Original Research Article

QSAR, molecular docking Studies of 6-(Amino methyl)-5-(2,4-dichlorophenyl)-7-methylimidazo[1,2-a] pyrimidine-2- carboxamides as Potent, Selective Dipeptidyl Peptidase-4 (DPP4) Inhibitors

Abstract:

Type 2 diabetes (T2DM) is a metabolic disorder disease and DPP-4 inhibitors are a class of oral hypoglycemics that block the enzyme dipeptidyl peptidase-4 (DPP-4). DPP-4 inhibitors reduce glucagon and blood glucose levels and don't have side effects such as hypoglycemia or weight gain. In this paper a series of imidazolopyrimidine amides analogues as DPP4 inhibitors were applied for quantitative structure–activity relationship (QSAR) analysis. A collection of chemometric methods such as multiple linear regression (MLR), factor analysis-based multiple linear regression (FA-MLR), principal component regression (PCR) , genetic algorithm for variable selection-MLR (GA-MLR) and partial least squared combined with genetic algorithm for variable selection (GA-PLS), were conducted to make relations between structural features and DPP4 inhibitory of a variety of imidazolopyrimidine amides derivatives. GA-PLS represented superior results with a high statistical quality (R^2 = 0.94 and Q^2 = 0.80) for predicting the activity of the compounds. Docking studies of these compounds reveals and confirms that, compounds 15, 18, 25, 26, and 28 are introduced as good candidates for DPP-4 inhibitors were introduced as a good candidate for DPP-4 inhibitory compounds.

Keywords:

Imidazo pyrimidine derivatives, DPP-4 inhibitors, QSAR, Molecular docking

1.Introduction

Diabetes Mellitus (DM) is a metabolic disorder disease that body doesn't have the ability to produce insulin or is resistant to insulin so it cannot function properly. Dipeptide peptidase 4 inhibitors are a new therapy Target that does not complicate previous medications such as hypoglycemia, weight gain and cardiovascular risk. DPP-4 is a membrane protease that has a specific selectivity on the secretion of incretins hormones in fact, these drugs can effective in controlling the secretion of insulin and reducing glucagon secretion on hemostasis of glucose. So reducing glucagon secretion and can also affect the process of gluconeogenesis in the liver. Therefore, by inhibiting this physiological pathway in the body, effectively reduce blood glucose levels.

The quantitative structure-activity relationship (QSAR) research field provides medicinal chemists with the ability to predict drug activity by mathematical equations which construct a relationship between biological activity of the molecules and descriptors [1, 2]. These mathematical equations are in the form of $y = Xb+e$ that describe a set of predictor variables (X) with a predicted variable (y) by means of a regression vector (b) [3]. The most important step in building QSAR models is the appropriate representation of the structural and physicochemical features of structures [4-10]. These features called molecular descriptors are the ones with higher impact on the biological activity of interest. Nowadays, a wide range of descriptors are being used in QSAR studies which can be classified into different categories according to the Karelson approach including; constitutional, geometrical, topological, quantum, chemical and so on [8].Hyperchem and Dragon are two well-known computational softwares provide us more than 1000 of these descriptors [11-12].There are different variable selection methods available including; stepwise multiple linear regression (MLR), genetic algorithm (GA), principal component or factor analysis (PCA) and so on. The mathematical relationships between molecular descriptors and activity are used to find the parameters affecting the biological activity and/or estimate the property of other molecules.

Here, we consider the DPP4 inhibitory activity of a novel series of imidazolopyrimidine amides which have been recently designed and synthesized by W. Meng [13]. Our research show that these series of compounds don’t evaluate for QSAR studies. Different statistical methods were applied to model the relationship between the structural features and the DPP-4 inhibitory activity of the studied compounds. These methods are: (i) multiple linear regression (MLR), (ii)

principal component regression (PCR), MLR with factor analysis as the data pre-processing step for variable selection (FA-MLR) (iii), genetic algorithm-multiple linear regression (GA-MLR) (iv), genetic algorithm - partial least squares (GA-PLS) (v). Molecular docking simulation technique was also performed on twenty-nine compounds to reach the details molecular binding models for these compounds interacting with the key active site DPP-4 inhibitors.

2. Materials and methods

2.1. Data set

The biological activity was used in this study, were the DPP-4 inhibitory activity of a set of thirty-one imidazolepyrimidine amides derivatives (13), which were designed, synthesized and evaluated for their ability as potential treatments for type II diabetes. The structural features and biological activities of these compounds are listed in Table 1. The biological data were converted to logarithmic scale (pIC50) and then used for subsequent QSAR analysis as dependent variable.

[Table 1. near here]

2.2 Molecular descriptors

All structures were generated with HyperChem program (Hyper-cube Inc., Version 8.0.3) [11] and optimized by MM+ method and then semi-empirical AM1 method in hyperchem software. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal mol^{-1}. Some chemical parameters including molar volume (V), molecular surface area (SA), hydrophobicity (logP), hydration energy (HE) and molecular polarizability were calculated by using Hyperchem software. The resulted geometry was transferred into Dragon program package, which was developed by Milano Chemometrics and QSAR Group. Dragon software (version 5.5) [12] calculated different topological, geometrical, charge, empirical and constitutional descriptors for each molecule. 2D autocorrelations aromaticity indices, atom-centered fragments and functional groups were also calculated by dragon software.

In the case of docking procedure, each optimized structures in HyperChem 8.0.3 program were there after converted to PDBQT using MGLtools 1.5.6 [14]. The three dimensional crystal structure of dipeptidyl peptidase iv human (PDB ID:5j3j) were retrieved from protein data bank

[15]. Co-crystal ligand molecules were excluded from the structures and the PDBs were corrected in terms of missing atom types by modeller9.12 [16]. An in house application (MODELFACE) was used for generation of python script and running modeler software [17]. Subsequently, the enzymes were converted to PDBQT and gasteiger partial charges were added using MGLtools1.5.6.

2.3. Data screening and model building

The calculated descriptors were collected in a data matrix, D whose number of rows and columns were the number of molecules and descriptors, respectively. First the descriptors were checked for constant or near constant values and those detected were removed from the original data matrix. The correlated descriptors with each other's and with the activity data were determined and removed from pool of descriptors.

Five different methods were used: (1) stepwise-multiple linear regression (2) MLR with factor analysis as the data pre-processing step for variable selection (FA-MLR), (3) principal component regression analysis and (4) genetic algorithm- multiple linear regression (GA-MLR) (5) genetic algorithm- partial least squares (GA-PLS).

MLR with stepwise selection and elimination of variables was applied for developing QSAR models by using SPSS software (SPSS Inc., version 21). The resulted models were validated by leave-one out cross-validation procedure by using MATLAB software version 2014. However, this procedure did not produce good results and therefore we used genetic algorithm (GA-PLS) to select the best variables. FA-MLR was performed on the dataset. Factor analysis was used to reduce the number of variables. Principal component regression analysis was also 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 [18]. In this method, factor scores, as obtained from FA, are used as the predictor variables [19]. In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors. Partial least squares (PLS) linear regression is a recent technique that generalizes and combines features from principal component analysis and multiple regressions. PLS is a method suitable for overcoming the problems in MLR related to multicollinear or over-abundant descriptors [20]. Application of PLS method thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most

variables. This method is normally used in combination with cross-validation to obtain the optimum number of components [21]. The PLS regression method used was the NIPALS-based algorithm existed in the chemometrics toolbox of MATLAB software (version 8.0.3.532 Math Work Inc.).

2.4. Docking procedures

An in house batch script (DOCK-FACE) for automatic running of AutoDock 4.2 was used to carry out the docking simulations [22] in a parallel mode [23]. To prepare the receptor structure, the three dimensional crystal structure of Dipeptidyl Peptidase-4 (PDB ID: 5j3j) was acquired from Protein Data Bank (PDB data base; http://www.rcsb.org) [24] and water molecules and co-crystal ligand were removed from the structure. The PDB were then checked for missing atom types with the python script as implemented in MODELLER 9.17 [25]. 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 30, 30, and 30 in x-, y-, and z directions 20.3, 3.7 and 51.3 were used. Number of points in x, y and z was and 65 respectively. All visualization of protein ligand interaction was evaluated using VMD software [26]. Cluster analysis was performed on the docked results using a root mean square deviation (RMSD) tolerance of 2.4 Å.

3. Results and Discussion

The structural feature and the experimental DPP-4 inhibitory activity (represented as pIC_{50}) of the molecules used in this study are shown in Table 1. To obtain the effects of the structural parameters of the investigated derivatives on their DPP-4 activity, QSAR analysis was performed with various molecular descriptors. Among the different chemometric tools available for modeling the relationship between the biological activity and molecular descriptors, five methods (i.e., stepwise MLR, PCR, FA-MLR, GA-MLR and GA-PLS) were applied and compared here. The calculated descriptors from whole molecular structures are briefly described in Table 2.

[Table 2. near here]

3.1. MLR models for subset of molecules

Firstly, separate stepwise selection-based MLR analyses were performed using different types of descriptors, and then, a MLR equation was obtained utilizing the pool of all calculated descriptors. First principal component analysis was done to detect outlier data and was drawn PC1 on PC2 (Figure 1), as it can show the molecule number of 1 and 16 are outlier data so omitted. Then Kennard stone algorithm was used to divide data set to calibration and prediction set. MLR models with maximum number of variables of 5 were selected. Statistical parameters such as correlation coefficient (R^2), correlation coefficient for test set (R^2test set or R^2predic), standard error of regression (SE), and Fisher ratio (F) at specified degrees of freedom, leave-one-out cross-validation correlation coefficient (Q^2) was shown in Table 3. Equation 1 was selected as the best equation in the MLR model because of its greatest statistical parameters. The selected variables demonstrate that 2D-autocorrelation (MATS1m), constitutional (Ms), topological charge indices (GGI5), topological (DELS), 3D-MORSE descriptors (Mor25m) effect on the inhibitory activity of the studied compounds.

A small difference between the conventional and cross-validate correlation coefficients of the different MLR equations (Table 4) reveals that none of the models are over fitted, which can be partially attributed to the absence of collinearity between the variables in one hand and use of no extra variables on the other hand. Equation 1 (as the best equation in this series) could explain 91% of the variance and predict 84% of the variance in ($-\log IC_{50}$) data. All of the descriptors

that used in this equation have positive effect on DPP-4 inhibitory expect DELS as topological descriptors. Figure 2 shows the plots of linear regression predicted versus experimental value of the DPP4 inhibitory activity of ligand. The plots for this model show to be more convenient with $R^2cv= 0.84$.

[Table 3. near here], [Table 4. near here]

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

3.2. PCR Analysis

When factor scores were used as the predictor parameters in a multiple regression equation (Table 5), a predictive QSAR model with factor scores of 1, 2, 3 and 7 as input variables, was obtained (Table 3, Equation 2). This equation shows statistical quantities similar to those obtained by the FA-MLR method.

Considering this information in modeling, it may apparently increase the model variances (i.e., R^2) but they are useful for prediction. Figure 2 shows the plots of linear regression predicted versus experimental value of the DPP-4 inhibitory activity of ligand. The plots for this model show to be more convenient with $R^2cv= 0.75$.

[Table 5. near here]

3.2. FA-MLR *analysis*

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. 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 [27]. In this method, factor scores, as obtained from FA, are used as the predictor variables [28]. In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors. Table 5 shows the two factor loadings of the variables (after VARIMAX rotation) for the compounds tested against dipeptidyl peptidase 4 inhibitors'. As it is observed, about 77% of variances in DPP4 inhibitors' could be explained by the selected two factors. It is observed; about 0.67 of variances in the original data matrix can be explained by selected 2 factors, MATS1m as 2D-autocorrelation descriptors and Ms as Constitutional descriptors. And also have

weak predicted variance in DPP4 inhibitory. Figure 2 shows the plots of linear regression predicted versus experimental value of the DPP4 inhibitory activity of ligand. The plots for this model show to be more convenient with $R^2cv= 0.53$.

3.3.GA-MLR analysis

Genetic algorithm technique was employed as a selection tool to select the most relevant descriptors with respect to an objective function. The genetic algorithm (GA) starts with the creation of a population of randomly generated parameter sets. the parameters set used for the GA includes population size (160), initial terms 18%, max generation (250) and %convergences (90%), These selected subsets of variables are further evaluated by their fitness to predict inhibitory activity values. multiple linear regression analysis was performed on the training set and then, evaluated by test set. Using genetic algorithm-multiple linear regression (GA-MLR) analysis resulted in the development of a predictive QSAR model with four descriptors with the following equation:

PIC50=2.072GGI7 ((±0.676)+4.427Ms((±0.678)+8.047BELm6(±1.753)-0.453MOr27u(±0.187)-14.411(±3.275)

The statistical parameters of GA-MLR model are shown in Table 3.and could explain 94% of the variance and predict 88% of the variance in ($-\log IC_{50}$) data. This equation describes the effect of GGI7(Topological charge indices), Ms (Constitutional), BELm6(Burden Eigenvalues) and MOr27U (3-D Morse Descriptors) in dpp4 inhibitory. All the descriptors have positive coefficient except MOR27u and indicated that increase this descriptor (MOR27u) could result in decreasing PIC50. Figure 2 shows the plots of linear regression predicted versus experimental value of the dpp4 inhibitory activity of ligand. The plots for this model show to be more convenient with $R^2cv= 0.88$.

3.4.GA-PLS analysis

In PLS analysis, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. Therefore, unlike MLR analysis, the multi collinearity problem in the descriptors is omitted by PLS analysis. Because a minimal number of latent variables are used for modeling in PLS; this modeling method coincides with noisy data better than MLR. In order to find the more convenient set of descriptors in PLS

modeling, genetic algorithm was used. To do so, many different GA-PLS runs were conducted using different initial set of populations.

The data set (n = 29) was divided into two group: calibration set (n = 20) and prediction set (n = 9). Given 20 calibration samples; the leave-one out cross-validation procedure was used to find the optimum number of latent variables for each PLS model.

[Table 6. near here]

The most convenient GA-PLS model that resulted in the best fitness contained 9 indices, five of them being those obtained by MLR. The PLS estimate of coefficients for these descriptors are given in Figure 3. As it observed, a combination of Constitutional, Topological, Connectivity indices, 2D-autocorrelation, Edge adjacency indices, Topological charge indices, 3-D Morse Descriptors, WHIM Descriptors have been selected by GA-PLS to account the Dipeptidyl Peptidase-4 (DPP4) inhibitory activity of imidazole derivatives. The resulted GA-PLS model possessed a high statistical quality $R^2 = 0.94$ and $Q^2 = 0.80$. The predictive ability of the model was measured by applying to 10 external tests set molecules. The squared correlation coefficient for prediction was 0.95 and standard error of prediction was 0.49. The values of pIC_{50} using GA-PLS model (refined from cross-validation or external prediction set) are shown in Table 1**.** This figure 3 describes the effect of Ms (Constitutional), X0A (Connectivity indices), MATSIM(2D-autocorrelation), EEig09d, EEig13d (Edge adjacency indices), GGi5(Topological charge indices), Mor25m (3-D Morse Descriptors), E2M (WHIM Descriptors) and DELS (Topological) on inhibitory DPP4 activity. And also describe that X0A, EEig13d and DELS have negative coefficient on DPP4 inhibitory but the other descriptors have positive effect on DPP4 activity. Figure 2 shows the plots of linear regression predicted versus experimental value of the DPP4 inhibitory activity of ligand. The plots for this model show to be more convenient with $R^2cv=$ 0.80.

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 [29]. 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 [30].

[Figure 4. Near here]

The VIP analysis of PLS equation is shown in Figure 4. VIP analysis shows that Ms which is constitutional descriptors, X0A as Connectivity indices descriptors, MATS1m which is 2D-autocorrelation and GGI5 which is topological charge indices parameter, are the most important indices in the QSAR equation derived by PLS analysis. In addition, the other descriptors have been found to be low influential parameters.

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 [31]. 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 6. As seen, the leverages of all test samples are lower than h^* for all models. This means that all predicted values are acceptable.

[Table.6 near here]

3.6. Docking Study

Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to predict the affinity and activity of the small molecule. Hence docking plays a great role in the rational design of drugs. Here, docking studies were carried out on our compounds to find their binding site, binding modes and the best direction on the base of their binding energy. Having completed the docking process, the protein– ligand complex was analyzed to investigate the type of interactions. The conformation with the lowest binding energy was considered as the best docking result in each case resource.

As it was shown in Table 1, Compounds 15, 18, 25, 26 and 28 based on their highest docking binding energy can be a good candidate for DPP-4 inhibitors.

On the other hand, promising results such as the ligand-receptor binding site and binding modes were obtained from docking analyses. The results for each ligand were compared to its corresponding co-crystal ligand.

[Figure 5. near hear]

On the other hand, promising results such as the ligand-receptor binding site and binding modes were obtained from docking analysis. The results for each ligand were compared to its corresponding co-crystal ligand. a hydrogen bond acceptor interaction between amino group of co-crystal ligand (HL1) and Glu 166, Glu 167 and Tyr 623 of the receptor (salt bridge). Trifluoro phenyl group was occupying S1hydrophobic pocket of dpp4 inhibitors with val 627, His701, Val617, Tyr 592, Tyr 627 and Tyr506 residue of receptor [32]. Dimethoxy phenyl group of co-crystal ligand π-π interaction with Phe 318 of the receptor figure 5a. Hydrogen bindings between docked potent agents such as 25 and the dpp4 receptor (5j3j) figure 5b. dichloro phenyl group π-π interaction with Tyr 508 and were occupying S1 pocket. Hydrogen bonding interaction between methyl of imidazole pyrimidine of 25 molecules and Trp 590 of receptor.

4. Conclusion

In this study, five different QSAR modeling methods, MLR, FA-MLR, PCR, GA-PLS and GA-MLR were used in the construction of a QSAR model for DPP4 inhibitory of imidazolopyrimidine amides and the resulting models were compared. The reliability, accuracy and predictability of the proposed models were evaluated by root mean square error of cross-validation (RMSECV) and cross- validation, the root mean square error of prediction (RMSEP). Results confirm that among the applied models, the GA-PLS is superior for the prediction of the pIC_{50} of imidazolopyrimidine amides analogues. All models represent high goodness of fit (measured by R^2), whereas that obtained from GA-PLS is significantly better than that of the other models. The cross-validation statistics reported suggested that the higher prediction ability of the GA-PLS model. This study suggests the importance of constitutional, topological, connectivity indices, 2D-autocorrelation, edge adjacency indices, topological charge indices, 3D Morse descriptors, WHIM Descriptors of molecules for imidazolopyrimidine amides derivatives.

Docking study reveals and confirms that, compounds 15, 18, 25, 26, and 28 are introduced as good candidates for DPP-4 inhibitors.

Quantitative relationships between molecular structure and anti-cancer activity of isatin derivatives were discovered by two chemometrics methods: MLR and GA-PLS. Different QSAR models revealed that topological parameters (X3v and PJI2) have significant impacts on the anti-cancer activity of the compounds. In this series a significant role of chemical (Polarizability and HE), constitutional (Sv and nX) and geometrical parameters (G1 and SPAM) on the inhibitory activity was observed. Using the pool of all types of calculated descriptors a new QSAR model was derived for these compounds. In this model the importance of effects of topological (X3v and PJI2) and functional groups parameter (nCs) on the cytotoxic activity was indicated. The positive effects of the number of halogen atoms and the number of total secondary carbons, and the negative effects of the number of secondary amides, and the number of ketones on the anti-cancer activity was in agreement with previous SAR studies.. GA-PLS model showed the effects of seven topological indices (X3v, PW4, PJI2, MSD, SEigZ, IC1 and BIC2), three constitutional descriptors (Ss, Me and nBr) and one chemical parameter (Vol) on the cytotxic activity of the compounds. A comparison between the two statistical methods employed indicated that MLR represented superior results. The resulted MLR model possessed a high statistical quality (R^2 = 0.92 and Q^2 = 0.90) for predicting the activity of the compounds.

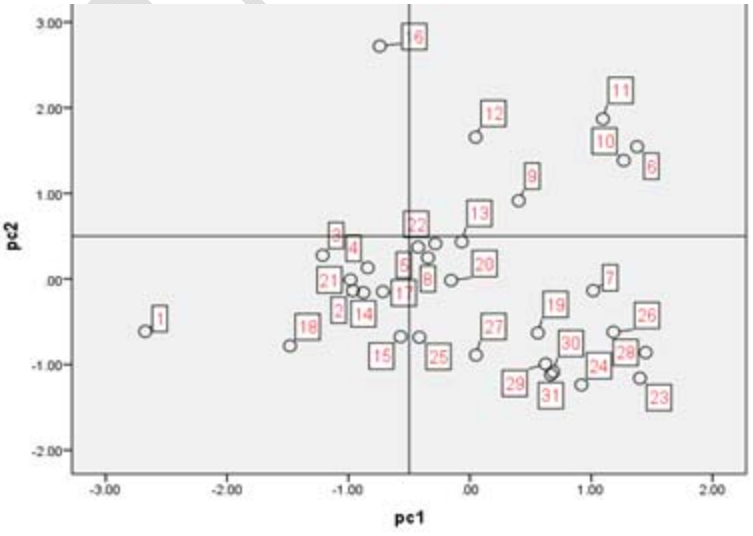
References

1. Hansch, C.; Hoekman D.; Gao, H. Comparative QSAR: Toward a Deeper Understanding of Chemicobiological Interactions. *Chem. Rev.* **1996**, *96*, 1045-1076.
2. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178-180.
3. 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.
4. 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.
5. Sabet, R.; Fassihi, A.; Moeinifard, B., QSAR study of PETT Derivatives as Potent HIV-Reverse Transcriptase Inhibitors. J. Mol. Graph & Model. 2009; 28: 146.
6. C. Hansch, D. Hoekman, H. Gao, Comparative QSAR: Toward a deeper understanding of chemicobiological interactions, Chem. Rev. 96 (1996) 1045-1075.
7. R. Todeschini, V. Consonni, Handbook of Molecular Descriptors. Wiley-VCH, Weinheim, 2000.
8. Sabet, R.; Fassihi, A., QSAR study of Isatin analogues as in vitro anti-cancer agents. Eur. J. Med. Chem. 2010; 45: 1113.
9. 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.
10. Karbakhsh, R.; Sabet, R.; Application of different chemometrics tools in QSAR Study of Azolo-adamantanes against influenza A virus. 2011**;**6,23**.**
11. Visit the Hyperchem official website at: http://www.hyper.com.
12. Todeschini, R. Milano Chemometrics and QSPR Group. http://michem.disat.unimib.it/
13. Wei Meng, Robert P. Brigance, Hannguang J. Chao, Aberra Fura, Discovery of 6-(Aminome thyl) -5-(2,4-dichlorophe nyl) -7-methy limidazo[1,2-a] pyrimidine-2-carbox amides as Potent, Selective Dipeptidyl Peptidase-4 (DPP4) Inhibitor s. J. Med. Chem. 2010, 53, 5620–5628.

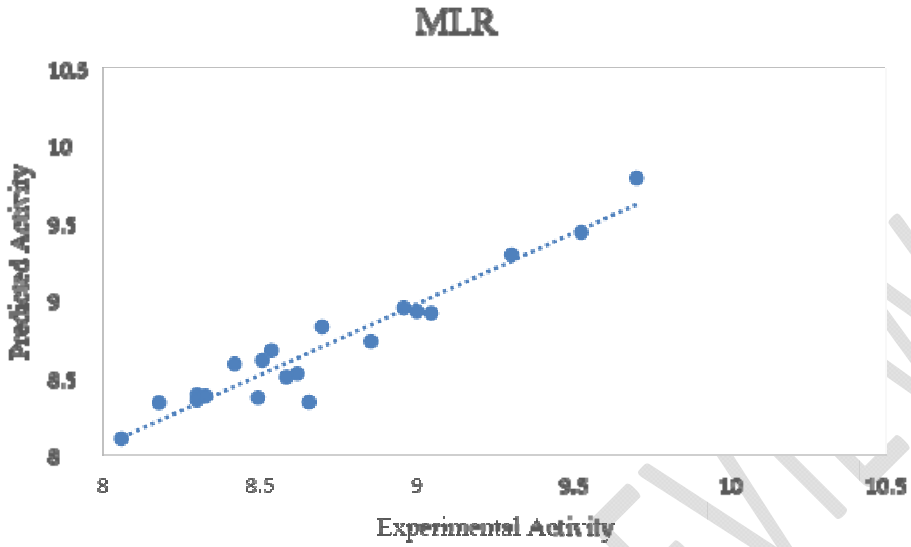
14. Morris GM, Huey R, Olson AJ. Using AutoDock for Ligand-Receptor Docking. *Curr Protoc Bioinformatics*. 2008; Chapter 8: Unit 8.14
15. Hikisz P, Szczupak Ł, Koceva-Chyła A, Oehninger L, Ott I, Therrien B, *et al*. Anticancer and Antibacterial Activity Studies of Gold (I)-Alkynyl Chromones. *Molecules.* 2015; 20:19699-718
16. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-y, *et al.* Comparative Protein Structure Modeling Using Modeller. *Curr Protoc Bioinformatics.* 2006; Chapter 5: Unit 5.6
17. Sakhteman A. PreAuposSOM, https:// www.biomedicale.univ-paris5.fr/aupossom/
18. Leardi, R. Application of Genetic Algorithm-PLS for Feature Selection in Spectral Data Sets. *J. Chemomtr.* **2000**, *14*, 643-655
19. Siedlecki, W.; Sklansky, J. On Automatic Feature Selection. *Int. J. Pattern Recog. Artif. Intell.*, **1988**, *2*, 197-220.
20. Schmidi, H. Multivariate Prediction for QSAR. *Chemom. Intell. Lab. Sys.* **1997**, *37*, 125-134
21. Hansch, C.; Kurup, A.; Garg, R.; Gao, H. Chem-bioinformatics and QSAR. A Review of QSAR Lacking Positive Hydrophobic Terms. *Chem. Rev.* **2001**, *101*, 619-672.
22. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. Journal of molecular graphics. 1996;14(1):33-8.
23. Fereidoonnezhad M, Faghih Z, Mojaddami A, Sakhteman A, Rezaei Z. A Comparative Docking Studies of Dichloroacetate Analogues on Four Isozymes of Pyruvate Dehydrogenase Kinase in Humans. Indian J Pharm Educ. 2016;50(2):S32-S8.
24. Mirjalili BF, Zamani L, Zomorodian K, Khabnadideh S, Haghighijoo Z, Malakotikhah Z, et al. Synthesis, antifungal activity and docking study of 2-amino-4H-benzochromene-3-carbonitrile derivatives. Journal of Molecular Structure. 2016; 1116:102-8.
25. Li Z, Gu J, Zhuang H, Kang L, Zhao X, Guo Q. Adaptive molecular docking method based on information entropy genetic algorithm. Applied Soft Computing. 2015; 26:299-302.
26. Feng J, Ablajan K, Sali A. 4-Dimethylaminopyridine-catalyzed multi-component one-pot reactions for the convenient synthesis of spiro[indoline-3,4′-pyrano[2,3-c]pyrazole] derivatives. Tetrahedron. 2014;70(2):484-9.

27. A. Fassihi, D. Abedi, L. Saghaie, R. Sabet, H. Fazeli, G. Bostaki, O. Deilami, H. Sadinpour, Eur. J. Med. Chem. (2008), doi: 10.1016/j.ejmech.2008.10.022.
28. Sharaf MA, Illman DL, Kowalski BR. Chemometrics. New York: *Wiley*. 1986;332.
29. 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.
30. Mohajeri, A.; Hemmateenejad, B.; Mehdipour A.; Miri, R. Modeling Calcium Channel Antagonistic Activity of Dihydropyridine Derivatives Using QTMS Indices Analyzed by GA-PLS and PC-GA-PLS. J. Mol. Graph. Model. 2008, 26, 1057-1065
31. Brereton R. Chemometrics Data Analysis for the Laboratory and Chemical Plant. Wiley. 2004:47–54.
32. Liu. Y and Liu.T, Recent in non-peptidomimetic dipeptidyl peptidase 4 inhibitors: Medicinal chemistry and preclinical aspects. Current Medicinal Chemistry, 2012, 19, 3982-3999.

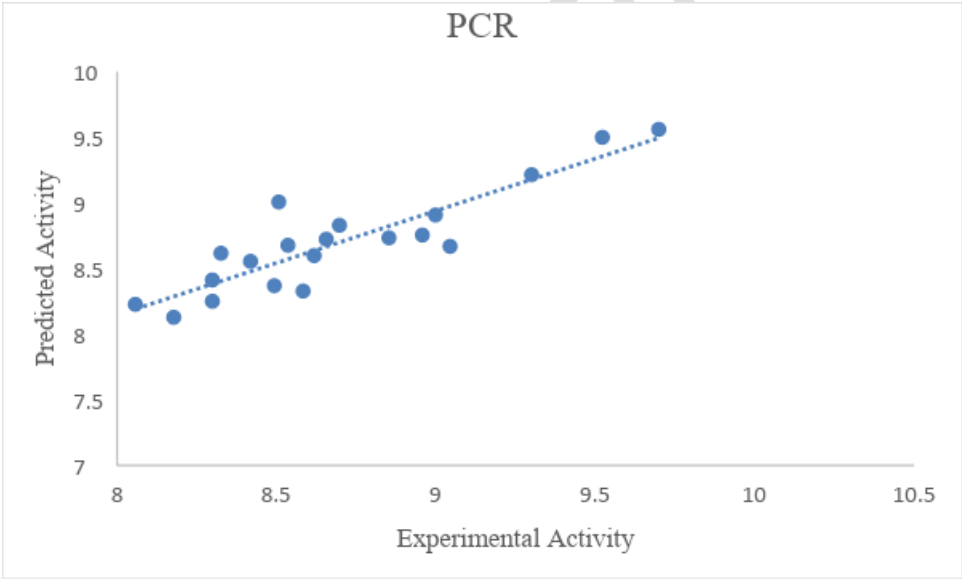
Figure 1. Principal component analysis diagram for detection of outlier data



433 **Figure 2.** Plots of the cross-validated predicted activity against the experimental activity for the
434 QSAR models obtained by different methods.

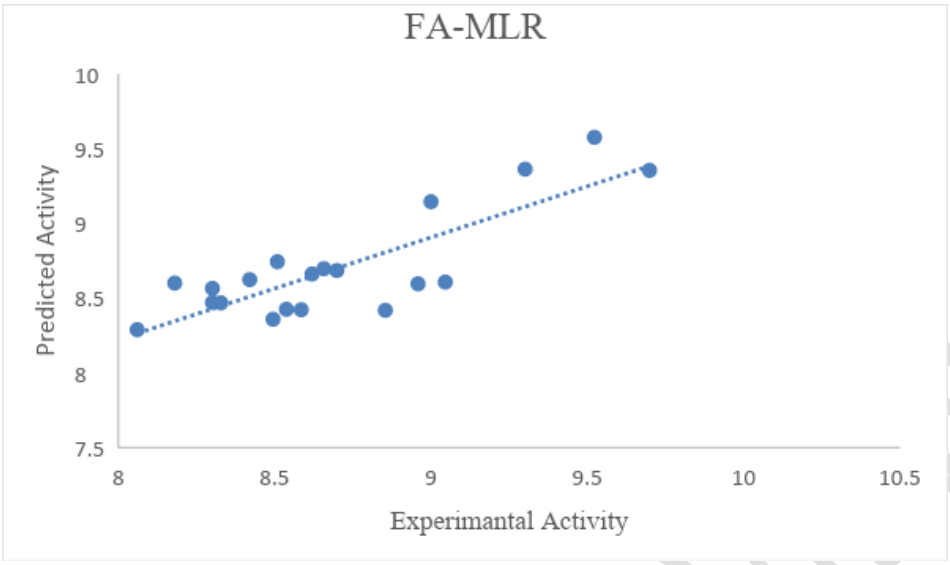


435



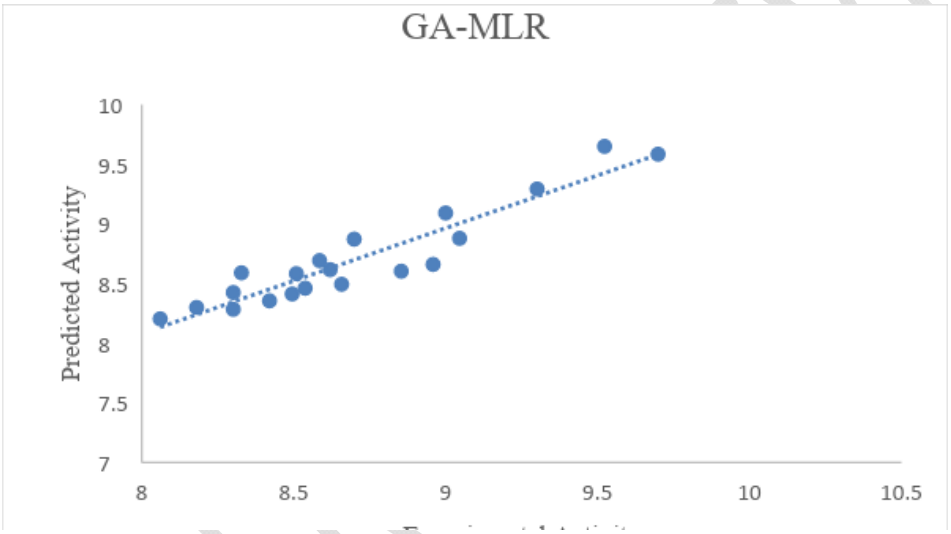
436

437



438

439



440

441

442

443

444

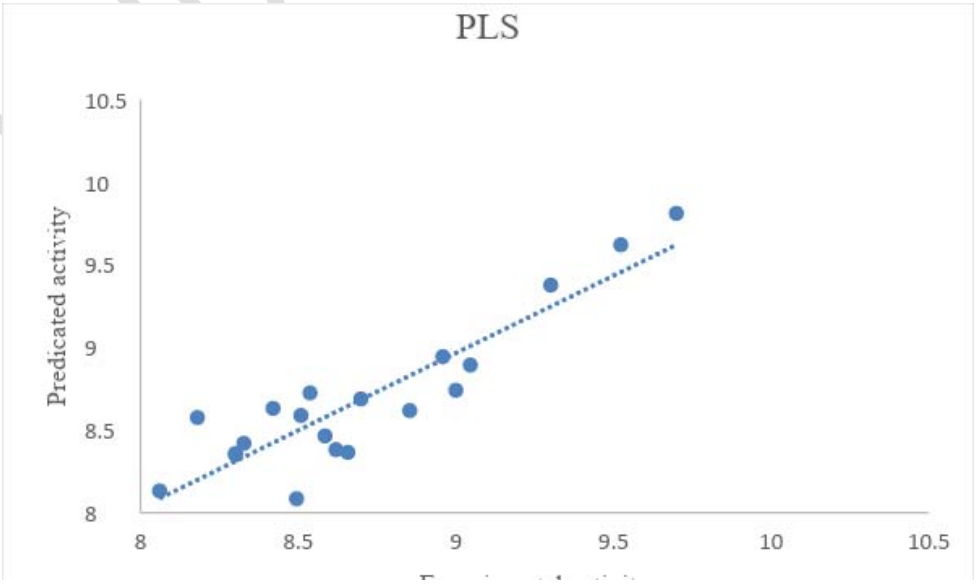
445

446

447

448

449



458

459

460

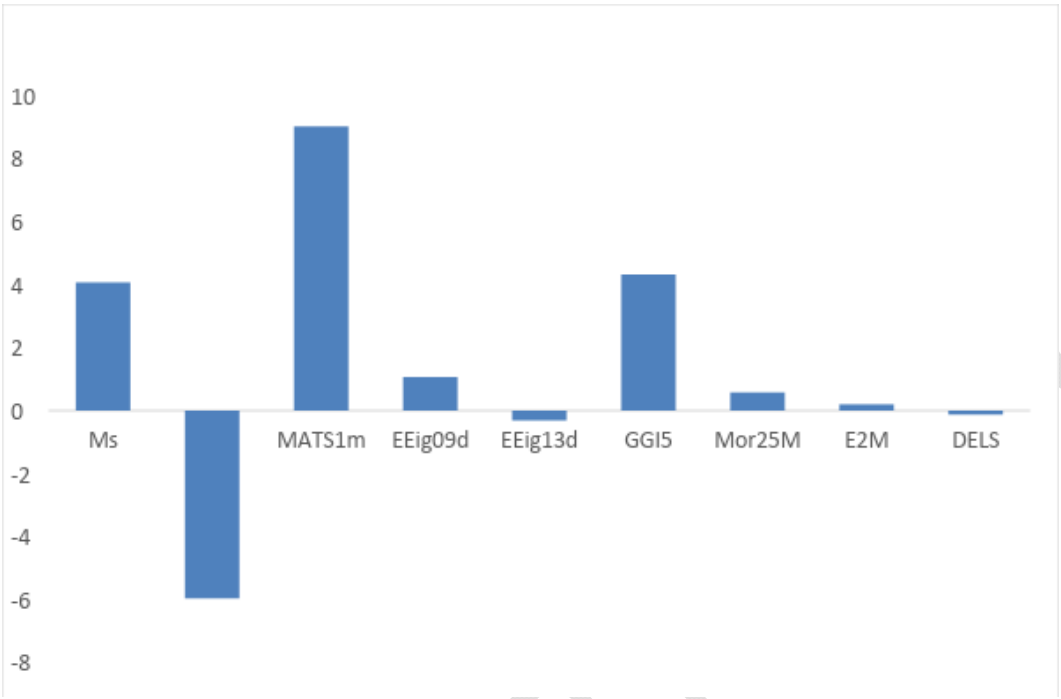


Figure 3. Plots of the cross-validated predicted activity against the experimental activity for the QSAR models obtained by GA-PLS methods.

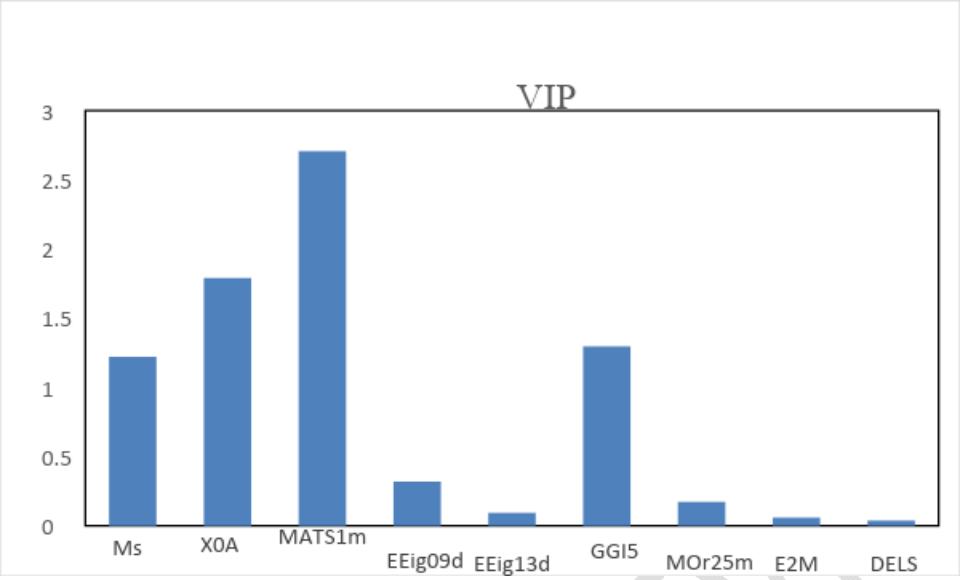


Figure 4. Plot of variables important in projection (VIP) for the descriptors used in GAPLS model.

Figure 5. Interactions of A) HL1 and B) compound 25 with the residues in the binding site of DPP4 (5j3j) receptor.

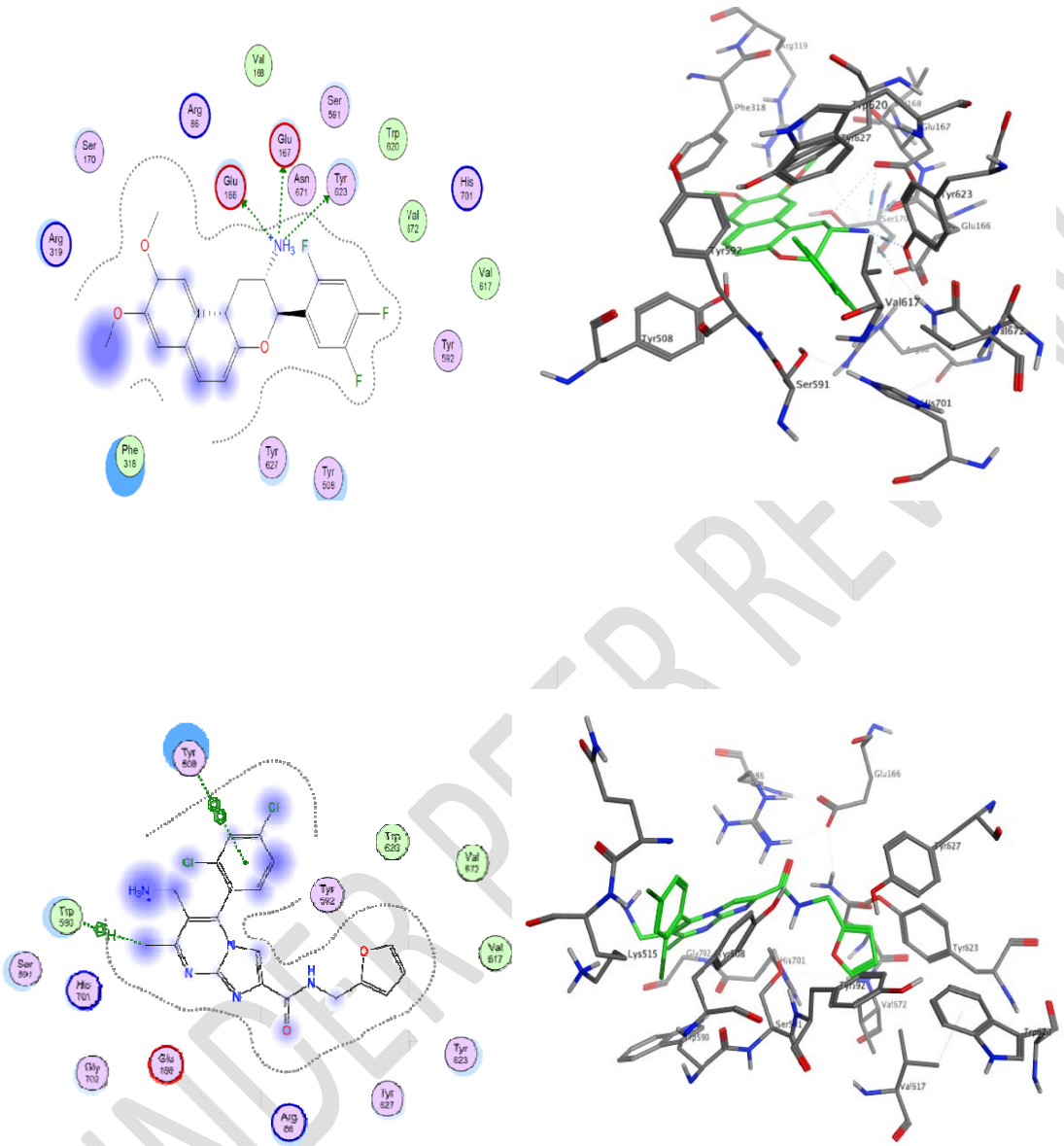
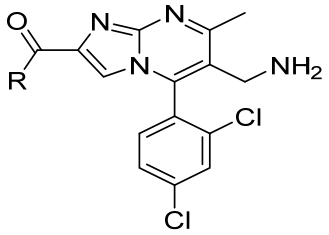
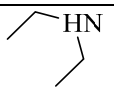
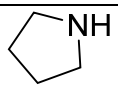
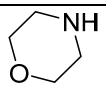
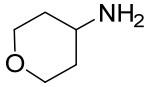
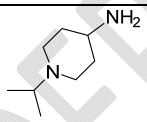
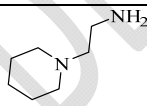
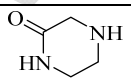
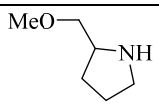
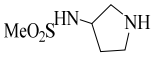
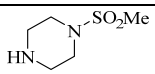
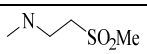
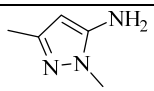
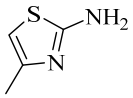
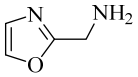
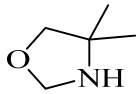
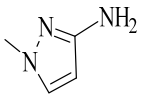
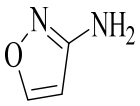
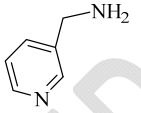
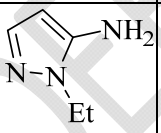
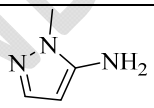
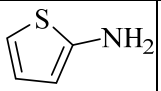
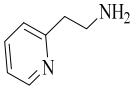


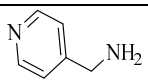
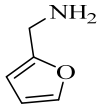
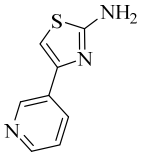
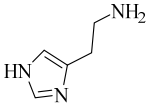
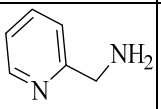
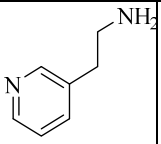
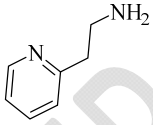
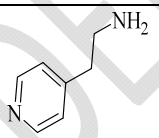
Table1. Chemical structure of imidazolopyrimidine amides analogues used and their experimental and cross validated-predicted activity by (GA-PLS) for DPP4 inhibitory and their docking bonding energies.



1-31

NO	R	Exp.pIC$_{50}$	Pred. pIC$_{50}$ by GA-PLS	Binding Energy (kcal/mol)
1*	OEt	9.39	------	------
2		8.6	8.38	-8.1
3		8.5	8.7	-8.7
4		8.6	8.3	-7.9
5**		8.5	8.65	-8.6
6		8.3	8.3	-8.1
7		8.06	8.1	-8.2
8		9	8.7	-8.1
9		8.5	8.5	-8
10		9.69	9.8	-8.2

11		9.3	9.3	-8.2
12		9.5	9.6	-7.9
13		9.04	8.9	-8.9
14		8.18	8.5	-8.4
15		8.69	8.7	-9
16*		------	------	------
17**		8.7	8.37	-8.5
18		8.40	8.60	-9.3
19**		8.58	8.3	-8.9
20		8.95	8.9	-8.6
21**		8.69	8.68	-8.6
22		8.49	8.48	-8.7
23**		9.15	9.1	-8.9

24		8.58	8.4	-8.4
25**		8.39	8.4	-9.4
26		8.3	8.35	-9.2
27		8.32	8.4	-8.3
28**		8.26	8.28	-9.4
29**		8.8	8.72	-8.8
30		8.8	8.6	-8.9
31**		8.8	8.73	-8.3

509 *: outlier data
510 **: molecules as test set
511
512

513

514 **Table 2.** Brief name of molecular descriptors was used in the models.

Descriptor type	descriptors	Brief description
Constitutional	Ms.	Mean electropological state
Topological	Jhetm	Balaban-type index from mass weighted distance matrix
	DELS	Molecular electropological variation
Connectivity indices	X0A	Average connectivity index chi-0
2D-autocorrelation	MATS1m	Moran autocorrelation – lag1/weighted by atomic Masses
Edge adjacency indices	EEig09d	Eigen values 09 from edge adj. matrix weighted by dipole moment
	EEig13d	Eigen values 13 from edge adj. matrix weighted by dipole moment
Burden Eigenvalues	BELm6	Lowest eigenvalue n.6 of burden matrix/weighted by atomic masses
Topological charge indices	GGI5	Topological charge index of order 5
	GGI7	Topological charge index of order 7
3-D Morse Descriptors	Mor27u	3D-MoRSE-signal 27/unweighted
	Mor25m	3D-MoRSE-signal 25/weighted by atomic masses
WHIM Descriptors	E2m	2nd component accessibility directional WHIM index/weighted by atomic masses

515

516

517

518

519

520 **Table 3.** The results of different QSAR model analysis

Models	Equation	N	R^2	Q^2	F	SE	R^2p
MLR	PIC_{50}=9.508 MATS1m (±2.252) +4.286 Ms (±0.78) +4.319 GGI5 (±0.816)-0.105 DELS (±0.028) +0.538 MOR25m (±0.182)-3.903(±1.868)	29	0.91	0.84	31.7	0.14	0.92
PCR	PIC_{50}= 0.245 PC1(±0.243) +0.13 PC3 (±0.043) +0.129 PC2(±0.043)-0.121 PC7(±0.043) + 8.695(±0.042)	29	0.77	0.75	14. 6	0.22	0.83
FA-MLR	PIC_{50}=11.953 MATS1m(±2.8) +2.65 Ms (±0.83) +2.61(±1.96)	29	0.67	0.53	13.2	0.12	0. 63
GA-MLR	PIC_{50}=2.072GGI7((±0.676)+4.427Ms((±0.678)+ 8.047BELm6(±1.753)-0.453Mor27u(±0.187)-14.411(±3.275)	29	0.94	0.88	28.5	0.16	0.91
GA-PLS	----------	29	0.94	0.80	-------	0.49	0.95

521

522

Table 4. Correlation coefficient (R2) matrix for descriptors represented in multiple linear regression

	MATS1m	Ms	GGI5	DELS	Mor25m
MATS1M	1	0.413	0.649	0.712	0.077
MS		1	0.345	0.668	0.250
GGI5			1	0.859	-0.125
DELS				1	0.079
Mor25M					1

Table 5. Numerical values of factor loading numbers 1–7for some descriptors after VARIMAX rotation (against DPP4 inhibitory activity).

	Component						
	1	2	3	4	5	6	7
volume	.524	.262	.073	.125	.214	-.110	-.325
Ms	.252	.792	.340	.019	-.208	.198	-.119
nH	.472	-.700	.088	.075	.369	-.295	.113
STN	-.163	.087	-.754	-.030	.464	.225	.058
DELS	.753	.448	.131	.132	.283	.005	.031
X0A	.335	.038	.881	.115	.084	-.127	.040
IVDE	.122	.190	.905	.185	.145	-.012	.150
IC0	.202	.925	.173	.205	.007	.043	.079
MATS1m	.849	.054	.267	-.049	.046	.041	-.155
GATS6m	-.179	-.646	-.430	-.251	-.120	.299	-.231
EEig09d	.670	.296	.117	.081	.372	-.295	.163
EEig13d	.252	-.104	-.570	-.063	.548	-.358	.176
GGI5	.598	.295	.270	.278	.508	-.093	.119
GGI4	-.054	.093	.319	.058	.785	.052	.239
JGT	.008	.319	.842	.136	.283	.035	.135
RDF015m	.874	-.157	-.142	-.121	.021	.311	.123
Mor04m	-.029	-.482	-.330	.096	-.442	.190	.180
Mor25m	.053	-.006	-.130	-.055	-.079	-.003	-.885
Mor19m	.243	.052	-.009	-.043	-.176	.905	.179
Mor18p	.383	-.212	.282	.034	-.226	-.699	.321
E2m	.822	.165	.140	.097	-.144	-.040	-.068
HATS3e	-.304	.081	.133	.023	-.752	.080	.171
R4e	.559	-.465	.027	.139	.355	-.399	.210
ALOGP2	-.075	-.110	-.117	-.975	.006	.043	-.062
TE1	.321	.162	.195	.900	.056	-.012	-.012
TPSA(Tot)	.206	.817	-.099	.148	.346	.065	-.014
F03[C-O]	.175	-.031	-.064	-.976	-.001	.027	-.011

Tabel 6. Leverage (*h*) of the external test set molecules for different models. The last row (*h**) is the warning leverage.

Molecular no	MLR	PCR	FA-MLR	GA-MLR	GA-PLS
5	0.13	0.07	0.049	0.37	1.09
17	0.26	0.10	0.052	0.26	0.3
19	0.13	0.20	0.048	0.20	0.65
21	0.23	0.15	0.052	0.19	0.30
23	0.089	0.25	0.047	0.32	1.0
25	0.076	0.065	0.052	0.137	0.37
28	0.129	0.23	0.047	0.114	0.52
29	0.24	0.095	0.048	0.100	0.34
31	0.3	0.105	0.048	0.10	0.48
h*	0.71	0.6	0.3	0.6	1.35