Original Research Article

QSAR and docking study of isatin analogues As cytotoxic agents

Abstract:

Computational chemistry is a unique method in the drug discovery process?? Explain Why?. In this study 109 molecules containing the isatin backbone were subjected to quantitative structure-activity relationship analysis to find the structure requirements for ligand binding. The structures were sketched and optimized in Hyperchem. The structural invariants used in this study were those obtained from whole molecular structures: by both hyperchem and dragon software (16 types of descriptors). Four chemometrics methods including MLR, FA-MLR, PCR and GA-PLS were employed to make connections between structural parameters and anticancer effects. MLR models revealed the effects of constitutional, functional, geometrical, WHIM and GETAWAY descriptors having higher impact on anticancer activity of the compounds. GA-PLS showed functional, constitutional and chemical descriptor indices to be the most significant parameters on anticancer activity. Moreover, the result of FA-MLR analysis revealed the effects of functional descriptors on the anticancer activity. A comparison between the different statistical methods employed and the results indicated that GA-PLS represented superior results and could explain and predict 81% and 78% variances in the PIC_{50} data, respectively. Docking studies of these compounds were also investigated and promising results were obtained showing that some compounds were introduced as a good candidate for cancer agents.

Introduction

The isatin (1*H*-indole-2,3-dione) derivatives show a broad spectrum of biological activities such as antibacterial, antifungal, antiviral and anticancer drug candidates in many synthetic compounds [1–5]. Among these properties antineoplastic activities of these moieties were of our interest to study the quantitative structure-activity relationships of a series of 109 isatin derivatives reported in literature.

Synthesis and evaluation of the biological activity of these novel compounds are usually time-consuming to make and is expensive.Hence the use of computational

techniques for designing biologically active compounds has opened a new window to drug discovery research. Computational methods can accelerate the procedure of discovering new drugs by designing new compounds and predicting activity of newly synthesised or even non-synthesized compounds. Quantitative structure activity relationships (QSAR) studies, is one of the most important subjects in chemometrics andplays an important role in predicting activity of novel compounds [6-10]. Linear QSAR models are mathematical equations that present us with good 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 [11-14]. These features named molecular descriptors have high impact on the biological activity of the compounds [15-18]. 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 which provide us more than 4000 of these descriptors [19,20].

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 the anti-cancer activity of compounds [21-24]. An important approach of the researchers in modifying 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 is associated with increased biological activity [25].The *in vitro* cytotoxic activities of isatin bromo-derivatives were determined against the human monocyte-like, histiocytic lymphoma cell line (U937), showing that the introduction of electron withdrawing groups at positions C5, C6, and C7 significantly increased the cytotoxic activity when compared with isatin molecules with the substitution at the 5-position being the best [26]. Introduction of an aromatic ring with one or three carbon atom linker at N_1 enhances the activity too [27]. In 2006, an isatin 5-fluoro-derivative (Sunitinib) was approved by FDA for the treatment of gastrointestinal tumours and advanced renal cell carcinoma [28,29]. Isatin bromo-derivatives have been shown to exhibit anticancer activity [30-32]. 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 a mathematical relationship between biological activities and computable parameters such as topological, quantum, physicochemical, stereo chemical or electronic indices. The QSAR study of halogenated isatin analogues was reported by Sabet *et al* [33] and showed that topological, chemical, geometrical and functional group were effective on the cytotoxic activity. QSAR analysis of novel N-alkyl substituted isatin derivatives were identified by RajK.Prasad *et al* [34] by using different multiple regression approach. Three-dimensional quantitative structure–activity relationship (3D-QSAR) and docking methods of isatin derivatives with anticancer activity against human monocyte-like histiocytic lymphoma human U937 cells was reported by Elidrissi B [35].

The molecular docking study helps us to understand the various interactions between the ligands and enzyme active sites in detail and also help to design novel potent inhibitors. Molecular docking simulation techniques were also performed on one-hundred and nine compounds to investigate the molecular binding models for these compounds interacting with the key active site of protein.

2.Methods

2.1. 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} [19]. 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 [20]. The calculated descriptors from whole molecular structures are briefly described in Table 2.

2.2. 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 the model's stability and prediction ability. We used cross-validation and external test set for these molecules. 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 [36]. 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 which exist 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 [37].

2.3. Docking procedures

An in house batch script (DOCK-FACE) for automatic running of AutoDock 4.2 was used to carry out the docking simulations [38] in a parallel mode [39]. To prepare the receptor structure, the three dimensional crystal structure of Caspase-3 inhibitory activity (PDB ID: 1GFW) was acquired from Protein Data Bank (PDB data base; http://www.rcsb.org) [40] and water molecules and co-crystal ligands were removed from the structure. The PDB were then checked for missing atom types with the python script as implemented in MODELLER 9.17 [41]. The ligand structures were made by Hyper Chem software package (Version 7, Hypercube Inc). For geometry optimization, Molecular Mechanic (MM^{+}), followed by semi empirical AM1 method was performed. The prepared Ligands were given to 100 independent genetic algorithm (GA) runs. 150 population size, a maximum number of 2,500,000 energy evaluations and 27,000 maximum generations were used for Lamarckian GA method. The grid points of 80, 80, and 80 in x-, y-, and z directions 38, 34 and 23 were used. Number of points in x, y and z were used respectively. All visualization of protein ligand interaction was evaluated using VMD software [42]. Cluster analysis was performed on the docked results using a root mean square deviation (RMSD) tolerance of 1.98 Å.

3. Results and discussion

3.1. Data set

The biological data used in this study was the anti-cancer activity against MCF7, (in terms of -log IC_{50}), of a set of 109 isatin derivatives [43-51]. The data set was classified into calibration and prediction set by kenardston algorithm of the 20 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]

3.2. 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 (89 molecules as calibration and 20 molecules as prediction sets) are listed in Table 3.

[Table 3. near here]

The equation E1 of Table 3 shows among chemical descriptors, the negative effect of surface area of the molecules on cytotoxicity which shows the positive effect of log p of the molecules on the activity. This equation indicates the hydrophilic molecules show better cytotoxic effect. The second equation of Table 3 demonstrated the effect of constitutional descriptors on the anti-cancer activity of these compounds. It shows that increasing the number of halogen atoms (nX, nF, nCl, nBr) of the compounds results in an activity enhancement, such as the molecular series 1-18, 89-109. It also shows that the halogen substitution is better on the 5 or 7 position of the isatin ring. If the substitution was Br, it gave the better the activity, confirming the E1 of this table because Br undergoes lipophilic substitution. It also explain the positive effect of nDB (number of double bonds), nCIC (number of rings), and nR09 (number of 9-membered rings) such as the indol ring on activity (such as molecule series 19-24 and 25-30 have good activity).

The effect of the topological group count parameter on anti-cancer activity of the studied compounds has been described by equation E_3 of Table 3. It shows that among the topological descriptors, the structural information content (SIC2) and spanning tree number (STN) have the positive effects on cytotoxic activity of the compounds.

The equation E_4 of Table 3 was found by using Mol-Walk descriptors (E_4), which explains the positive effect of MWC03 index (molecular walk count of order 03) and negative effect of MWC10 (molecular walk count of order 10) and PIPC09 (molecular multiple path count of order 09) of the studied compounds on the anti-cancer activity. It can explain and predict more than 61% of variances in the biological activity data. The equation E_5-E_{14} and E_{16} 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 MLR equation of Table 3 obtained from the pool of functional group descriptors, E_{15}, explained the positive effect of the n oxim (number of oxim substitution) , n pyridine (number of pyridine substitution), n isothiocyanate and n thiocyanate

(number of isothiocyanate and thiocyanate substitution) (such as molecules of 25-30, 78, and 79) on the anti-cancer activity. The nC=S (number of C=S substitution) , $nArNO_2$ (number of aromatic nitro groups), n oxazole (number of oxazole substitution), nThiazol (number of thiazole substitution), nCOOH (number of COOH group), nCOOCH (number of ester group) (molecules series 33-34, 55-56, 74- 76 and 77-84) have negative effects on the anti-cancer activity. The negative sign of this group proposed that a decrease in the number of these descriptors resulted in an activity enhancement. This equation, has a high statistical quality (R^2 = 0.77, Q^2 = 0.72).

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 0.74, which means that the resulted QSAR model could predict 74% of variances in the anti-cancer activity data. It has root mean square error of 0.21.

3.3. 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 = 109) was divided into two groups: calibration set (n = 89) and prediction set (n = 20). Given 89 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 8 descriptors including, three constitutional descriptor (nR09, nC=S, nX) and one chemical (logp) parameter and

four functional descriptors (n isothiocyanate, nCOOH, npyridine, $nArNO_2$). The majority of these descriptors are functional indices, 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 statistic, but also represents a high ability for modeling external test samples. It could explain and predict about 78% 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 8 selected PLS descriptors in the protein tyrosine kinase inhibitory activity it was important to investigate the relative importance of the variable which appeared in the final model obtained by GA-PLS method, variable important in projection (VIP) was employed [52]. VIP values reflect the importance of terms in the 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 signifying highly or less influential, respectively, and 0.8 < VIP< 1.0 meaning moderately influential. The VIP analysis of PLS equation is shown in Figure 2. As it is observed, logp, nCOOH and nR09 indices represent the most significant contribution in the resulted QSAR model. In addition, functional group parameter such as nC=S, n isothiocyanate and $nArNO_2$ have been found to be moderately influential parameters.

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

3.4. 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 [53]. 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 [54]. In this method, factor scores, as obtained from FA, are used as the predictor variables [53]. 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 82% 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).

$Y= -4.456(\pm1.004) -0.383(\pm0.077)\ nArNO_2+2.234(\pm0.432)\ nR09+ 5.417(\pm1.643)\ n\ COOH$

$R^2 = 0.657\ \ S.E = 0.32\ \ F = 24.74\ \ \ Q^2 = 0.62\ \ RMScv = 0.15$

This equation could explain about 65.7% of the variance and predict 62% of the variance in pIC_{50} data. It has a root mean square error of 0.18. This equation describes the effect of functional descriptors ($nArNO_2$, nR09 and n COOH) 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):

$Y= 4.742(\pm0.043) +.654(\pm0.043)\ F1 +0.756\ (\pm0.043)\ F6 - 0.456(\pm0.043)\ F3 +.321(\pm0.043)\ F2$

$R^2 = 0.73\ \ S.E. = 0.23\ \ F = 15.54\ \ \ Q^2 = 0.70\ \ RMScv = 0.18$

This equation could explain and predict 73% and 70% of the variances in pIC_{50} data, respectively. The root mean square error of PCRA analysis was 0.18. 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 more 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 autocorrelation indices whereas information about the Connectivity indices, 3D WHIM, MoRSE descriptors and Functional descriptors are highly incorporated in factor 3 and 4. Factor score 5, 6 and 7 signify the importance of GETAWAY و 2D autocorrelations, Functional and Atom-center descriptors.

3.5. Robustness and applicability domain of the models

Leverage is one of the 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 [55]. 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]

3.6. Molecular Docking Studies

The docking study was performed using the AutoDock 4.2. All the one-hundred and nine isatin derivatives were docked into the active site of the enzymes Caspase-3 inhibitory (PDBID:1GFW) (How did you choose this enzyme?). All the docking protocols were done on validated structures, with RMSD values below 2 Å. The conformation with the lowest ones was considered as the best docking result. Docking binding energies of these active compounds were summarized in Table 1. Our results indicated that 23 compounds, number 38-49 and 66-76 showed better docking scores than corresponding co-crystal ligands. These compounds could be considered as possible hits as cancer agents. Compounds having two indolin rings with electron withdrawing groups at C-5 and C-7 position showed good docking scores. In general,

increase in the number of the ring especially indolin ring and substitutions in C-5 and C-7 such as halogen and ester on indolin moieties can cause better interaction with the receptor. The interaction modes of 39,46 and 68-69 those with the best docking scores are shown in Figure 3. Binding interaction of 4 compounds are presented in Table 9. The NH and oxygen atom which exist in carbonyl group of indolin of ligand 39 had H-bonding with Gly 122 and His 121 at receptor site, also NH atom of pyrrole ring had H-bonding with Cys 163 and indolin ring showed Arene-Arene interaction with Phe 256 at distance 3.65A^0. At 46 compound, exist H-bond between carbonyl group of indolin and Arg207, also NH group of chain formed H-bond with Phe 250 at distance 2.90 A^0. At 68 compound, NH and carbonyl group of indolin and NH group of benzimidazole had H-bonding with Glu 248, Phe250, Ser 249 amino acid in order side, the chlorine atom in position 5 of indolin showed hydrophobic interaction with Gln 217 at distance 3.26 A^0 and also benzene thiol ring formed Arene-Arene interaction with Trp 206 at distance 3.76A^0. at 69 compound exist five H-bond between NH, carbonyl group of indolin and NH group of chain with Trp 214, Asn 208, Ser 209, Arg 207 and Phe 250 respectively.

[Table 9near here], [Figure 3 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 n oxim, n pyridine, n isothiocyanate, n thiocyanate on the anti-cancer activity and it also indicate the nC=S, $nArNO_2$, n oxazole, nThiazol, nCOOH, nCOOCH have negative effects on activity. GA-PLS analysis indicated that three constitutional descriptor (nR09, nC=s, nX) and one chemical (log p) indices and four functional descriptors (n isothiocyanate, nCOOH, npyridine, $nArNO_2$ parameters were the most significant parameters on cytotoxicity activity of studied compound. The FA-MLR describes the effect of functional descriptors ($nArNO_2$, nR09 and n COOH 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 autocorrelation indices whereas information about the connectivity indices, 3D WHIM, MoRSE descriptors and functional descriptors are highly incorporated in factor 3 and 4 Factor score 5, 6 and 7 signify the importance of GETAWAY و 2D autocorrelations, functional and

atom-center descriptors. A comparison between the different statistical methods employed revealed that GA-PLS represented superior results and it could explain and predict 81% and 78% of variances in the pIC_{50} data, respectively. As docking studies revealed, 23 compounds, number 38-49 and 66-76 are introduced as good candidates for cancer agents and the docking results show that increase in number of the ring especially indolin ring and substitutions such as halogen and ester at C-5 and C-7 on indolin moieties can cause better interaction with the receptor.

References:

1. S.N. Pandeya, S. Smitha, M. Jyoti, S.K. Sridhar, Acta Pharm. 55, (2005) 27–46.
2. 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.
3. 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.
4. 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.
5. A. Gursoy, N. Karali, Eur. J. Med. Chem. 38, (2003) 633–643.
6. H. Schmidi, Multivariate prediction for QSAR, Chemom. Intell. Lab. Syst. 37 (1997) 125-134.
7. 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.
8. S. Wold, J. Trygg, A. Berglund, H. Antii, Some recent developments in PLS modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131-150.
9. R. Sabet, A. Fassihi, B. Hemmateenejad, L. Saghaie, R. Miri, L. Gholami, Computer-aided drug design of novel antibacterial 3-hydroxypyridine-4-ones: application of QSAR methods based on the MOLMAP approch. J Comput Aid Mol Des. 26 (2012) 349-361.

10. R. Sabet, A. Fassihi, B. Moeinifard, QSAR study of PETT Derivatives as Potent HIV-Reverse Transcriptase Inhibitors. J. Mol. Graph & Model. 28 (2009) 146-155.
11. 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.
12. 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.
13. C. Hansch, D. Hoekman, H. Gao, Comparative QSAR: Toward a deeper understanding of chemicobiological interactions, Chem. Rev. 96 (1996) 1045-1075.
14. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, 2000.
15. 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.
16. S. Putta, J. Eksterowicz, C. Lemmen, R. Stanton, A novel subshape molecular descriptor, J. Chem. Inf. Comput. Sci. 43 (2003) 1623-1635.
17. 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.
18. 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.
19. HyperChem, Release 8.0 for Windows, Molecular Modeling System: HyperCube.
20. Todeschini, R. Milano Chemometrics and QSAR Group. http://michem.disat.unimib.it/.
21. A. Fassihi, R. Sabet. QSAR Study of $p56^{lck}$ Protein Tyrosine Kinase Inhibitory Activity of Flavonoid Derivatives Using MLR and GA-PLS. Int. J. Mol. Sci. 9 (2008) 1876-1892.

22. R. Sabet, A. Fassihi. QSAR Study of Antimicrobial 3-Hydroxypyridin-4-one and 3-Hydroxypyran-4-one Derivatives Using Different Chemometric Tools. Int. J. Mol. Sci. 9 (2008) 2407-2423.
23. A. Fassihi, D. Abedi, L. Saghaie, R. Sabet, H. Fazeli, Gh. Bostaki, O. Deilami, H. Sadinpour. Synthesis, Antimicrobial Evaluation and QSAR Study of Some 3-hydroxypyridine-4- one and 3-hydroxypyran-4-one Derivatives. Eur. J. Med. Chem. 44 (2009) 2145-2157.
24. V. Consonni, R. Todeschini, M. Pavan, J. Chem. Inf. Comput. Sci. 42 (2002) 693-705.
25. K.L. Vine, J.M. Locke, M. Ranson, K. Benkendorff, S.G. Pyne, J.B. Bremner, Bioorg. Med. Chem. 15, (2007) 931–938.
26. K.L. Vine, J.M. Locke, M. Ranson, S.G. Pyne, J.B. Bremner, Bioorg. Med. Chem. 15(2007) 931.
27. K.L. Vine, J.M. Locke, M. Ranson, S.G. Pyne, J.B. Bremner, J. Med. Chem. 50, (2007) 5109–5117.
28. K. Kumar, S. Sagar, L. Esau, M. Kaur, V. Kumar, Eur. J. Med. Chem. 58 (2012) 153.
29. R. Roskoko Jr., Biochem. Biophys. Res. Commun. 356 (2007) 323.
30. R. Sabet, M. Mohammadpour, A. Sadeghi, A. Fassihi, Eur. J. Med. Chem. 45 (2010)1113.
31. K.L. Vine, J.M. Locke, M. Ranson, S.G. Pyne, J.B. Bremner, Bioorg. Med. Chem. 15(2007) 931.
32. K.L. Vine, L. Matesic, J.M. Locke, M. Ranson, D. Skropeta, Anti-Cancer Agents Med.Chem. 9 (2009) 397.
33. R. Sabet, M. Mohammadpour, A. Sadeghi, A. Fassihi. QSAR study of isatin analogues as in vitro anti-cancer agents. Eur J Med Chem. 45 (2010) 1113–1118.
34. Raj K. Prasada, T. Narsinghanib, R. Sharmab, QSAR analysis of novel N-alkyl substituted isatins derivatives as anticancer agents. J Chem Pharm Res, 1 (2009) 199-206.
35. B. Elidrissi, A. Ousaa A. Aouidate, H. Zaki, MA. Ajana, T. Lakhlifi and M. Bouachrine. 3D-QSAR Studies of Isatin Derivatives with Anti-Cancer in

Vitro: Advanced CoMFA, CoMSIA and Docking Methods. Chem Sci J. 8(2) (2017) 158.

36. W. Siedlecki, J. Sklansky. On Automatic Feature Selection. Int. J. Pattern Recog. Artif. Intell., 2 (1988) 197-220.
37. R. Leardi, Genetic Algorithms in Chemometrics and Chemistry: A Review. J. Chemometrics. 15 (2001) 559-569.
38. R. Sabet, A. Fassihi, L. Saghaie. Octanol-water partition coefficients determination and QSPR study of some 3-hydroxy pyridine-4-one derivatives, J Pharm Res Int. 22 (2018) 1-15.
39. W. Humphrey, A. Dalke, K. Schulten. VMD: visual molecular dynamics. J Mol Graph Model. 14 (1) (1996) 33-41.
40. M. Fereidoonnezhad, Z. Faghih, A. Mojaddami, A. Sakhteman, Z. Rezaei. A Comparative Docking Studies of Dichloroacetate Analogues on Four Isozymes of Pyruvate Dehydrogenase Kinase in Humans. Indian J Pharm Educ. 50(2) (2016) S32-S8.
41. BF. Mirjalili, L. Zamani, K. Zomorodian, S. Khabnadideh, Z. Haghighijoo . Z. Malakotikhah, et al. Synthesis, antifungal activity and docking study of 2-amino-4H-benzochromene-3-carbonitrile derivatives. J. Mol. Struct. 1116 (2016) 102-110.
42. Z. Li, J. Gu. H. Zhuang, L. Kang, X. Zhao, Q. Guo. Adaptive molecular docking method based on information entropy genetic algorithm. Applied Soft Computing. 26 (2015) 299-302.
43. S. Reddy, R. Pallela, D. Kim, M. Won, Y. Shim. Synthesis and Evaluation of the Cytotoxic Activities of Some Isatin Derivatives. Chem Pharm Bull. 61(11) (2013) 1105–1113.
44. N. Evdokimov, I. Magedov, D. McBrayer, A. Kornienko. Isatin derivatives with activity against apoptosis-resistant cancer cells. Bioorg Med Chem Lett. 26 (6) (2016) 1558-1560.
45. HS. Ibrahim, SM. Abou-seri, NS. Ismail, NM. Elaasser, MH. Aly, HA. Abdel-Aziz. Bis-isatin hydrazones with novel linkers: Synthesis and biological evaluation as cytotoxic agents. Eur J Med Chem. 108 (2016) 415-422.

46. Ö. Akgül, AH. Tarıkoğulları, F. Aydın Köse, P. Kırmızıbayrak, M. Pabuççuoğlu. Synthesis and cytotoxic activity of some 2-(2,3-dioxo-2,3-dihydro-1H-indol-1-yl)acetamide derivatives.Turkish J Chem. 37 (2013) 204 – 212.

47. K. Vine, J. Locke, M. Ranson, S. Pyne, J. Bremner. In vitro cytotoxicity evaluation of some substituted isatin derivatives. Bioorg Med Chem. 15 (2) (2007) 931–938.

48. KB. Priyanka, C. Manasa, G. Sammaiah. Synthesis and evaluation of new isatin derivatives for cytotoxic activity. J Pharm Pharm Sci. 3 (2) (2013) 2393-2402.

49. G. Krishnegowda, AP. Gowda, HR. Tagaram O. Staveley, RB. Irby, AK. Sharma, S. Amin. Synthesis and biological evaluation of a novel class of isatin analogs as dual inhibitors of tubulin polymerization and Akt pathway. Bioorg Med Chem. 19 (20) (2011) 6006-6014.

50. M. Farooq, Z.M. Almarhoon, NA. Taha, AA. Baabbad, MA. Al-Wadaan, A. El-Faham. Synthesis of novel class of N-alkyl-isatin-3-iminobenzoic acid derivatives and their biological activity in zebrafish embryos and Human cancer cell lines. Biol Pharm Bull. (2018) b17-00674.

51. K. Beckman, Isatin Derivatives as Inhibitors of Microtubule Assembly [Thesis]. Kansas: University of Kansas; 2008.

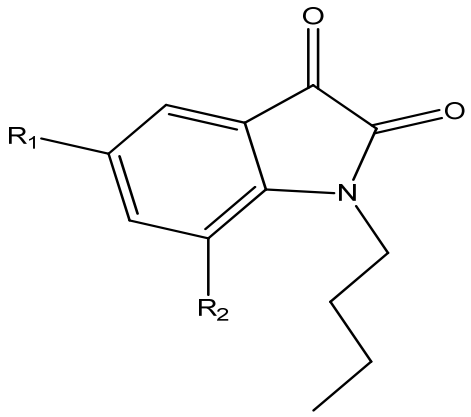
52. M. Olah, C. Bologa, T.I. Oprea. An Automated PLS Search for Biologically Relevant QSAR Descriptors. J. Comput. Aided Mol. Des. 18 (2004) 437-449.

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

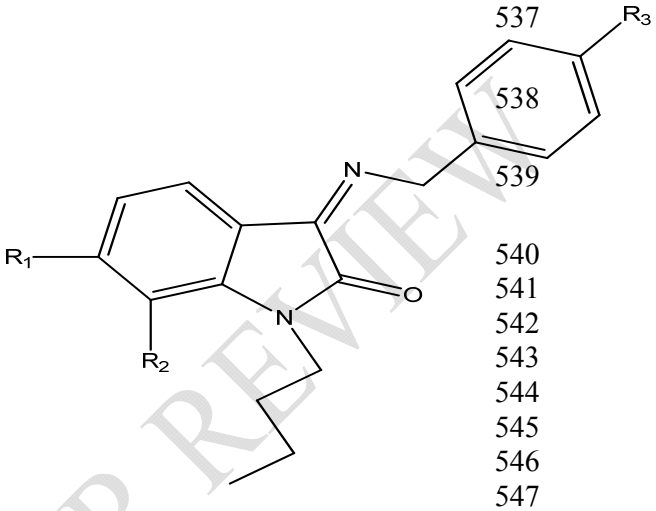
54. H. Kubinyi, The quantitative analysis of structure-activity relationships, in: M.E. Wolff, (Ed.), Burger's Medicinal Chemistry and Drug Discovery, 5th Ed.; Wiley, New York, 1 (1995) 506-509.

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

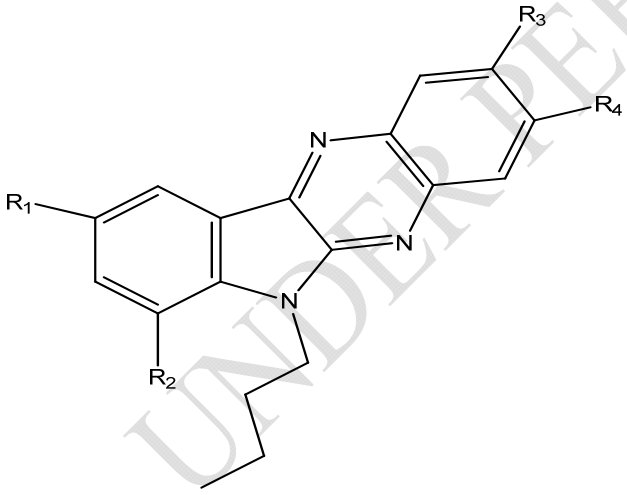
Table1 1. Chemical structure of isatin derivatives used in this study



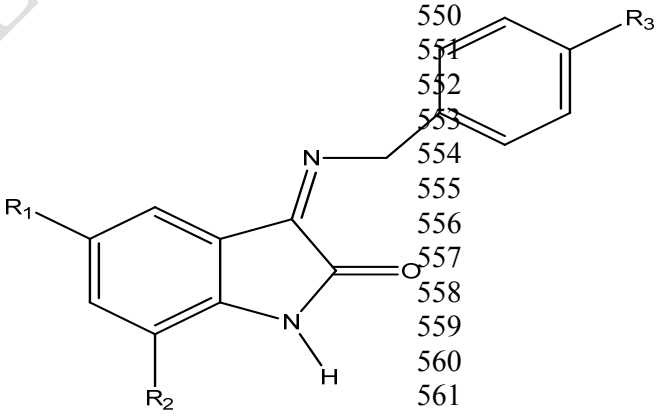
1-3



4-6



7-12

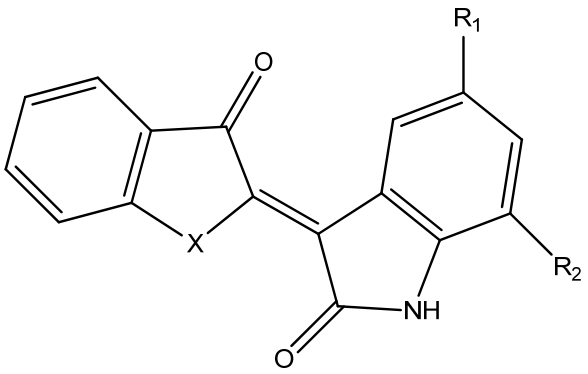


13-18

574

Compound	R_1	R_2	R_3	R_4	PIC_{50}	Binding Energy (kcal/mol)
1	Cl	H	-	-	4.16	-6
2	H	Cl	-	-	4.12	-6.4
3	H	F	-	-	4.16	-6.4
4	Cl	H	OCH_3	-	4.50	-6.9
5	H	Cl	OCH_3	-	4.76	-6.9
6	H	F	OCH_3	-	4.10	-6.9
7	Cl	H	CH_3	CH_3	4.42	-7.4
8	H	Cl	CH_3	CH_3	4.49	-7.3
9	H	F	CH_3	CH_3	4.14	-7.5
10	Cl	H	Cl	Cl	4.47	-7.3
11	F	H	Cl	Cl	4.08	-7.2
12	H	F	Cl	Cl	4.61	-7.2
13	Cl	H	OCH_3	-	4.50	-6.8
14	H	Cl	OCH_3	-	4.48	-6.8
15	F	H	OCH_3	-	4.24	-6.8
16	H	F	OCH_3	-	4.10	-6.8
17	H	Cl	H	-	5.28	-7
18	F	H	H	-	4.30	-6.9

575
576
577
578
579
580
581
582



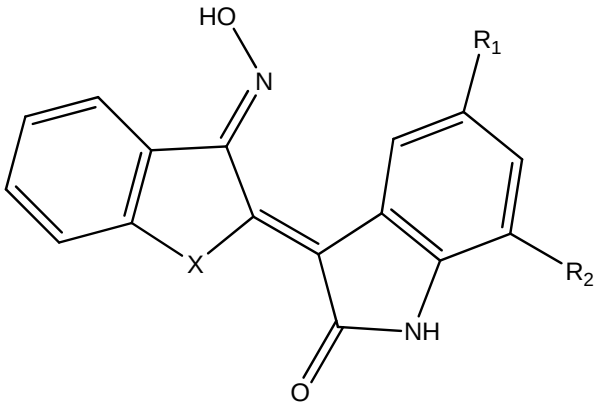
19-24

583

584

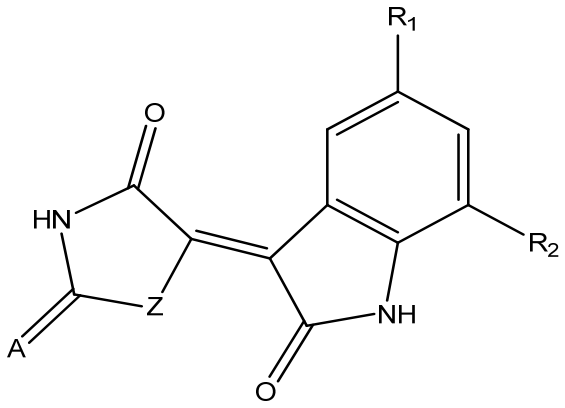
Compound	R_1	R_2	X	PIC_{50}	Binding Energy (kcal/mol)
19	Br	H	NH	4.43	-7.6
20	H	F	NH	4.35	-7.5
21	H	Br	NH	4.28	-7.6
22	H	H	CH_2	4.15	-8.1
23	Br	H	CH_2	4.19	-7.9
24	H	H	O	6.52	-7.7

585
586
587
588
589
590
591
592
593
594



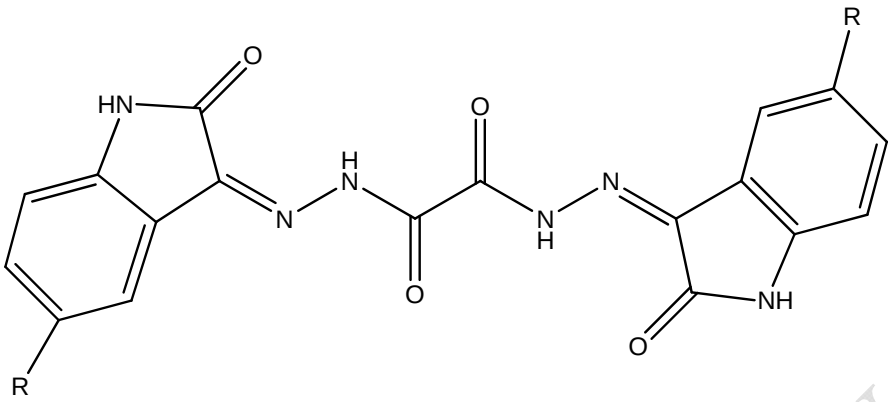
25-30

Compound	R_1	R_2	X	PIC_{50}	Binding Energy (kcal/mol)
25	H	H	NH	5.04	-8.1
26	Br	H	NH	5.24	-8.2
27	H	F	NH	4.58	-8.3
28	H	Cl	NH	4.56	-7.8
29	H	Br	NH	5.31	-7.6
30	H	H	CH_2	4.41	-8.2



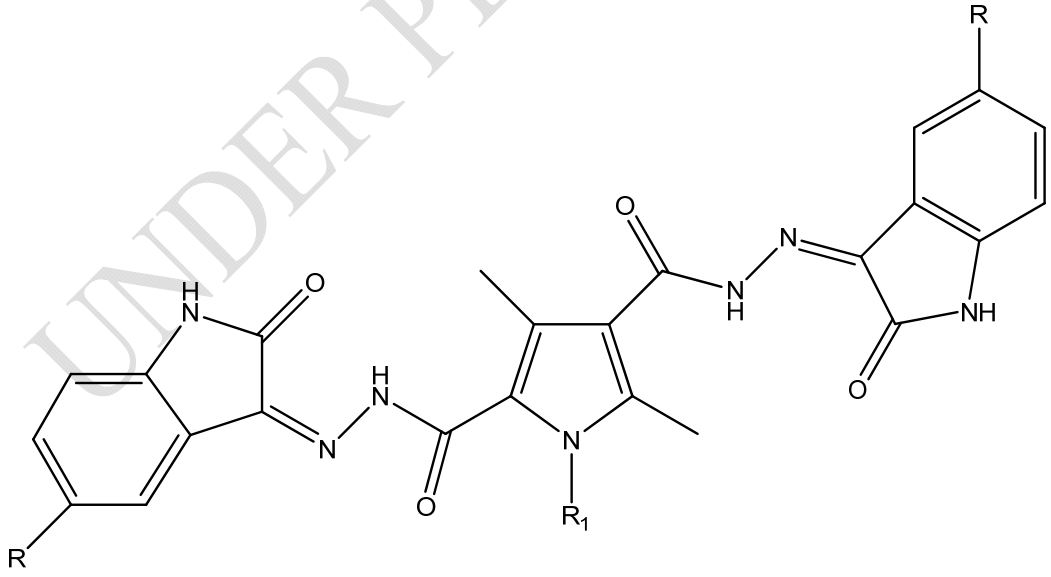
31-34

Compound	R_1	R_2	A	Z	PIC_{50}	Binding Energy (kcal/mol)
31	H	H	O	NH	4.02	-7.5
32	H	H	S	NH	4.06	-7.5
33	H	Br	S	NH	4.29	-6.7
34	Br	H	S	S	4.08	-7.6



35-37

Compound	R	PIC_{50}	Binding Energy (kcal/mol)
35	Br	4.04	-8.4
36	NO_2	4.04	-8.2
37	CH_3	4.25	-8.4

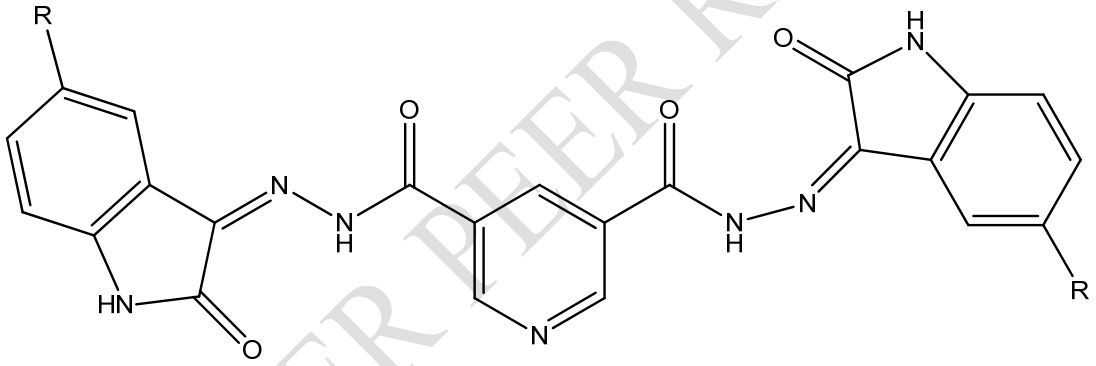


38-44

642

Compound	R	R_1	PIC_{50}	Binding Energy (kcal/mol)
38	H	H	4.16	-9.9
39	F	H	4.12	-10.2
40	Br	H	4.44	-9.3
41	CH_3	H	4.34	-9.5
42	OCH_3	H	4.10	-9.3
43	CH_3	CH_3	4.52	-9.4
44	OCH_3	CH_3	5.74	-8.9

643
644
645



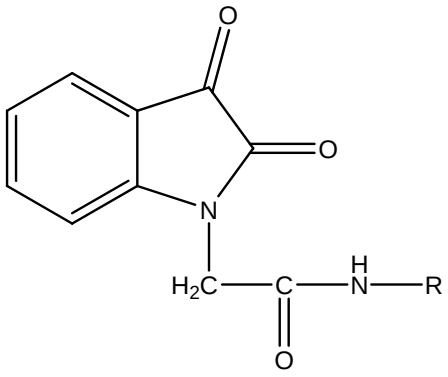
45-49

646
647
648
649
650

Compound	R	PIC_{50}	Binding Energy (kcal/mol)
45	H	4.41	-9.8
46	F	4.42	-9.9
47	Br	4.46	-9
48	NO_2	4.05	-8.6

49	OCH_3	4.18	-9.4

651

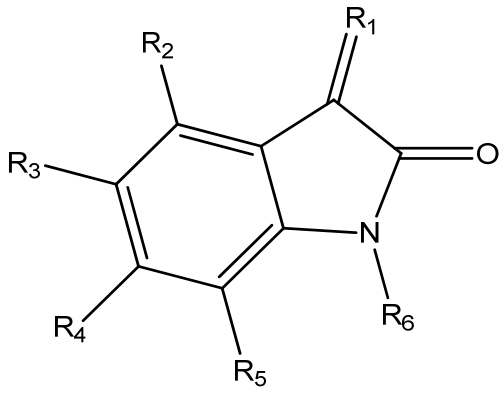


50-61

652
653
654

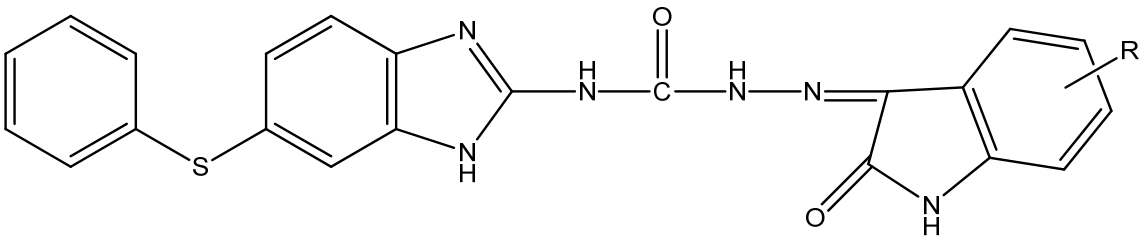
Compound	R	PIC_{50}	Binding Energy (kcal/mol)
50	4-methylphenyl	4.06	-8.1
51	2-methoxyphenyl	4.96	-7.8
52	4-methoxyphenyl	4.07	-7.8
53	2-chlorophenyl	4.49	-7.9
54	3-chlorophenyl	4.21	-7.9
55	2-nitrophenyl	4.96	-8.2
56	4-nitrophenyl	4.17	-8.1
57	2-ethylphenyl	4.31	-8
58	2-isopropylphenyl	4.74	-7.9
59	2,6-dimethylphenyl	4.19	-8.4
60	2,6-dichlorophenyl	4.22	-8
61	benzyl	4.33	-8.3

655
656



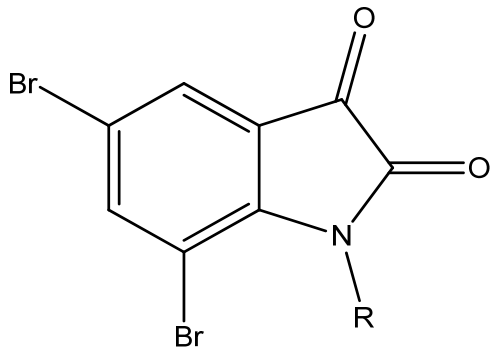
62-65 657

Compound	R_1	R_2	R_3	R_4	R_5	R_6	PIC_{50}	Binding Energy (kcal/mol)
62	O	H	Br	H	Br	H	4.50	-5.4
63	O	H	Br	Br	H	H	4.69	-5.6
64	O	H	I	H	I	H	4.74	-5.4
65	O	H	Br	Br	Br	H	4.88	-6

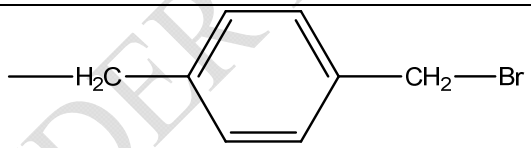
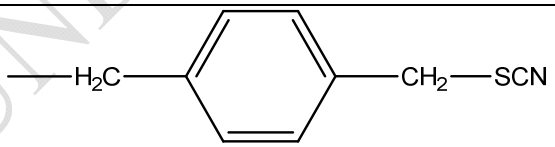
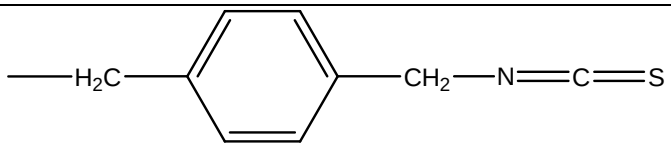


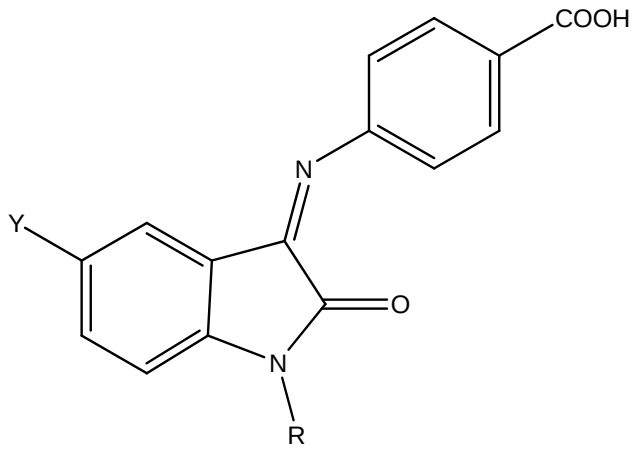
66-76

Compound	R	PIC_{50}	Binding Energy (kcal/mol)
66	H	4.64	-9.8
67	5-F	4.65	-9.9
68	5-Cl	4.63	-10
69	7-Cl	4.71	-10.1
70	5-Br	4.72	-9.7
71	6-Br	4.34	-9.5
72	5-NO_2	4.47	-9.6
73	7-NO_2	4.39	-9.7
74	5-COOH	4.35	-10
75	5-$COOCH_3$	4.28	-9.8
76	7-$COOCH_3$	4.32	-9.6



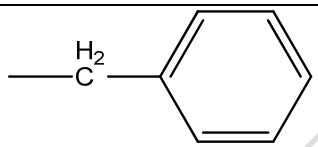
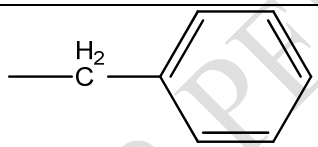
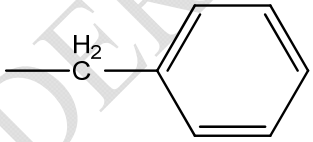
77-84

Compound	R	PIC_{50}	Binding Energy (kcal/mol)
77	$-(CH_2)_3$-Cl	4.67	-5.9
78	$-(CH_2)_3$-SCN	5.01	-5.7
79	$-(CH_2)_3$-N=C=S	5.05	-5.8
80	$-(CH_2)_4$-Cl	4.83	-5.9
81	$-(CH_2)_4$-SCN	4.66	-5.8
82		4.56	-6.9
83		4.61	-6.9
84		4.92	-6.8

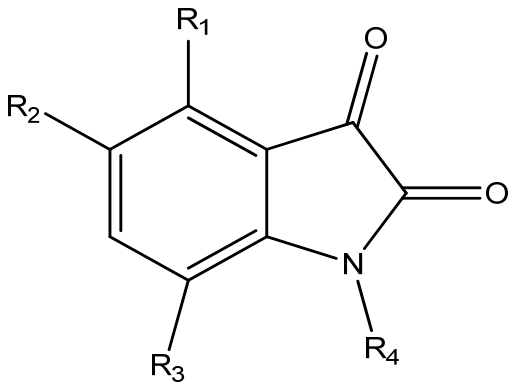


85-88

671
672

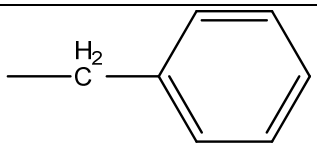
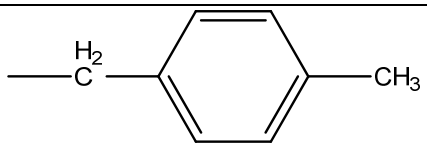
Compound	R	Y	PIC_{50}	Binding Energy (kcal/mol)
85	CH_3	H	4.18	-7.4
86		H	4.60	-8.2
87		Cl	4.63	-8
88		F	4.46	-8.2

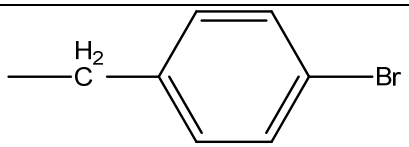
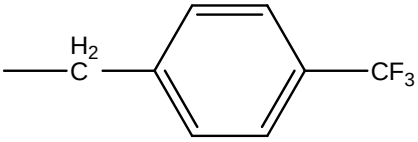
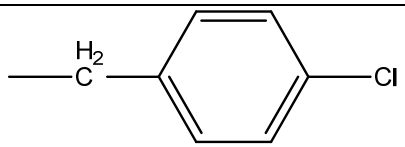
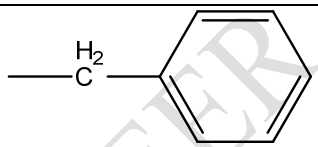
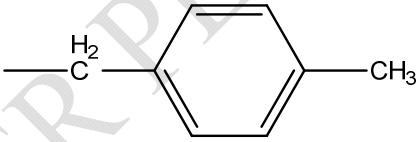
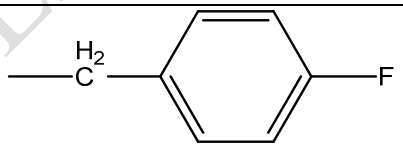
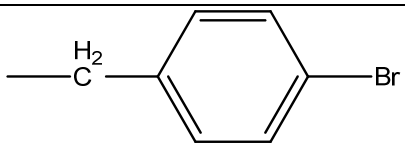
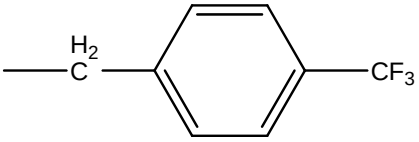
673

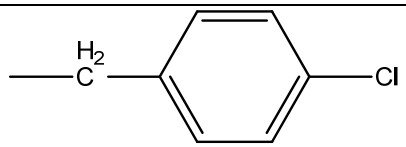
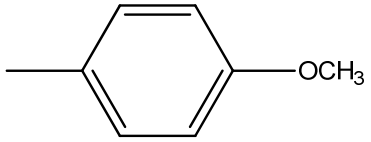


89-109

674
675
676

Compound	R_1	R_2	R_3	R_4	PIC_{50}	Binding Energy (kcal/mol)
89	H	CH_3	H	$-(CH_2)_2-CH_3$	5.05	-6.1
90	H	Cl	Cl	H	5.22	-5.8
91	H	Cl	H	H	4.96	-5.8
92	Cl	Cl	H	H	4.70	-6.1
93	Cl	H	Cl	H	4.62	-6
94	H	OCH_3	H	H	4.66	-5.6
95	H	Cl	H			-7.2
96	H	Cl	H		5.30	-6.7

97	H	Cl	H		4.62	-6.7
98	H	Cl	H		4.20	-7.7
99	H	Cl	H		4.85	-6.9
100	H	Cl	Cl	$-CH_2-CH_3$	4.74	-5.7
101	H	Cl	Cl	$-(CH_2)_2-CH_3$	4.89	-6.1
102	H	Cl	Cl	$-(CH_2)_3-CH_3$	5.22	-6.2
103	H	Cl	Cl		5.10	-7.6
104	H	Cl	Cl		5.40	-7.2
105	H	Cl	Cl		5.40	-7.6
106	H	Cl	Cl		5.70	-7
107	H	Cl	Cl		4.72	-7.6

108	H	Cl	Cl		4.40	-7
109	H	CH_3	H		4.74	-7.6

716
Table 2. Brief description of some descriptors used in this study 717
718

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 van der 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 index,RCI;Jug RC index HOMA 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

719
720

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

Eq.	Descriptors	(+) effect	(-) effect	R^2	F	Q^2	SE
1	Chemical	logp	SA	0.489	16.28	0.40	0.37
2	constitutional	nF, nDB, nCl, nR09, nX, nCIC,nBr	--	0.611	17.78	0.58	0.21
3	Topological descriptors	--	SIC2, STN	0.613	23.18	0.58	0.23
4	Molecular walk counts	MWC03	MWC10, PIPC09	0.618	13.276	0.59	0.321
5	BCUT descriptors	BELm3	BELv8	0.416	15.655	0.39	0.226
6	Galvz topol. Charge in dices	GGI7	JGI3	0.473	15.765	0.43	0.480
7	2D autocorrelations	GATS1M	ATS6e, MATS3E	0.567	17.564	0.52	0.337
8	Charge descriptors	Qpos	SPP	0.347	14.674	0.29	0.308
9	Burden eigenvalues	BEHm1	----------	0.546	21.567	0.51	0.112
10	Geometrical descriptors	H3D, G(Cl..Cl)	DISPV, MAXDP	0.578	13.478	0.52	0.214
11	RDF descriptors	RDF085m, RDF110u	RDF100e	0.567	18.543	0.53	0.336
12	3D MoRSE descriptors	MOR30M, Mor31u	Mor06v	0.543	23.432	0.52	0.454
13	WHIM descriptors	E1m, P1P	G2M	0.654	32.678	0.61	0.241
14	GETAWAY descriptors	R3v+,R1p+	HATS5e ,HATS6n	0.673	32.451	0.63	0.242
15	Fuctional group counts	noxim, n pyridine, n isothiocyanate, nthiocyanate	nC=S, $nArNO_2$, noxazole, nThiazol, nCOOH, nCOOCH3	0.77	30.211	0.72	0.340
16	Charge descriptors	QMEAN, QPOS	--	0.55	34.231	0.51	0.321

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	0.71	0.67	0.23	0.74	**0.21**
GA-PLS	0.81	0.78	0.31	0.85	**0.17**
PCR	0.73	0.70	0.15	0.75	**0.20**
FA-MLR	0.657	0.62	0.31	0.74	**0.32**

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

738 **Table 5.** Numerical values of factor loading numbers 1–4 for descriptors after
VARIMAX rotation 739

740

	Component						
	1	2	3	4	5	6	7
SIC2	-0.617	0.109	0.094	-0.364	-0.199	0.012	0.097
nC=S	0.948	-0.406	0.103	-0.032	-0.036	-0.092	0.155
logp	0.697	0.316	-0.673	0.084	0.050	-0.312	0.397
nF	0.164	0.555	-0.146	0.170	0.088	-0.047	0.029
nDB	-0.123	0.047	0.286	0.109	0.035	-0.039	-0.036
G(Cl..Cl)	0.883	-0.031	0.853	0.009	0.109	0.053	-0.152
nCl	0.762	0.454	0.041	-0.081	0.099	0.017	0.106
nArNO2	0.609	0.067	0.159	0.039	-0.181	-0.106	0.856
nR09	0.807	0.134	-0.105	-0.159	-0.055	-0.157	0.017
nX	0.858	0.080	0.261	0.075	-0.106	-0.017	0.195
SA	-0.779	0.229	0.232	-0.003	0.009	0.209	-0.001
Qpos	0.334	0.409	0.272	-0.017	-0.081	-0.028	0.155
nCIC	-0.292	-0.073	-0.251	-0.163	0.039	0.114	0.397
STN	0.163	0.022	-0.195	-0.070	-0.159	0.077	0.029
MWC03	-0.858	-0.188	0.100	0.827	0.075	0.262	-0.036
MWC10	-0.065	-0.130	-0.126	0.791	-0.003	0.277	-0.152
PIPC09	0.518	0.107	0.853	-0.102	-0.017	-0.028	0.106
G(Cl..Cl)	-0.123	0.134	0.041	-0.061	-0.163	0.114	0.856
BELm3	0.883	0.080	0.159	-0.651	-0.070	0.077	0.017
BELv8	0.762	0.229	-0.105	-0.007	0.827	0.262	0.195
GGI7	0.609	0.409	0.261	0.520	0.791	0.277	-0.001
JGI3	0.807	-0.073	0.232	0.149	-0.102	-0.023	0.016
GATS1M	0.858	0.022	0.272	-0.052	-0.061	-0.066	-0.028
ATS6e	-0.779	-0.188	-0.251	-0.175	0.046	-0.072	-0.076
MATS3E	0.334	-0.130	-0.195	-0.002	-0.033	0.072	0.084
JGI5	-0.292	0.107	0.100	0.261	0.008	0.026	-0.004
SPP	0.163	-0.017	-0.126	-0.651	-0.087	0.241	-0.023
SA	-0.858	0.057	0.014	-0.007	0.078	-0.089	-0.010

n pyridine	-0.065	0.653	0.177	0.520	-0.056	0.039	0.122
nROR	0.518	0.734	0.161	0.149	0.046	0.138	0.005
Noxim	-0.781	0.258	-0.085	-0.141	-0.033	0.156	0.108
isothiocyanate	-0.927	0.009	-0.183	0.053	0.008	0.007	0.066
nArNO2	0.127	-0.038	0.086	-0.921	-0.087	0.084	-0.001
nAzole	-0.865	0.124	-0.181	0.226	0.078	-0.024	0.258
nThiazol	-0.629	-0.149	-0.312	-0.257	-0.056	-0.441	-0.043
nCOOH	0.044	0.066	-0.108	-0.359	0.039	0.770	0.111
nCOOCH3	0.022	0.447	-0.069	0.464	-0.365	0.199	0.008
nthiocyanate	0.677	0.528	0.186	0.164	-0.030	0.347	0.036
N piperidine	0.110	0.760	-0.081	0.458	-0.021	0.178	0.128
R3v+	0.891	0.075	-0.279	-0.122	-0.048	0.195	0.031
HATS5e	-0.629	0.266	-0.349	0.358	0.027	-0.163	0.085
HATS6n	0.275	0.645	0.125	-0.071	0.099	0.279	-0.340
% variances	37.86	15.85	7.91	7.65	4.45	4.28	3.15

741
742
743
744
745

Table 6. The results of FA-MLR analysis with different types ofdescriptors

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				

746
747
748

Table 7. The results of PCR analysis

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	R^2	F	Q^2	SE
	B	Std.Error	Beta						
(Constant)	4.742	0.043		105.268	0.000	0.73	15.54	0.70	0.23
F1	0.654	0.043	0.518	6.602	0.000				
F6	0.765	0.043	0.241	3.078	0.003				
F3	-0.456	0.043	-0.239	-3.050	0.003				
F2	0.321	0.043	0.157	1.998	0.049				

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
6	0.158855	0.101806	0.041009	0.060281
8	0.045048	0.13409	0.022111	0.063121
10	0.109807	0.227308	0.018691	0.025659
16	0.102708	0.198805	0.021734	0.045611
17	0.105906	0.127991	0.022526	0.016686
20	0.117418	0.084609	0.026426	0.014426
23	0.058532	0.058078	0.03644	0.028202
27	0.087443	0.084802	0.101804	0.034729
30	0.087529	0.067963	0.092915	0.035335
59	0.04769	0.157524	0.03296	0.021066
60	0.081846	0.093302	0.016547	0.037432
70	0.077447	0.058078	0.026426	0.068055
73	0.109807	0.07017	0.022111	0.063121
75	0.102708	0.084802	0.06149	0.056011
90	0.105906	0.127991	0.106844	0.036003
96	0.081846	0.084609	0.10121	0.040156
102	0.071099	0.08314	0.102167	0.056011
104	0.054337	0.077263	0.06149	0.036003
105	0.081619	0.134119	0.023009	0.068055
108	0.097168	0.144921	0.023009	0.022631
h*	0.33707	0.2696	0.13483	0.10112

Table9. binding interaction of compounds 39,46 and 68-69 in active site of enzyme

Compounds	Hydrogen bonds		Aromatic bonds		Hydrophobic interaction	
	Amino acid	Distance	Amino acid	Distance	Amino acid	Distance
39	Cys163 His121 Gly122	3.62 3.05 2.85	Phe 256	3.65		
46	Phe 250 Arg207	2.90 2.93				
68	Phe250 Ser 249 Glu 248	2.66 3.03 3.01	Trp206	3.76	Gln217	3.26
69	Trp214 Asn208 Ser209 Arg207 Phe250	3.16 3.08 3.06 2.80 3.79				

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

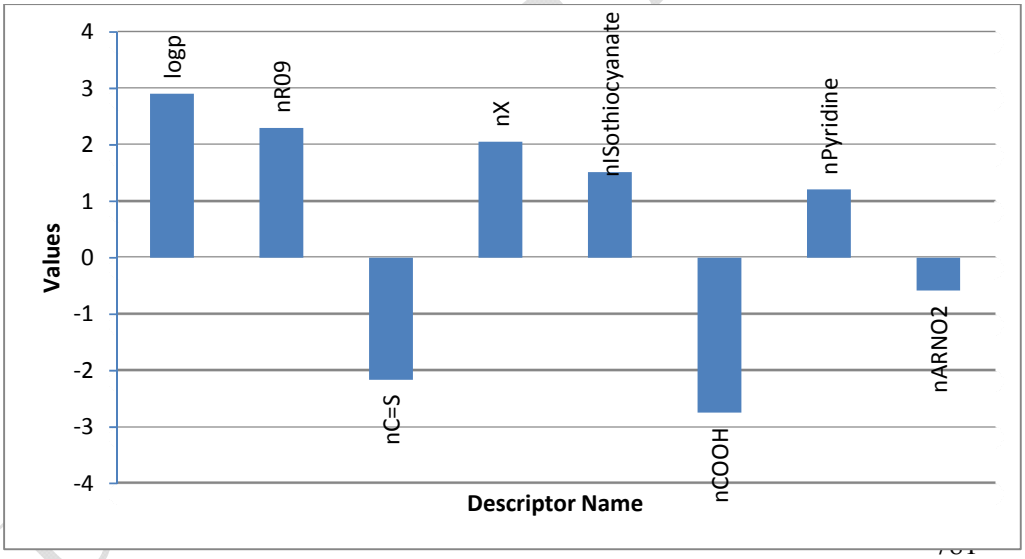


Figure 2. Plot of variables important in projection (VIP) for the descriptors used in GA-PLS model

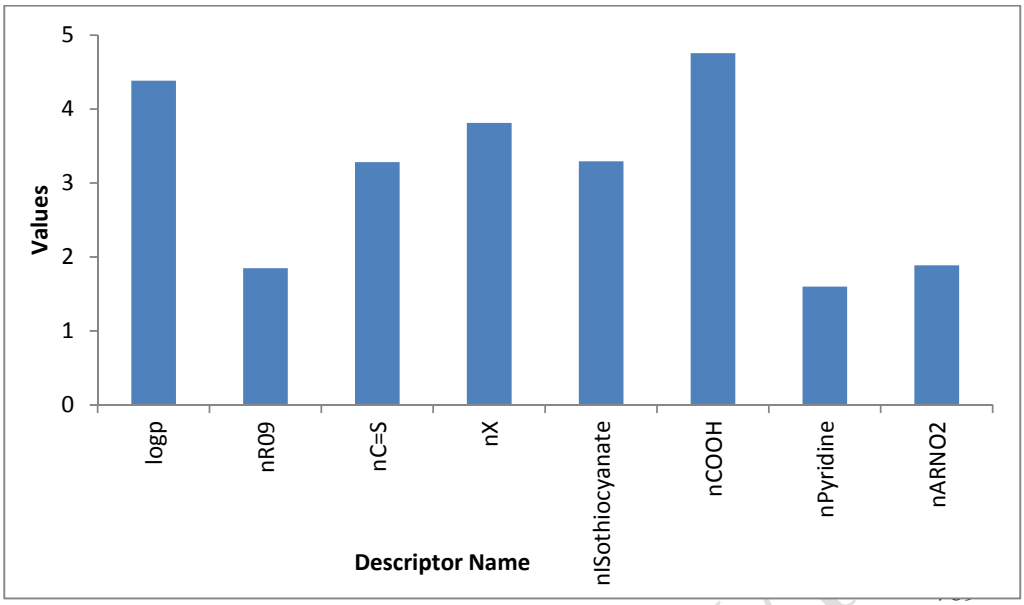


Figure 3. The docked configuration of 39 (A), 46(B), 68(C) and 69 (D) in the binding site of 3GFW