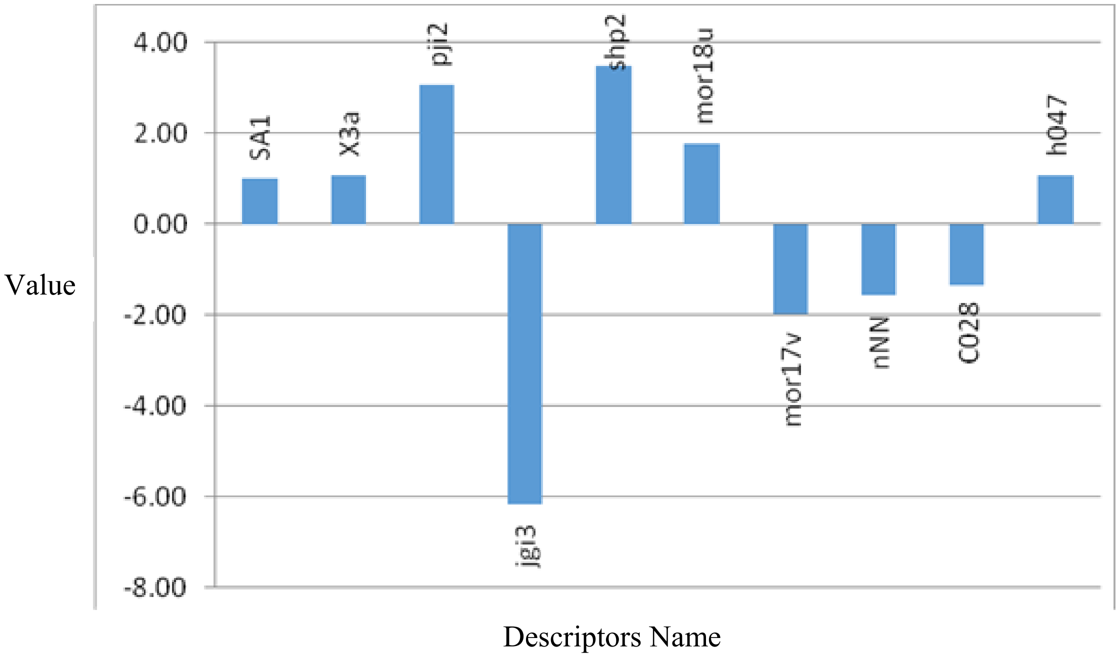


Figure 2. Plot of variables important in projection (VIP) for the descriptors used in GA-PLS model.

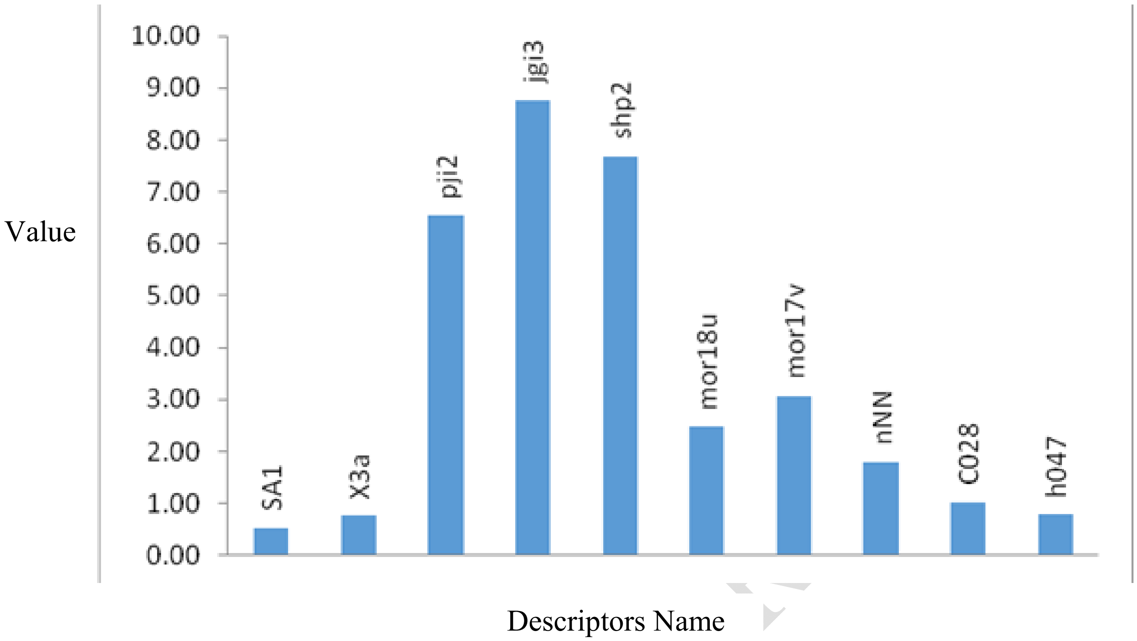
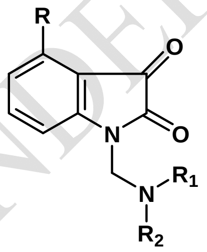
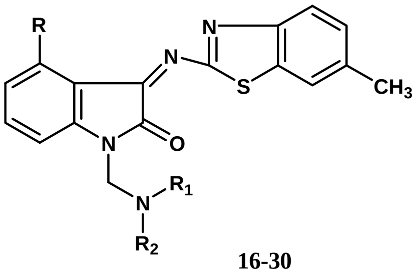


Table1. Chemical structure of isatin derivatives used in this study



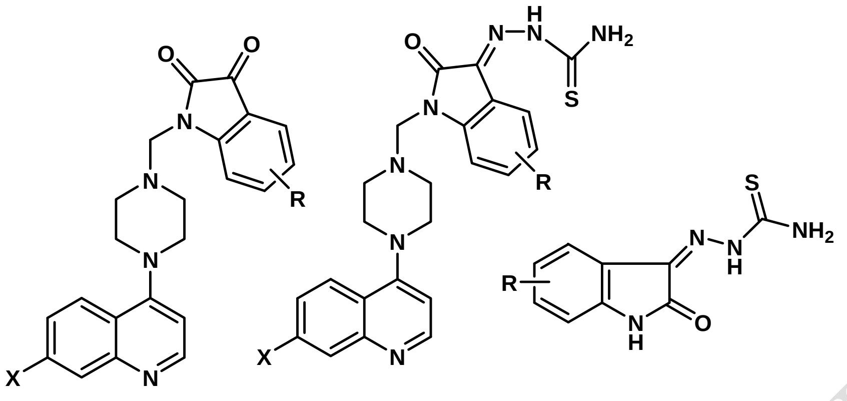
1-15



16-30

Compound	R	R_1 , R_2	pIC_{50}
1	H	$(CH_3)_2$	4.39
2	H	$(CH_2CH_3)_2$	4.36
3	H	$(C_6H_5)_2$	4.62
4	H	Piperidinyl	4.38
5	H	Morpholinyl	4.28

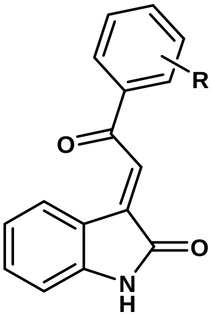
6	Cl	$(CH_3)_2$	4.06
7	Cl	$(CH_2CH_3)_2$	4.47
8	Cl	$(C_6H_5)_2$	4.53
9	Cl	Piperidinyl	4.29
10	Cl	Morpholinyl	4.53
11	Br	$(CH_3)_2$	4.24
12	Br	$(CH_2CH_3)_2$	4.69
13	Br	$(C_6H_5)_2$	3.96
14	Br	Piperidinyl	4.50
15	Br	Morpholinyl	4.18
16	H	$(CH_3)_2$	4.53
17	H	$(CH_2CH_3)_2$	4.13
18	H	$(C_6H5)_2$	4.41
19	H	Piperidinyl	4.42
20	H	Morpholinyl	4.84
21	Cl	$(CH_3)_2$	4.68
22	Cl	$(CH_2CH_3)_2$	4.72
23	Cl	$(C_6H_5)_2$	4.61
24	Cl	Piperidinyl	4.51
25	Cl	Morpholinyl	4.67
26	Br	$(CH_3)_2$	4.75
27	Br	$(CH_2CH_3)_2$	4.52
28	Br	$(C_6H_5)_2$	4.52
29	Br	Piperidinyl	4.41
30	Br	Morpholinyl	4.92



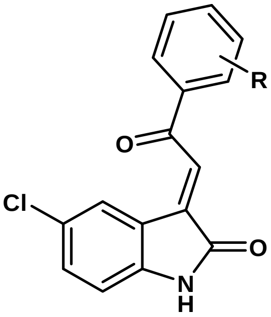
31-40 **41-50** **51-55**

Compound	X	R	PIC$_{50}$
31	Cl	4-Cl	4.41
32	Cl	4-Br	4.82
33	Cl	6-Cl	4.58
34	Cl	6-Br	4.75
35	Cl	H	4.32
36	CF3	4-Cl	4.77
37	CF3	4-Br	4.18
38	CF3	6-Cl	4.80
39	CF3	6-Br	4.55
40	CF3	H	4.47
41	Cl	4-Cl	4.24
42	Cl	4-Br	4.33
43	Cl	6-Cl	4.65
44	Cl	6-Br	4.94
45	Cl	H	4.64
46	CF3	4-Cl	4.67
47	CF3	4-Br	4.30
48	CF3	6-Cl	4.61
49	CF3	6-Br	4.45
50	CF3	H	4.86
51	-	4-Cl	4.52
52	-	4-Br	4.83
53	-	6-Cl	4.76

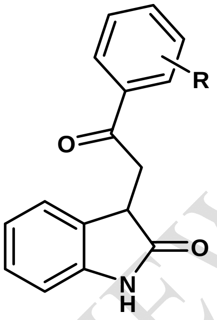
54	-	6-Br	4.75
55	-	H	4.43



56-65

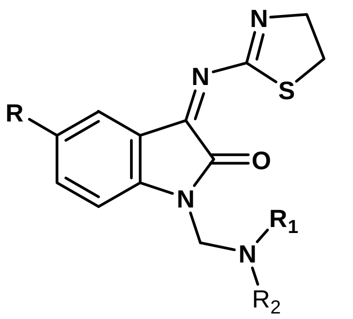


66-68

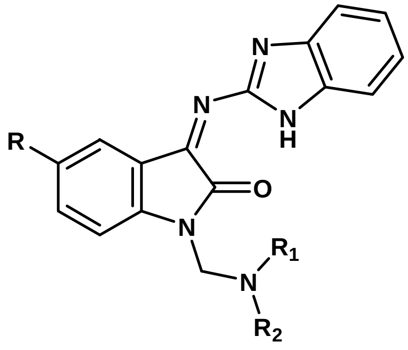


69

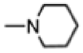
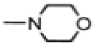
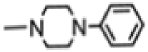
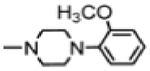
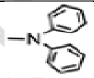
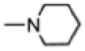
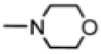
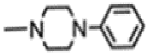
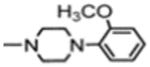
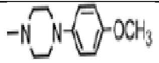
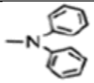
Compound	R	PIC$_{50}$
56	H	5.37
57	4-F	5.25
58	4-Cl	5.09
59	4-Br	4.98
60	4-NO$_2$	5.02
61	3-NO$_2$	5.11
62	4-OCH$_3$	5.09
63	2-OCH$_3$	5.17
64	3,4-OCH$_3$	5.20
65	4-CH$_3$	5.09
66	H	5.08
67	4-F	5.27
68	3,4-OCH$_3$	5.44
69	H	5.20

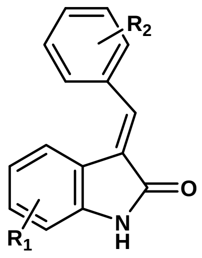


70-74

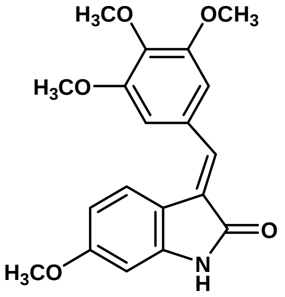


75-80

Compound	R	R_1 , R_2	pIC_{50}
70	H		7.22
71	H		7.32
72	H		7.42
73	H		7.29
74	H		7.37
75	H		7.46
76	H		7.60
77	H		7.19
78	H		7.58
79	H		7.37
80	H		7.65

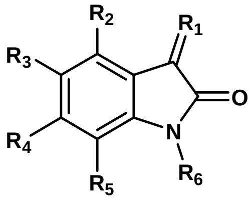


81-86



87

Compound	R_1	R_2	pIC_{50}
81	6-OCH3	3’-OH	6.41
82	6-OCH3	4’-OH	6.64
83	6-H	3’-OH	4.50
84	6-H	4’-OH	4.59
85	6-OH	3’-OH	5.07
86	6-OH	4’-OH	4.71
87	-	-	9.15



88-105

Compound	**R_1**	**R_2**	**R_3**	**R_4**	**R_5**	**R_6**	**pIC_{50}**
88	O	H	H	H	H	H	3.25
89	O	Br	H	H	H	H	3.67
90	O	H	Br	H	H	H	4.18
91	O	H	H	Br	H	H	4.13
92	O	H	H	H	Br	H	4.08
93	O	H	F	H	H	H	4.01
94	O	H	I	H	H	H	4.27
95	O	H	NO_2	H	H	H	3.88
96	O	H	OCH_3	H	H	H	3.38
97	O	H	Br	H	Br	H	4.98
98	O	H	Br	Br	H	H	4.94
99	O	H	I	H	I	H	5.11
100	O	H	Br	H	NO_2	H	3.59
101	O	H	NO_2	Br	H	H	4.77
102	O	H	Br	Br	Br	H	5.17
103	$N–C_6H_5$	H	H	H	H	H	4.17
104	$N–C_6H_5$	H	Br	H	Br	H	4.86
105	O	H	H	H	H	CH_3	3.62

Table 2. Brief description of some descriptors used in this study

Descriptor type	Molecular Description
Chemical	LogP (Octanol-water partition coefficient), Hydration Energy (HE), Polarizability (Pol), Molar refractivity (MR), Molecular volume (V), Molecular surface area (SA).
Constitutional	mean atomic vander Waals volume (MV), no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds (nBM), no. of aromatic bonds, no. of functional groups (hydroxyl, amine, aldehyde, carbonyl, nitro, nitroso, etc.), no. of rings, no. of circuits, no of H-bond donors, no of H-bond acceptors, no. of Nitrogen atoms (NN), chemical composition, sum of Kier-Hall electrotopological states (Ss), mean atomic polarizability (Mp), number of rotable bonds (RBN), mean atomic Sanderson electronegativity (Me), number of Chlorine atoms (NCl), number of 9-membered rings (NR09), etc.
Topological	Molecular size index, molecular connectivity indices (X1A, X4A, X2v, X1Av, X2Av, X3Av, X4Av), information content index (IC), Sum of topological distances between F..F (T(F..F)), Ratio of multiple path count to path counts (PCR), Mean information content vertex degree magnitude (IVDM), Eigenvalue sum of Z weighted distance matrix (SEigZ), reciprocal hyper-detour index (Rww), Eigenvalue coefficient sum from adjacency matrix (VEA1), radial centric information index, 2D petijean shape index (PJI2), mean information index on atomic composition(AAC), Kier symmetry index(S0K), mean information content on the distance degree equality (IDDE), structural information content (neighborhood symmetry of 3-order) (SIC3), Randic-type eigenvector-based index from adjacency matrix (VRA1), sum of topological distances between N..N (T(N..N)), sum of topological distances between O..O(T(O..O)),etc.
Geometrical	3D-Balaban index (J3D), span R (SPAN), length-to-breadth ratio by WHIM (L/BW), sum of geometrical distances between N..N (G(N..N)), sum of geometrical distances between N..O (G(N..O)), sum of geometrical distances between O..O (G(O..O)), ect.
Walk-Mol	molecular walk count of order 08 (MWC08), self-returning walk count of order 05 (SRW05), total walk count (TWC), etc.
Burden matrix	highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses (BEHM1), highest eigenvalue n. 7 of Burden matrix / weighted by atomic masses (BEHM7), lowest eigenvalue n. 1 of Burden matrix / weighted by atomic masses (BELM1), highest eigenvalue n. 1 of Burden matrix / weighted by atomic van der Waals volumes (BELV1), highest eigenvalue n. 2 of Burden matrix / weighted by atomic Sanderson electronegativities (BEHE2), etc.
Galvez	topological charge index of order 1 (GGI1), topological charge index of order 6 (GGI6),topological charge index of order 7 (GGI7), global topological charge index (JGT), etc.

2D autocorrelation	Broto-Moreau autocorrelation of a topological structure - lag 7 / weighted by atomic Sanderson electronegativities (ATS7E), Moran autocorrelation -lag 4 / weighted by atomic Sanderson electronegativities (MATS4E), Broto-Moreau autocorrelation of a topological structure - lag 3 / weighted by atomic Sanderson electronegativities (ATS3E), Broto-Moreau autocorrelation of a topological structure - lag 3 / weighted by atomic van der Waals volumes (ATS3V), etc.
Charge	maximum positive charge (QPOS), partial charge weighted topological electronic charge (PCWTE), etc.
Aromaticity	Harmonic Oscillator Model of Aromaticity HOMA index,RCI;Jug RC index aromaticity indices,HOMT;HOMA total (trial) , etc.
Randic	DP0;molecular profile, SP0;shape profile; SHP;average shape profile index , etc.
RDF	Radial Distribution Function - 7.0 / unweighted(RDF070U),Radial Distribution Function - 13.5 / unweighted(RDF135U),Radial Distribution Function - 1.0 / weighted by atomic masses(RDF010M),Radial Distribution Function - 3.0 / weighted by atomic masses(RDF030M),Radial Distribution Function - 4.5 / weighted by atomic masses(RDF045M),Radial Distribution Function - 12.5 / weighted by atomic masses(RFD125M),Radial Distribution Function - 2.0 / weighted by atomic van der Waals volumes(RDF020V),Radial Distribution Function - 8.5 / weighted by atomic van der Waals volumes(RDF085V),Radial Distribution Function - 1.0 / weighted by atomic Sanderson electronegativities(RDF010E), etc.
3D-MoRSE	3D-MoRSE - signal 01 / unweighted (MOR01U)(01U,02U,…,32U), 3D-MoRSE - signal 01 / weighted by atomic van der Waals volumes (MOR01V)(01V,02V,…,32V), ect.
WHIM	1st component symmetry directional WHIM index / weighted by atomic polarizabilities (G1P), 2st component symmetry directional WHIM index / weighted by atomic electrotopological states (G2S), D total accessibility index / weighted by atomic van der Waals volumes (DV), etc.
GETAWAY	H autocorrelation of lag 1 / lag2/ lag3 weighted by atomic Sanderson electronegativities (H1E,H2E,H3E), total information content on the leverage equality (ITH), R maximal autocorrelation of lag 3 / lag4 unweighted (R3U+,R4U+), R maximal autocorrelation of lag 6 / weighted by atomic masses (R6M+), R maximal autocorrelation of lag 5 / weighted by atomic van der Waals volumes (R5V+), R maximal autocorrelation of lag 1 / lag 4 weighted by atomic Sanderson electronegativities (R1E+), R maximal autocorrelation of lag 3 / weighted by atomic polarizabilities (R3P+), etc.
Functional	number of total secondary C(sp3) (NCS), number of ring tertiary C(sp3) (NCRHR), number of secondary C(sp2) (n=CHR), number of tertiary amines (aliphatic) (NNR2), number of N hydrazines (aromatic) (nN-NPH), number of nitriles (aliphatic) (NCN), number of phenols (NOHPH), number of ethers (aromatic) (NRORPH),

	number of solfures (NRSR), etc.
Atom-Centred	**CHR3 (C-003), CR4 (C-004), X--CR..X (C-034), Ar-C(=X)-R (C-039), R-C(=X)-X / R-C#X / X-=C=X (C-040), X--CH..X (C-042), H attached to C1(sp3) / C0(sp2) (H-047), RCO-N< / >N-X=X (N-072),R2S / RS-SR (S-107), etc.**
connectivity indices	**X0(connectivity index chi-0), connectivity index chi-1(x1), average connectivity index chi-0(XOA)**
information indices	**Uindex(Balaban U index), IC0(information content index), TIC0(total information content index)**
edge adjacency indices	**EEig01x(Eigenvalue 01),EEig01r(Eigenvalue 01 from edge)**
eigenvalue-based indices	**Eig1v(Leading eigenvalue from van der Waals weighted distance Eigenvalue sum from mass weighted distance matrix),SEigm matrixeigenvalue-based indices**

Table 3. The results of MLR analysis with different types of descriptors.

Eq.	Descriptors	Equation	R^2	F	Q^2	SE
1	Chemical	Y=2.734(±0.471)+0.006 (±0.001)SA1	0.59	29.02	0.51	0.23
2	constitutional	Y=4.505(±0.390)+3.092 (±0.533)nR05+2.552(±0.529)nR09 –0.698(±0.297)ns	0.54	17.22	0.49	0.23
3	Topological descriptors	Y=–21.196(±6.273)+0.011 (±0.002)DDr05+129.265(±34.518)X3A +3.174(±1.352)PJI2	0.56	14.15	0.49	0.13
4	Geometrical descriptors	Y=5.409(±0.396) + 0.001(±0.000)DDI–0.035(±0.015)G1	0.21	5.75	0.17	0.21
5	Fuctional group counts	Y=4.777(±0.181)–0.824 (±0.307Nnn+1.026(±0.475) nCHR	0.56	14.28	0.50	0.16
6	Charge descriptors	Y=4.097(±0.478)+0.052 (±0.023)PCWTe	0.35	11.20	0.28	0.17
7	Molecular walk counts	Y=4.142(±0.424)+0.076 (±0.030)SRW05	0.15	18,21	0.13	0.24
8	BCUT descriptors	Y=4.669(±0.931)–0.260 (±0.114)BEHm1+1.143(±0.534)BELm4	0.51	17.43	0.45	0.16
9	Galvz topol. Charge in dices	Y=7.557(±0.782)–23.375 (±7.436)JGI3	0.21	9.88	0.17	0.31
10	2D autocorrelations	Y=3.366(±0.757)+1.111 (±0.377)GATS3e+4.035(±1.204)MATS8e +1.621(±0.780)MATS5e	0.32	10.39	0.28	0.25
11	RDF descriptors	Y=4.901(±0.167)+0.031 (±0.012)RDF110u	0.17	6.86	0.14	0.25
12	3D MoRSE descriptors	Y=4.468(±0.245)–3.429 (±0.545)Mor17v+1.965 (±0.484)Mor28u–1.795 (±0.465)Mor27u–2.336 (±0.721)Mor24m+1.183 (±0.567)Mor24u	0.45	11.62	0.36	0.34

13	WHIM descriptors	Y=4.354(±0.343)+0.166(±0.066)L2u	0.11	6.30	0.09	0.35
14	GETAWAY descriptors	Y= -20.489(±12.001)–28.370(±6.263)R3v-A+4.019(±1.525)HATS2u+26.884(±12.154)ISH	0.25	7.37	0.17	0.21
15	Atom-centered	Y=4.481(±0.275)+0.070(±0.018)H047–0.432(±0.153)C028	0.21	8.19	0.17	0.27
16	Aromaticity	Y=5.171(±0.153)–0.333(±0.444)HOMT	0.14	0.57	0.11	0.16
17	Randic molecular	Y=4.046(±2.600)+0.150(±0.188)DP06–0.290(±3.420)SHP2	0.23	1.64	0.17	0.19

Table 4. Statistical parameters for testing prediction ability of the MLR, GA-PLS, PCR, and FA-MLR models

Model	R^2	R^2_{LOOCV}	RMSEcv	R^2p	RMSEp
MLR	0.65	0.59	0.23	0.80	**0.18**
GA-PLS	0.80	0.72	0.17	0.89	**0.25**
PCR	0.77	0.72	0.14	0.82	**0.21**
FA-MLR	0.66	0.61	0.18	0.71	**0.15**

R^2: Regression Coefficient for Calibration set
R^2_{LOOCV}: Regression Coefficient for Leave One Out Cross Validation
$RMSE_{cv}$: Root Mean Square Error of cross validation
R^2p: Regression Coefficient for prediction set
RMSEp: Root Mean Square Error of prediction set

Table 5. Numerical values of factor loading numbers 1–4 for descriptors after VARIMAX rotation

Descriptors	**Component**			**Extraction**
	F1	**F2**	**F3**	
SA1	**.650**	-.308	.190	.554
X3A	-.190	**.904**	-.031	.855
PJI2	.093	-.178	-.549	.342
JGI3	-.123	-.126	**.900**	.840
SHP2	-.456	.597	.596	.919
MOR28U	-.559	.550	.141	.634
MOR17V	**-.831**	-.022	.306	.785
nNN	**.830**	-.039	-.241	.748
C028	-.003	**-.898**	-.140	.826
H047	**.616**	-.433	-.476	.794
%variance	27.619	26.139	19.215	72.973

Table 6. The results of PCR analysis

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	R^2	SE	Q2	F
	B	Std. Error	Beta						
(Constant)	5.026	.096		52.150	.000	0.77	0.23	0.72	12.97
F1	.508	.097	.464	5.243	.000				
F2	.250	.097	.229	2.584	.011				
F3	-.211	.097	-.193	-2.180	.032				

Table 7. The results of FA-MLR analysis with different types of descriptors

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	R^2	F	Q^2	SE
	B	Std.Error	Beta						
(Constant)	-4.456	1.004		-3.354	.001	0.657	24.74	0.62	.32
nArNO2	-0.383	0.077	0.367	5.511	.000				
nR09	2.234	0.432	0.305	3.372	.001				
n COOH	5.417	1.643	0.178	2.080	.000				

Tabel 8. Leverage (*h*) of the external test set molecules for different models. The last row (*h**) is the warning leverage.

Molecule .no	MLR	GA-PLS	PCR	FA-MLR
17	0.09746	0.17875	0.08226	0.06075
18	0.06096	0.09862	0.04997	0.06450
19	0.05874	0.11962	0.03309	0.05698
20	0.07026	0.23594	0.06298	0.05577
29	0.02240	0.08769	0.03078	0.05129
30	0.02604	0.09271	0.03261	0.04644
32	0.02844	0.06821	0.02755	0.05590
34	0.09571	0.08710	0.03276	0.04359
35	0.08727	0.08929	0.03351	0.04380
39	0.04394	0.06628	0.02476	0.04393
44	0.06903	0.10634	0.02965	0.04266
45	0.07313	0.10575	0.02842	0.04241
51	0.06198	0.12216	0.02116	0.04210
54	0.09509	0.12957	0.04365	0.02684
56	0.04910	0.09905	0.04313	0.02418
57	0.05542	0.10247	0.04563	0.02332

58	0.05739	0.10751	0.04836	0.02326
60	0.04911	0.08442	0.04949	0.02545
62	0.07641	0.09106	0.06323	0.03819
80	0.04640	0.12240	0.00384	0.04589
85	0.04854	0.16624	0.08482	0.02662
87	0.06481	0.11755	0.02738	0.02128
93	0.04033	0.17742	0.05823	0.03038
h*	0.25714	0.42857	0.12857	0.21429