

# SWFB and GA Strategies for Variable Selection in QSAR Studies for the Validation of Thiazolidine-2,4-Dione Derivatives as Promising Antitumor Candidates

Vivek Asati<sup>1\*</sup>, Sanjay Kumar Bharti<sup>1</sup>, Ankita Rathore<sup>2</sup>, Debarshi Kar Mahapatra<sup>1\*</sup>

<sup>1</sup>Institute of Pharmaceutical Sciences, Guru Ghasidas Vishwavidyalaya (A Central University), Bilaspur - 495009, Chhattisgarh, India

<sup>2</sup>Buraydah College of Dentistry and Pharmacy, Buraidah, Al-Qassim, Saudi Arabia

## ABSTRACT

**Objective:** Thiazolidine-2,4-dione (TZD) are the well known anti-diabetic scaffold. Very recently, several TZD based anti-cancer agents have come into limelight for treating mutant cancer forms. In order to establish and understand the relationship of biological activity with that of physiochemical parameters associated with the structure, two-dimensional (2D-QSAR), group-based (G-QSAR), and three-dimensional (3D-QSAR) were performed which may be useful for (medicinal) chemists in selecting the most suitable substituent for the development of more potent, effective and selective TZD based anticancer agents in future. **Methods:** A series of TZD derivatives were subjected to 2D-QSAR, G-QSAR, and 3D-QSAR studies. The following studies were performed using partial least square regression, multiple regressions and k-nearest neighbor methodology coupled with various feature selection methods, viz. stepwise forward backward (SWFB) and genetic algorithm (GA) to derive QSAR models which were further validated for statistical significance and predictive capability by internal and external validation. **Results:** The results were expressed for both SWFB and GA consecutively. The statistically significant best 2D QSAR model has  $r^2=0.90$ ,  $0.89$  and  $q^2=0.86$ ,  $0.84$  with  $\text{pred}_r^2=0.87$ ,  $0.82$  for PLSR with whereas  $r^2=0.97$ ,  $0.91$  and  $q^2=0.95$ ,  $0.86$  with  $\text{pred}_r^2=0.86$ ,  $0.77$  were predicted for MLR. G-QSAR model has  $r^2=0.92$ ,  $0.81$  and  $q^2=0.90$ ,  $0.76$  with  $\text{pred}_r^2=0.77$ ,  $0.77$  for PLSR whereas  $r^2=0.92$ ,  $0.88$  and  $q^2=0.87$ ,  $0.71$  with  $\text{pred}_r^2=0.73$ ,  $0.87$  were predicted for MLR. The 3D-QSAR studies were performed by using of kNN-MFA approach; a leave-one-out cross-validated correlation coefficient  $q^2=0.85$ ,  $0.84$  and  $\text{pred}_r^2=0.94$ ,  $0.77$  were obtained. Contour maps using this approach showed that steric, electrostatic, and hydrophobic effects dominantly determine binding affinities. The docking study revealed the binding orientations of these inhibitors at active site amino acid residues (ARG281 and ARG 852) of PI3K $\alpha$  enzyme (PDB ID: 3ZIM). **Conclusion:** The present research represents an effort to recognize the necessary structural requirements of TZD derivatives to be potential anticancer agents.

**Key words:** G-QSAR, kNN-MFA, Thiazolidine-2,4-dione, 2D/3D QSAR, VLife MDS.

Submission Date: 15-11-2016;  
Revision Date: 03-01-2017;  
Accepted Date: 15-02-2017

DOI: 10.5530/ijper.51.3.72

**Correspondence:**  
Debarshi Kar Mahapatra,  
Institute of Pharmaceutical  
Sciences, Guru Ghasidas  
Vishwavidyalaya (A Central  
University), Bilaspur -  
495009, Chhattisgarh, India  
Phone: 8357821881  
E-mail: dkmbbsp@gmail.com

## INTRODUCTION

Heterocycles play an imperative role in the therapy of multi-drug resistant cancer of different origin. Particularly, the five-membered ring heterocycles like thiazoles which comprise of all three major components; carbon atoms, nitrogen atom, and sul-

fur atom, are having a high significance in medicinal chemistry.<sup>1,2</sup> Thiazolidine-2,4-dione (TZD), alone or in combination with other heterocyclic rings have been reported to exhibit pronounced anticancer activity. GSK1059615 and its analogs are PI3K $\alpha$



www.ijper.org

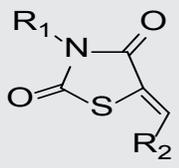
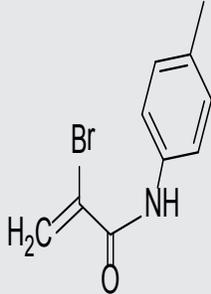
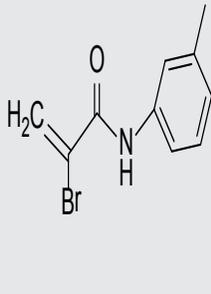
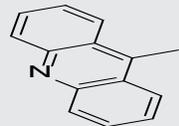
inhibitors containing TZD moiety, act as potential anticancer agents.<sup>3-4</sup> Recently, substituted-TZDs have been reported as PIM-2 enzyme inhibitors which have perspectives in the management of mutant cancer forms.<sup>5</sup> In various literature, 5-arylidene-thiazolidine-2,4-dione derivatives consisting of substituents at *para* or *meta* position of the benzylidene moiety have been identified to produce anticancer activity.<sup>6-8</sup> Xia *et al.* reported a series of 5-(3-trifluoromethylbenzylidene)thiazolidine-2,4-dione as the Pim-1 inhibitor and Barros *et al.* reported a series of 5-acridin-9-ylmethylene-3-benzylthiazolidine-2,4-dione analogs as moderate antiproliferative agents against a wide panel of cancer cell lines.<sup>9-10</sup> In past few years, several TZD derivatives have been designed and explored for their anti-cancer potentials. In the present study, QSAR analysis was performed for 24 previously synthesized 5-benzylidene thiazolidine-2,4-dione derivatives to establish a quantitative relation-

ship between the biological activity of the compounds with their physicochemical properties.<sup>10-11</sup> The present work represents an effort to generate the best predictive and validated 2D, 3D and G-QSAR models which may be helpful for the medicinal chemist in the development of novel anticancer inhibitors of TZD scaffold. The generated models may provide insights into the influence of various interactive fields on the biological activity. Therefore, it can help in designing and forecasting the inhibitory activity of novel anticancer agents.

## COMPUTATIONAL METHODS

### Dataset

Molecular modeling studies (2D, 3D, and G-QSAR) were performed using the VLife Molecular Design Suite.<sup>12</sup> Selected data set and their biological activity is shown in Table 1. Biological data presented as IC<sub>50</sub> (μM)

Table 1: Structure and anticancer activity (pIC <sub>50</sub> ) of thiazolidine-2,4-dione derivatives				
				
S No.	R <sub>1</sub>	R <sub>2</sub>	IC <sub>50</sub> (μM)	pIC <sub>50</sub> (M)
1	CH <sub>3</sub>		0.36	6.44
2	C <sub>2</sub> H <sub>5</sub>		0.87	6.06
3	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub>		2.3	5.64
4	C <sub>6</sub> H <sub>5</sub> (CH <sub>2</sub> ) <sub>2</sub>		4.2	5.38
5	4-F-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		2.4	5.62
6	3-F-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		3.2	5.49
7	4-Cl-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		2.8	5.55
8	3-Cl-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		3.7	5.43
9	4-CF <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		8	5.10
10	4-CH <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		3.7	5.43
11	4-C(CH <sub>3</sub> ) <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		0.83	6.08
12	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub>		0.51	6.29
13	C <sub>6</sub> H <sub>5</sub> (CH <sub>2</sub> ) <sub>2</sub>		0.94	6.03
14	4-F-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		0.68	6.17
15	3-F-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		0.69	6.16
16	4-Cl-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		1.4	5.85
17	3-Cl-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		1.4	5.85
18	4-CF <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		0.92	6.04
19	4-CH <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		1.2	5.92
20	4-C(CH <sub>3</sub> ) <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>		3	5.52
21	4-CH <sub>3</sub> -C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>			60.9
22	4-Br-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>	52.6		4.28
23	4-Cl-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>	58		4.24
24	4-F-C <sub>6</sub> H <sub>4</sub> CH <sub>2</sub>	60.3		4.22

were converted into  $\log(1/IC_{50})$  for computational work.

### Molecular modeling for 2D-QSAR

The molecular structures of all the 24 molecules were built using VLife MDS 4.4 software, in the 2D builder module. The structures were then transformed into 3D space for further analysis. All the compounds were batch optimized for energy minimization along with geometry using the standard Merck Molecular Force Field (MMFF) followed by allocating the distance-dependent dielectric constant of 1.0, convergence criterion or root-mean-square (RMS) gradient at 0.01 kcal/mol Å° and the iteration limit to 10,000.<sup>13-15</sup> The most stable structure for each compound was generated and used for the determination of different 2D descriptors like physicochemical and Baumann alignment-independent topological descriptors.<sup>16</sup> Preprocessing of the independent variables (i.e. descriptors) was done by removing the non-variables (descriptors that are constant for all the molecules), which resulted in 346 descriptors in the descriptor pool.

The energy-minimized geometry was used for the estimation of the different 2D descriptors including Individual, Chi, ChiV, Path count, ChiChain, ChiVChain, Chain path count, Cluster, Element Count, Estate number, Estate contribution, Hydrophilic-hydrophobic, Kappa, Path cluster, Polar surface area, and Semi-empirical. Alignment-independent descriptors attribute is "T-attribute" to thoroughly characterize the topology of the molecule. The second attribute is the atom type. In this study to calculate alignment independent descriptors, the following attributes have been used; the 2 (double bonded atom), 3 (triple bonded atom), T (any), C, N, O, S, H, F, Cl, Br and I at a distance of 0-7 was considered as independent variables.

#### Creation of training and test set for 2D-QSAR

The data set is randomly divided into two subsets: a training set for the dramatization of QSAR model and a test set for authenticating the superiority of QSAR model. Data set included six compounds, namely, 10, 13, 17, 9, 22 and 23 were used as test set for model 1, 2 and 3 and compounds namely, 10, 17, 18, 8, 22 and 23 were used as test set for model 4 while remaining all 18 molecules were used as the training set (Table 1). In order to evaluate the resemblance of the distribution pattern of the compounds in the generated sets, statistical parameters (with respect to the activity), i.e. mean and standard deviation were calculated for both test and training sets (Table 2). In order to make the prediction statistics to be more reliable, the test set comprised of minimum five compounds.<sup>17</sup>

**Table 2: Unicolumn statistics of the training and test sets for thiazolidine-2,4-dione derivatives in different models**

Data set	Average	Max.	Min.	SD	Sum
<b>2D-QSAR Model 1, 2 and 3</b>					
Training	5.67	6.44	4.22	0.62	102.09
Test	5.15	6.03	4.24	0.77	30.93
<b>2D-QSAR Model - 4</b>					
Training	5.65	6.44	4.22	0.63	101.75
Test	5.21	6.04	4.24	0.78	31.27
<b>G-QSAR Model - 5</b>					
Training	5.53	6.44	4.22	0.70	99.54
Test	5.58	6.16	4.28	0.69	33.47
<b>G-QSAR Model - 6</b>					
Training	5.60	6.44	4.22	0.62	100.86
Test	5.36	6.16	4.24	0.88	32.16
<b>G-QSAR Model - 7</b>					
Training	5.53	6.44	4.22	0.70	99.54
Test	5.58	6.16	4.28	0.69	33.47
<b>G-QSAR Model - 8</b>					
Training	5.54	6.44	4.22	0.69	99.79
Test	5.54	6.29	4.28	0.70	33.22
<b>3D-QSAR Model - 9</b>					
Training	5.59	6.44	4.22	0.60	100.70
Test	5.39	6.29	4.22	0.93	32.32
<b>3D QSAR Model - 10</b>					
Training	5.50	6.44	4.22	0.67	99.01
Test	5.67	6.29	4.24	0.75	34.00

Molecular modeling for group-based QSAR (G-QSAR) A recent QSAR method known as Group-based QSAR (G-QSAR) deals with the problems related to QSAR model interpretation and the inverse QSAR problems. The method assists in establishing the relationship of chemical groups or fragment variation at different molecular sites with the biological activity. Therefore, the method deals with the chemical groups or molecular fragments instead of the whole molecule. G-QSAR method comprises of three steps: (a) Generation of molecule fragments using a set of predefined chemical rules, (b) Calculation of descriptors for the generated fragments, and (c) Build statistical models using the calculated fragment descriptors and their interactions.<sup>18</sup> The fragment descriptors and their interactions are related to the biological activity, resulting in model (s) that highlight important site (s) for substitution along with probable interactions and chemical nature. The suggested important fragments can be used as the building blocks to design novel molecules.<sup>19</sup> For G-QSAR analysis, a range of 2D descriptors was calculated for diverse

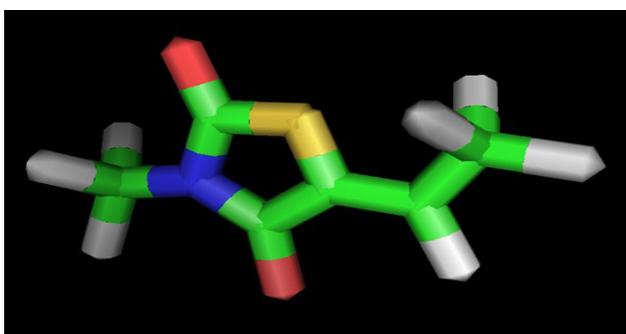
groups present at different sites for substitution in the compounds (i.e., Fragments  $R_1$  and  $R_2$ ). The removal of the invariable group descriptors resulted in a total of 468 group descriptors which can be used further. Since the same descriptors are calculated for different groups at diverse sites, the following nomenclature is used for naming a descriptor at a particular position, for example,  $R_1$ -XAMostHydrophobic descriptor signifies most hydrophobic value on the van der Waals (vdW) surface present at substitution site  $R_1$ .<sup>19</sup>

#### Creation of training and test set for G-QSAR

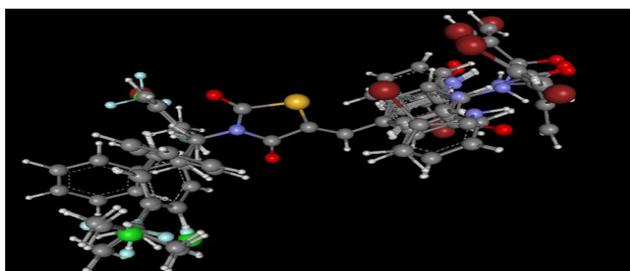
In G-QSAR data set, selected randomly given best result included six compounds, namely, 15, 18, 19, 20, 7 and 22 for model 5; 15, 18, 19, 20, 22 and 23 for model 6; 15, 18, 19, 7, 22 and 20 for model 7; 12, 2, 20, 3, 8 and 22 for model 8 were used as test sets while remaining 18 molecules were used as the training sets.

#### Molecular modeling for 3D-QSAR

For 3D-QSAR study, molecular alignment is a crucial step to obtain meaningful results. In the present study, using the template-based method, the compounds of the dataset were aligned in VLife MDS 4.4 software.<sup>20</sup> A template structure is defined and used for the alignment of molecules whereas a reference molecule is chosen on which the other molecules get aligned based on the chosen template. The template structure, i.e. 5-ethylidene-3-methyl-1,3-thiazolidine-2,4-dione was used for the alignment. This template was selected by taking some common elements of the series and is shown in Figure 1. The reference molecule, compound 1 has been cho-



**Figure 1: 5-ethylidene-3-methyl-1,3-thiazolidine-2,4-dione ring as a template.**



**Figure 2: Template based alignment of all the 24 thiazolidine-2,4-dione derivatives.**

sen in such a way that it is the most active among the series of molecules considered. The superimposition of molecules based on the minimizing root mean square deviation (RMSD) is shown in Figure 2.

The results of Molecular field analysis (MFA) may provide predictive and sufficiently reliable information to the researchers in design and development of novel anticancer agents. This approach is effective for the analysis of data sets, where information about the biological activity is reported but the structure of the receptor site is unknown. It attempts to assume and signifies the vital elements of a receptor site from the aligned common features of the molecules that participates in the interaction. The MFA computes the probe interaction energies on a rectangular grid around a bundle of active molecules. In the contributing models, the atomic coordinates were used to compute field values on each point of a 3D grid. The fields of molecules were symbolized using lattice framework or grid and the related energy with each grid point serve as an input for the calculation of 3D-QSAR. These obtained energies were added to the table to form new columns head according to the probe type. The molecular field was produced with methyl group as a probe, which represents steric, electrostatic and hydrophobic fields, respectively. For the determination of field descriptor values; electrostatic, steric and hydrophobic fields with cutoff values were selected as 10.0 and 30.0 kcal/mol, respectively, whereas the charge was selected as Gasteiger and Marsili.<sup>21</sup> The dielectric constant was set as 1.0 considering the distance-dependent dielectric function. Probe setting was a carbon atom with charge 1.0. This resulted in the calculation of 3,028 field descriptors (electrostatic, steric and hydrophobic) for all the molecules in different columns after excluding descriptors having zero values or same values. Data set selected manually given the best result in which six compounds, namely, 12, 15, 16, 7, 24 and 23 for model 9 and 12,17,18, 2, 20 and 23 for model 10 were used as test set while the remaining 18 molecules as the training set (Table 1). The unicolon statistics has been given in Table 2.

#### Feature selection and model development

The different statistical methods were used to generate 2D and G-QSAR models are multiple linear regressions (MLR) and partial least squares regression (PLSR) coupled with various feature selection methods, viz. stepwise forward backward (SWFB) and genetic algorithm (GA) which were further validated for statistical significance and predictive ability by internal and external validation.<sup>22</sup>

In 2D and G-QSAR some parameters which have been selected for best QSAR models included cross-correla-

tion limits at 1.0, variance cutoff at 0.0, with auto scaling in which the number of random iterations were fixed at 100. The descriptor subset size was increased for model establishment until no improvement was observed. It must be kept in view that the number of molecules in the training set should not be smaller than five times the descriptor frequency. The variance–covariance matrices were calculated for each of the descriptors in all of the resulting linear models. The descriptors which had multi collinearity were discarded. The tolerance and Variation Inflation Factor (VIF) were selected as the parameter for the estimation of colinearity among the variables. A VIF value in range of 1-4 indicates non-collinearity between the variables. Among the remaining models after the elimination process, the one that had the minimum SEE was chosen as the best. To establish the significance of QSAR models, following statistical parameters were considered: squared correlation coefficient ( $r^2$ ), F-test (F-test for statistical significance of the model), cross-validated squared correlation coefficient ( $q^2$ ) and predicted correlation coefficient ( $pre\_r^2$ ).

The 3D-QSAR was executed using k-Nearest Neighbor Molecular Field Analysis (kNN-MFA) methodology which relies on a distance learning approach in which an unidentified compound is classified according to the majority of its k-nearest neighbors in the training set.<sup>20</sup> The nearness is measured by an appropriate distance metrics (e.g. the molecular similarity was measured using field interactions of molecular structures). The standard kNN method is implemented simply as follows: (i) calculate distances between an unknown object (u) and all the objects in the training set. (ii) according to the calculated distances, selecting the k objects from the training set which are quite similar to object u. (iii) classifying the object u with the group in which the maximum of the k objects belongs.<sup>23</sup> An optimal k value is selected by optimization through the classification of the test set samples or using leave-one-out cross (LOO) validation. In many cases setting of k to 1 provides reasonably good predictive performance for classification purposes. In general, optimal values of k are acquired by the trial and error method. Another approach is to use a cross-validation scheme to obtain the best value of k for a given dataset. The kNN-MFA models were developed using the SWFB method with the cross-correlation limit set to 1.0 and the term selection criterion as  $q^2$ . F test 'in' was set to 4.0, and F test 'out' to 3.99. Few supplementary parameters like setting the variance cutoff at 0.0, and scaling to autoscaling were performed. Additionally, setting the kNN parameter within the limits 2–5 along with the selection of prediction method named distance-based weighted average.

In SWFB variable selection algorithm, the search procedure begins with developing a trial model step by step with an independent variable and in every step, the independent variables are added one at a time, examining the fit of the model. In this method, the model is repetitively changed from the previous one by adding or removing a predictor variable according to the stepping criteria (where  $F = 4$  for inclusion;  $F = 3.99$  for exclusion for the forward–backward selection method).<sup>20</sup> The procedure is carried out until there is no more significant variable remaining outside the model.

In the QSAR studies, the Genetic Algorithm (GA) represents an extensively used method for variable selection.<sup>24</sup> It is a group of algorithms derived by the progression of natural evolution in which species having a high fitness under some conditions can overcome and carry on to the subsequent generations. In this process, the finest species can be adapted by crossover and/or mutation in the search for superior characters. The method represents that a chromosome and its fitness in the species embodies a set of molecular descriptors along with the cross-validated predictive accuracy of the derived QSAR model, respectively.

For constructing QSAR models the feature selection methods used in conjunction with multiple linear regression (MLR), partial least square (PLS) and kNN regression analysis. MLR is the standard method for multivariate data analysis. By the application of least squares curve fitting method, it estimates the values of the regression coefficients. For getting reliable results, usually the dataset comprises of 5 times as many molecules as independent variables (descriptors) is required. PLS method attributes in simplification of regression, which can handle data with strongly correlated and/or noisy or numerous X variables which are statistically robust several times as compared to the MLR. The linear PLS model finds “new variables” (latent variables or X scores) that consist of linear combinations of variables. To avoid over-fitting, a strict test for the significance of every successive PLS constituent is needed and subsequently stopping when the components are non-significant. Cross-validation is a very convenient and

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2} \quad 1$$

dependable process for testing the significance. PLS is normally used in combination with cross-validation to obtain the optimum number of components.<sup>25</sup>

## Model evaluation and validation

The model evaluation and validation required to test

$$\text{Pred}_r^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2} \quad 2$$

the internal stability and predictive ability of the QSAR models. Leave-one-out (LOO) method has been applied for internal validation of data set. The cross-validated coefficient,  $q^2$ , was calculated using the following Eq. 1: where  $y_i$  and  $\hat{y}_i$  are the actual and predicted activity of the  $i$ th molecule in the training set and  $y_{\text{mean}}$  is the average activity of molecules in the training set. For external validation, the activity of molecule in the test set was predicted using the model generated from the training set. The  $\text{pred}_r^2$  value is calculated by using Eq. 2.

where  $y_i$  and  $\hat{y}_i$  are the actual and predicted activity of the  $i$ th molecule in the test set and  $y_{\text{mean}}$  is the average activity of molecules in the training set.

$$Z_{\text{score}} = \frac{q_{\text{org}}^2 - q_{\text{a}}^2}{q_{\text{std}}^2} \quad 3$$

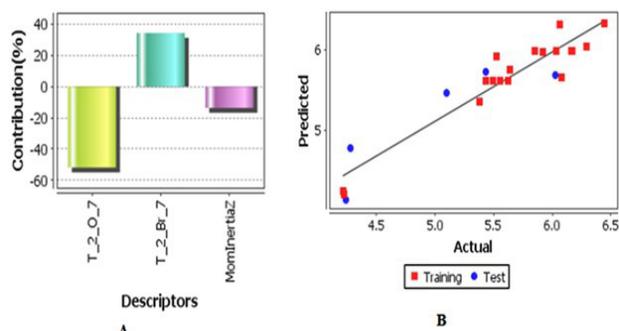
In an order to determine the robustness of the QSAR models, the experimental training sets were examined by comparing these models to those derived from random data sets (generated by rearranging biological activities of the training set molecules). The significance of the obtained models was derived based on the calculated Z-score. Z-score (Eq. 3) can be defined as the absolute difference between the value of the model and the activity field, divided by the square root of the mean square error of the dataset.

However, a QSAR model is considered to be predictive, if the following conditions are satisfied:  $r^2 > 0.6$ ,  $q^2 > 0.6$ , and  $\text{pred}_r^2 > 0.5$ .

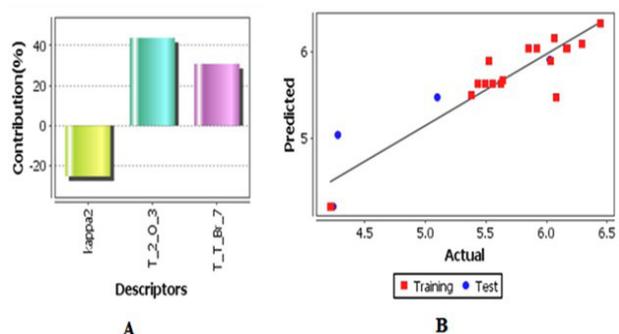
## Molecular docking

The molecular docking studies were performed using AutoDock Vina Software which provided improves average accuracy of the binding mode predictions. It has been tested against a virtual screening benchmark called the Directory of Useful Decoys by the Watowich group, and the result outcomes were found to be several times better than other software. Some important attributes for ligand and macromolecular receptor are necessary before carrying out the docking study are as follows.<sup>26,27</sup> For ligand (i) Add all hydrogens, compute Gasteiger charges, and merge non-polar H; (ii) Ensure total charge corresponds to tautomeric state; (iii) Choose torsion tree

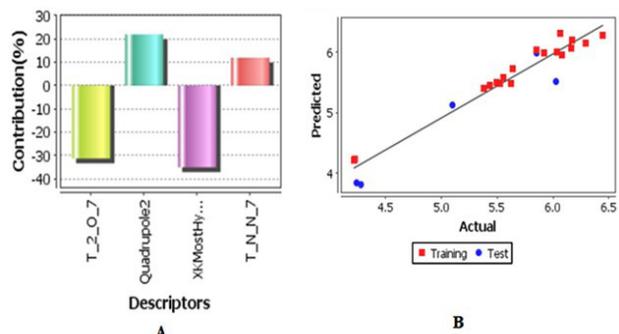
root & rotatable bonds. For macromolecule (i) Add all hydrogens, compute Gasteiger charges, and merge non-polar H; (ii) Assign Stouten atomic solvation parameters; (iii) Optionally, create flexible residues PDBQT in addition to the rigid PDBQT file. The final estimation was performed with AutoDock Score (docking score). Each docking produced multiple conformations along



**Figure 3 A: Contribution plot between selected descriptors for model 1 B. Fitness plot between actual and predicted activities for model 1.**



**Figure 4 A: Contribution plot between selected descriptors for model 2 B. Fitness plot between actual and predicted activities for model 2.**



**Figure 5 A: Contribution plot between selected descriptors for model 3 B. Fitness plot between actual and predicted activities for model 3.**

with corresponding binding energy scores which were computed using AutoDock scoring function.<sup>28</sup> The conformations were ranked based on the scores; a lower scoring conformation was ranked higher. An RMSD value was also computed based on the conformation of the bound ligand (true conformation) and the docked conformation of the ligand. The RMSD value measures

the distance between the docked conformation and the true conformation.<sup>29,30</sup> Auto Dock generated different file format were loaded in Python Molecular Viewer (PyMOL) and visualized diverse binding site attributes provide valuable insights for structure-based drug design (SBDD). The RMSD values lower than 2 Å were judged to be a successful docking.<sup>31</sup>

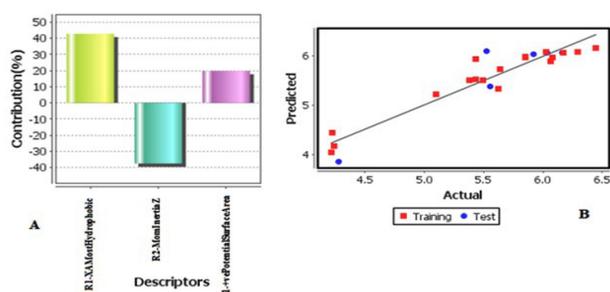
## RESULTS AND DISCUSSION

QSAR study of a series of TZD derivatives was performed by using VLife MDS 4.4 software. In this study, biological activity ( $pIC_{50}$ ) as dependent and various physiochemical with topological descriptors taken as the independent variable and regression were established

using PLS and MLR analysis. The models were selected on the basis of its statistical significance for further study. A data set of 24 compounds were divided into training (18 molecules) and test sets (6 molecules) using random selection method. Selection of molecules in the training and test set is a key and significant trait for any QSAR analysis. For that reason, it was of such concern that the biological activities of all compounds in test set lie within the maximum and minimum value range of biological activities of the training set of compounds. A Uni-Column statistics for test and training set was generated to check the correctness of selection criteria for both set of molecules (Table 2). After regression analysis, the best equations were selected, on the basis of  $r^2$ ,  $q^2$  and  $pred\_r^2$  values (Table 3 and 4).

**Table 3: Statistical results of best 2D-QSAR models of thiazolidine-2,4-dione derivatives**

S No.	Statistical parameter	2D-QSA			
		(Model 1) SWFB	(Model 2) GA	(Model 3) SWFB	(Model 4) GA
1	$r^2$	0.90	0.89	0.97	0.91
2	$q^2$	0.86	0.84	0.95	0.86
3	$pred\_r^2$	0.87	0.82	0.86	0.77
4	$r^2\_se$	0.21	0.22	0.13	0.21
5	$q^2\_se$	0.25	0.27	0.17	0.26
6	$pred\_r^2se$	0.35	0.40	0.86	0.44
7	F_test	68.84	57.74	98.50	45.43
8	Zscore	6.88	5.69	7.14	6.32
9	Best-rand- $q^2$	0.19	0.57	0.09	0.52
10	$n_{training}$	18	18	18	18



**Figure 6 A : Contribution plot between selected descriptors for model 4 B. Fitness plot between actual and predicted activities for model 4.**

**Table 4: Statistical results of best G-QSAR and 3D models of thiazolidine-2,4-dione derivatives**

S No.	Statistical parameter	G-QSAR				3D-QSAR	
		(Model 5) SWFB	(Model 6) GA	(Model 7) SWFB	(Model 8) GA	(Model 9) SWFB	(Model 10) GA
1	$r^2$	0.92	0.81	0.92	0.88	-	-
2	$q^2$	0.90	0.76	0.87	0.71	0.85	0.84
3	$pred\_r^2$	0.77	0.77	0.73	0.87	0.94	0.77
4	$r^2\_se$	0.20	0.29	0.22	0.27	-	-
5	$q^2\_se$	0.23	0.32	0.28	0.41	0.23	0.27
6	$pred\_r^2se$	0.33	0.44	0.36	0.25	0.24	0.37
7	F_test	68.84	31.43	54.71	33.22	-	-
8	Zscore	7.06	4.94	6.95	4.54	-	-
9	Best-rand- $q^2$	0.52	0.72	0.31	0.70	-	-
10	$n_{training}$	18	18	18	18	18	18

## Interpretation of 2D-QSAR

The 2D-QSAR study of 24 compounds (divided into 6 test and 18 training) for anticancer activity through PLS (models 1 and 2) and MLR (models 3 and 4) analysis coupled with SW-FB and GA, variable selection resulted in some statistical models summarized in Table 5. The inter-correlation matrix between four descriptors with the biological activity for the models 1, 2, 3 and 4 are presented in Table 6. Comparative observed and predicted activities of TZD derivatives by best 2D QSAR models are presented in Table 7. The contribution of descriptor plot for models 1, 2, 3 and 4 are depicted in Figures 3A, 4A, 5A, and 6A. The graph of actual versus predicted activity for the PLSR and MLR analysis are given in Figures 3B, 4B, 5B, and 6B. The present QSAR model reveals that Baumann's alignment-independent topological descriptor has a major role in the explanation of variance in activity. A descriptor  $T\_X\_Y\_Z$  can be defined as a count of fragments formed with atom types X and Y separated by distance of Z bonds. The definitions for the descriptors that were found to be dominating in the developed 2D-QSAR models are given below.

Mom Inertia Z : This descriptor suggests the moment of inertia at Z-axis.

kappa2: This descriptor signifies second kappa shape index:  $(n-1)^2 / m^2$

XKMostHydrophilic: Most hydrophilic value on the vdW surface.

Quadrupole2: This descriptor signifies magnitude of second tensor of quadrupole moments.

XAMostHydrophilic: Most hydrophilic value on the vdW surface.

chi3Cluster: This descriptor denotes 3rd order cluster chi index in a compound.

$T\_2\_O\_7$ : This is the count of number of double bounded atoms (i.e. any double bonded atom,  $T\_2$ ) separated from oxygen atom by 7 bonds in a molecule.

$T\_2\_Br\_7$ : This is the count of number of double bounded atoms (i.e. any double bonded atom,  $T\_2$ ) separated from bromine atom by 7 bonds in a molecule.

$T\_2\_O\_3$ : This is the count of number of double bounded atoms (i.e. any double bonded atom,  $T\_2$ ) separated from oxygen atom by 3 bonds in a molecule.

$T\_T\_Br\_7$ : This descriptor means the count of pair of any heavy atom and any bromine atom separated by 7 bond distance.

$T\_N\_N\_7$ : This is the count of number of Nitrogen atoms (single double or triple bonded) separated from any other Nitrogen atom (single double or triple bonded) by 7 bonds in a molecule.

The careful examination of the descriptors in the best model 1 suggests that alignment-independent descriptor  $T\_2\_O\_7$  is the most influential for the anticancer activity and negative contribution (~50%) of it showed that increase in the values of this descriptor would be detrimental for the anticancer activity of TZD derivatives. The descriptor in the best model 2 suggests that alignment-independent descriptor  $T\_2\_O\_3$  with a positive contribution (~42%) showed enhancement in the values of the descriptor seems to be beneficial for activity. The descriptor in the best model 3 suggests  $T\_2\_O\_7$  (~32%) and XKMostHydrophilic (~32%) showed negative contribution but Quadrupole2 (~21%) and  $T\_N\_N\_7$  (~11%) showed positive contribution for anticancer activity. The descriptor in the best model 4 suggests  $T\_T\_Br\_7$  (~30%) showed positive contribution but chi3Cluster (~20%) and XAMostHydrophilic (~50%) showed negative contribution for anticancer activity.

## Interpretation of G-QSAR

On the basis of obtained information of 2D descriptors, the models do not exactly indicate the part of the molecule where the alterations are necessary to enhance the activity, thus posing a hurdle in the complete structural interpretation. As a result, in a way to get approach to the crucial molecular part(s), in terms of their chemical information responsible for the disparity in the biological activity, the fragment descriptors and their interactions (cross terms) were developed in G-QSAR models. Table 4 reports the statistical parameters ( $r^2$ ,  $q^2$  and  $pred\_r^2$ ) and Table 5 reports the best G-QSAR models namely 5, 6, 7 and 8. Contributions of descriptors in the development of models are depicted in Figures 7A, 8A, 9A and 10A. The graph of actual versus predicted activity for the PLSR and MLR analysis are given in Figures 7B, 8B, 9B and 10B. Inter-correlation matrix between the descriptors with the biological activity for the models 5, 6, 7 and 8 are presented in Table 6. The comparative observed and predicted activities of TZD derivatives by best G-QSAR models are presented in Table 8. The definitions for the descriptors that were found to be dominating in the developed G-QSAR models are given below.

$R_1$ -XAMostHydrophobic: Most hydrophobic value on the vdW surface in part-  $R_1$ .

$R_2$ -MomInertia Z : This descriptor signifies moment of inertia at Z-axis in part-  $R_2$ .

$R_1$ -+vePotentialSurfaceArea: This descriptor signifies total van der Waals surface area with positive electrostatic potential of the molecule in part-  $R_1$ .

$R_2$ -Epsilon3: Measure of electronegative atom count (including the hydrogen atoms) with respect to the saturated hydrocarbon created from the molecule/fragment under consideration in part-  $R_2$ .

$R_1$ -polarizabilityAHC: This descriptor evaluates molecular polarizability using sum of atomic polarizabilities using the atomic hybrid component (AHC) in part-  $R_1$ .

$R_1$ -T\_C\_O\_6: This is the count of the number of carbon atoms (single, double or triple bonded) separated from oxygen atom (single or double bonded) by 6 bond distance in a molecule in part-  $R_1$ .

$R_1$ -SKMostHydrophilic: Most hydrophilic value on the vdW surface in part-  $R_1$ .

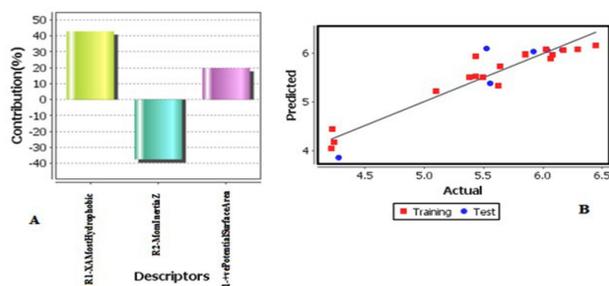
$R_2$ -XAAverage: Average hydrophobicity function value in part-  $R_2$ .

The most contributing descriptor in part-  $R_1$  with positive coefficient value are XAMostHydrophobic and +vePotentialSurfaceArea suggests that increase in Hydrophobicity or positive electrostatic potential of fragment  $R_1$  may lead to an increase in the anticancer activity. The descriptors polarizabilityAHC, SKMostHydrophilic and T\_C\_O\_6 were found to be inversely relative to the biological activity. This indicates that decrease in hydrophilic value, polarizability and number of carbon atoms (single, double or triple bonded) separated from any oxygen atom of fragment  $R_1$  may lead to an increase in the biological activity.

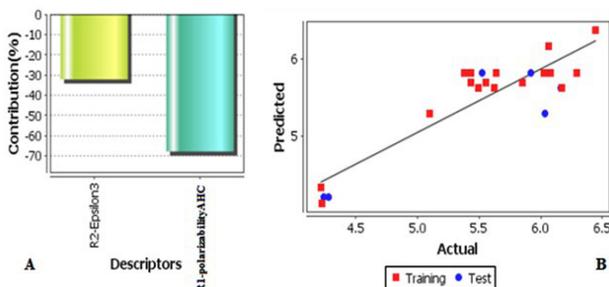
The most contributory descriptors in part- $R_2$  with negative coefficient values are MomInertiaZ, Epsilon3 and XA Average which signifies an increase in moment of inertia at Z-axis, electronegative atom count and hydrophobicity of fragment  $R_2$  may lead to a decrease in anticancer activity.

Interpretation of 3D-QSAR

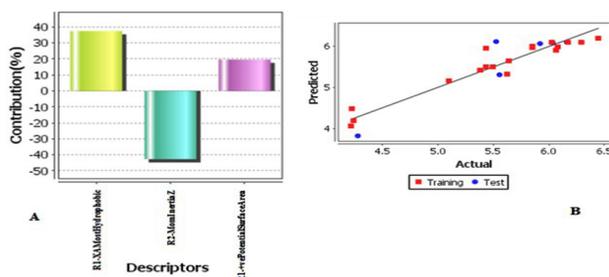
The 3D-QSAR studies were performed by using of k-nearest neighbor molecular field analysis approach with SWFB and GA variable selection methods. The model 9 generated through SWFB showed good statistical parameters including  $q^2$  value of 0.85 and  $\text{pred}_r^2$  value of 0.94 (Table 4). The model 10 generated through GA has a  $q^2$  value of 0.84 and  $\text{pred}_r^2$  value of 0.77 (Table 4). The careful analysis of comparison of biological activities ( $\text{pIC}_{50}$ ), predicted activities for training and test set molecules presented in Table 8. Therefore it may be said that the predictive abilities of both models are good. The results are supported by the statistical values presented in Table 4. The graph of actual versus predicted activity for the series is mentioned in Figure 11 for model 9. For model 10, the graph of actual versus predicted activity is mentioned in Figure 12. Inter-correlation matrix between descriptors with the biological



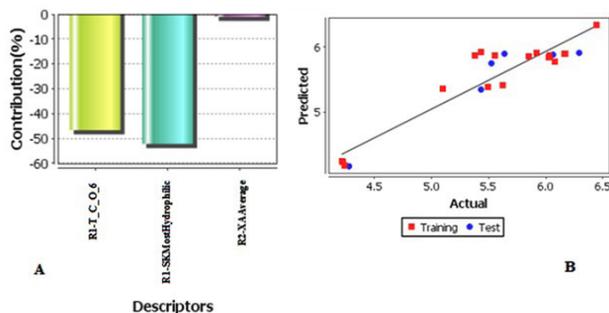
**Figure 7 A: Contribution plot between selected descriptors for model 5 B. Fitness plot between actual and predicted activities for model 5.**



**Figure 8 A: Contribution plot between selected descriptors for model 6 B. Fitness plot between actual and predicted activities for model 6.**



**Figure 9 A: Contribution plot between selected descriptors for model 7 B. Fitness plot between actual and predicted activities for model 7.**



**Figure 10 A: Contribution plot between selected descriptors for model 8 B. Fitness plot between actual and predicted activities for model 8.**

activity for the models 9 and 10 is presented in Table 6. In 3D-QSAR studies, 3D data points generated around TZD pharmacophore for model 9 are  $S_{1324}$  (-0.1464 -0.1446) and  $E_{1018}$  (-10.0000 -10.0000) that is, steric and electrostatic interaction fields at lattice points 1324 and 1018, respectively have been provided in Figure 13 and 14. This helps in identification of varied molecular characteristics responsible for activity variation and

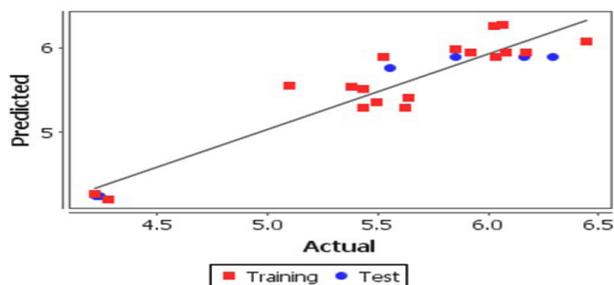


Figure 11: Fitness plot between actual and predicted activities for model 9

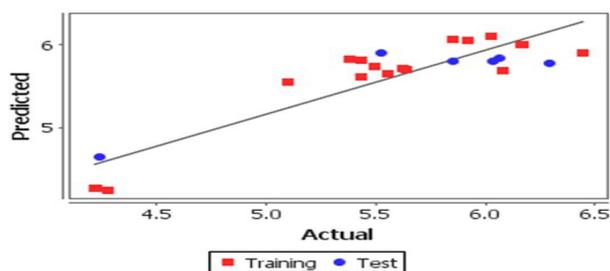


Figure 12: Fitness plots between actual versus predicted activities for model 10 of 3D-QSAR studies. Here red points showed training set of molecule whereas blue points showed test set of molecules.

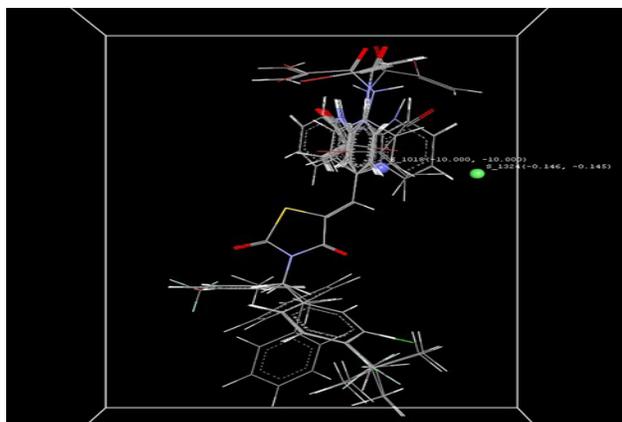


Figure 13: In 3D-QSAR studies, 3D data points generated around thiazolidine-2,4-dione pharmacophore for model 9 are S\_1324 (-0.1464 -0.1446) and E\_1018 (-10.0000 -10.0000) represent steric and electrostatic interaction fields at lattice points 1324 and 1018 for all 24 molecules.

hence aid in the design of novel anticancer agents. It is observed that electrostatic descriptors like E\_1018 with negative coefficients indicating that electronegative groups are favorable on this site and would lead to increase in the anticancer activity of these compounds. The steric descriptor S\_1324 with negative coefficients indicates that bulky groups are unfavorable on this site and presence of bulky groups would decrease the anticancer activity. In 3D-QSAR studies, another model 10 showed 3D data points created around the TZD pharmacophore are S\_1511 (-0.0250 -0.0231), S\_486 (-0.0054 -0.0051) and S\_1328 (-0.1516 -0.1136) that is, steric field at lattice points 1511, 486, and 1328 which

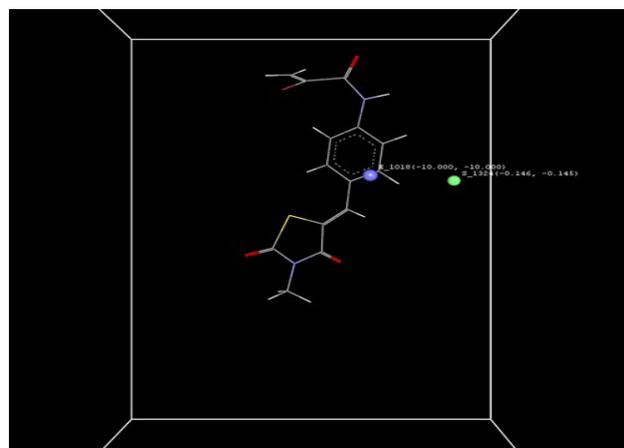


Figure 14: In 3D-QSAR studies, 3D data points generated around thiazolidine-2,4-dione pharmacophore for model 9 are S\_1324 (-0.1464 -0.1446) and E\_1018 (-10.0000 -10.0000) represent steric and electrostatic interaction fields at lattice points 1324 and 1018 for most active molecule

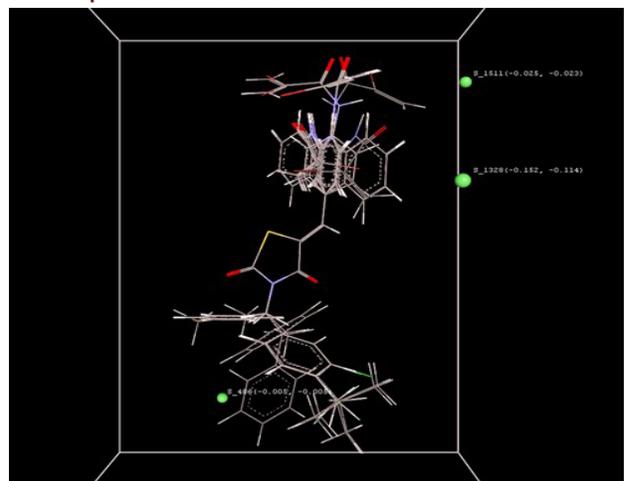


Figure 15 In 3D-QSAR studies, 3D data points generated around thiazolidine-2,4-dione pharmacophore for model 10 are S\_1511 (-0.0250 -0.0231), S\_486 (-0.0054 -0.0051) and S\_1328 (-0.1516 -0.1136) represent steric fields at lattice points 1511, 486 and 1328 for all 24 molecules.

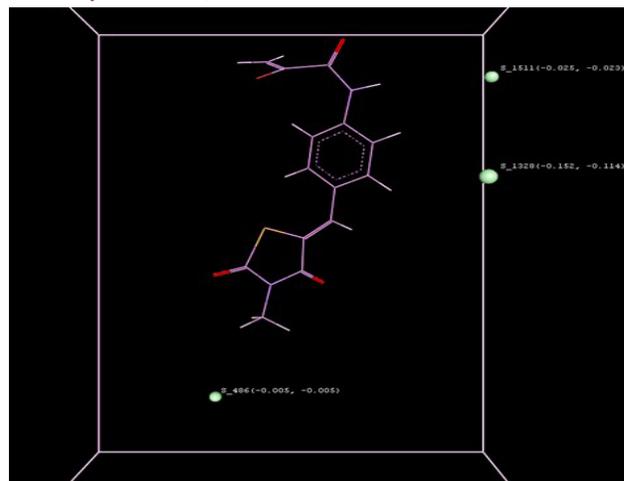


Figure 16: In 3D-QSAR studies, 3D data points generated around thiazolidine-2,4-dione pharmacophore for model 10 are S\_1511 (-0.0250 -0.0231), S\_486 (-0.0054 -0.0051) and S\_1328 (-0.1516 -0.1136) represent steric fields at lattice points 1511, 486 and 1328 for most active molecule.

**Table 5: List of predictive 2D and G-QSAR models generated from various regression methods**

Model No.	Method	Equation
01	2D/Random/ PLSR/SWFB	$IC_{50} =$
		$- 0.2172 T\_2\_O\_7$
		$+ 0.3996 T\_2\_Br\_7$
		$- 0.0000 MomInertiaZ$
		$+ 7.3995$
02	2D/Random/ PLSR/GA	$IC_{50} =$
		$-0.2463 \text{ kappa}2$
		$0.6039 T\_2\_O\_3$
		$0.4099 T\_T\_Br\_7$
		$+4.1417$
03	2D/Random/ MLR/SWFB	$IC_{50}$
		$- 0.1999(\pm 0.0017) T\_2\_O\_7$
		$+ 0.0115(\pm 0.0000) \text{ Quadrupole}2$
		$- 2.3828(\pm 0.1949) \text{ XKMostHydrophilic}$
		$+ 0.2736(\pm 0.0357) T\_N\_N\_7$
04	2D/Random/ MLR/GA	$IC_{50}$
		$- 1.0142(\pm 0.2787) \text{ chi}3\text{Cluster}$
		$- 26.6769(\pm 3.1982) \text{ XAMostHydrophilic}$
		$+ 0.3740(\pm 0.0033) T\_T\_Br\_7$
		$-1.4075$
05	GQSAR/ Random/ PLSR/SWFB	$IC_{50} =$
		$+ 4.9606 R1\text{-XAMostHydrophobic}$
		$- 0.0000 R2\text{-MomInertiaZ}$
		$+ 0.0222 R1\text{-+vePotentialSurfaceArea}$
		$+ 1.9003$
06	GQSAR/ Random/ PLSR/GA	$IC_{50} =$
		$- 10.5913 R2\text{-Epsilon}3$
		$- 0.3304 R1\text{-polarizabilityAHC}$
		$+ 16.8743$
07	GQSAR/ Random/ MLR/SWFB	$IC_{50}$
		$+ 4.4004(\pm 0.7336) R1\text{-XAMostHydrophobic}$
		$- 0.0000(\pm 0.0000) R2\text{-MomInertiaZ}$
		$+ 0.0223(\pm 0.0001) R1\text{-+vePotentialSurfaceArea}$
		$+ 2.1259$
08	GQSAR/ Random/ MLR/GA	$IC_{50}$
		$- 13.3617(\pm 0.0571) R1\text{-T\_C\_O\_6}$
		$- 59.2870(\pm 0.8982) R1\text{-SKMostHydrophilic}$
		$- 3.4045(\pm 1.8355) R2\text{-XAAverage}$
		$+ 37.8118$

**Table 6: Correlation matrix for descriptors used in different models**

<b>Model-1</b>				
	T_2_O_7	T_2_Br_7	Mom Inertia Z	
T_2_O_7	1			
T_2_Br_7	-0.58	1		
MomInertiaZ	0.049	0.55	1	
<b>Model-2</b>				
	T_2_O_3	T_T_Br_7	kappa2	
T_2_O_3	1			
T_T_Br_7	0.69	1		
kappa2	0.24	0.39	1	
<b>Model-3</b>				
	T_2_O_7	Quadrupole2	XK Most Hydrophilic	T_N_N_7
T_2_O_7	1			
Quadrupole2	0.48	1		
XK Most Hydrophilic	0.87	0.66	1	
T_N_N_7	0.37	0.65	0.60	1
<b>Model-4</b>				
	chi3Cluster	XA Most Hydrophilic	T_T_Br_7	
chi3Cluster	1			
XA Most Hydrophilic	-0.16	1		
T_T_Br_7	0.32	-0.61	1	
<b>Model-5</b>				
	R1-XA MostHydrophobic	R2-Mom Inertia Z	R1-+vePotential Surface Area	
R1-XAMostHydrophobic	1.00			
R2-MomInertiaZ	-0.67	1.00		
R1-+vePotentialSurfaceArea	0.49	0.07	1.00	
<b>Model-6</b>				
	R2-Epsilon3	R1-polarizability AHC		
R2-Epsilon3	1.00			
R1-polarizabilityAHC	0.08	1.00		
<b>Model-7</b>				
	R1-XAMostHydrophobic	R2-MomInertiaZ	R1-+vePotentialSurfaceArea	
R1-XAMostHydrophobic	1			
R2-MomInertiaZ	0.49	1		
R1-+vePotentialSurfaceArea	0.07	-0.67	1	
<b>Model-8</b>				
	R1-T_C_O_6	R1-SKMostHydrophilic	R2-XAAverage	
R1-T_C_O_6	1.000			
R1-SKMostHydrophilic	-0.99	1.00		
R2-XAAverage	-0.03	0.03	1.00	
<b>Model-9</b>				
	E_1018	S_1324		
E_1018	1.00			
S_1324	0.06	1.00		

Model-10				
	S_486	S_1511	S_1328	
S_486	1.00			
S_1511	0.15	1.00		
S_1328	0.13	0.61	1.00	

**Table 7: Comparative observed and predicted activities (LOO) of thiazolidine-2,4-dione derivatives by best 2D-QSAR models**

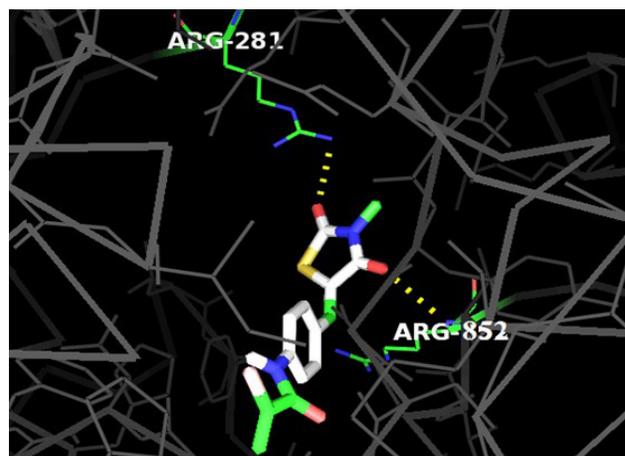
Compounds	Exp. pIC <sub>50</sub> (M)	2D-QSAR (PLSR)		2D-QSAR (MLR)	
		SWFB	GA	SWFB	GA
1_opt.mol2	6.44	6.32	6.32	6.28	6.09
2_opt.mol2	6.06	6.31	6.15	6.31	6.15
3_opt.mol2	5.64	5.75	5.67	5.74	5.65
4_opt.mol2	5.38	5.35	5.50	5.40	5.53
5_opt.mol2	5.62	5.61	5.63	5.48	5.52
6_opt.mol2	5.49	5.61	5.63	5.50	5.52
7_opt.mol2	5.55	5.61	5.63	5.58	5.65
8_opt.mol2	5.43	5.61	5.63	5.45	5.51
9_opt.mol2	5.10	5.45	5.48	5.13	5.54
10_opt.mol2	5.43	5.73	5.63	5.46	5.66
11_opt.mol2	6.08	5.65	5.48	5.95	5.80
12_opt.mol2	6.29	6.03	6.08	6.15	6.26
13_opt.mol2	6.03	5.68	5.91	5.51	6.29
14_opt.mol2	6.17	5.98	6.04	6.19	6.01
15_opt.mol2	6.16	5.98	6.04	6.07	5.97
16_opt.mol2	5.85	5.98	6.04	6.03	6.01
17_opt.mol2	5.85	5.97	6.04	5.99	5.97
18_opt.mol2	6.04	5.98	5.89	6.00	5.36
19_opt.mol2	5.92	5.96	6.04	5.99	5.98
20_opt.mol2	5.52	5.92	5.88	5.49	5.39
21_opt.mol2	4.22	4.24	4.22	4.22	4.18
22_opt.mol2	4.28	4.77	5.04	3.82	4.94
23_opt.mol2	4.24	4.14	4.22	3.84	4.19
24_opt.mol2	4.22	4.20	4.22	4.24	4.20

M = Molar concentration

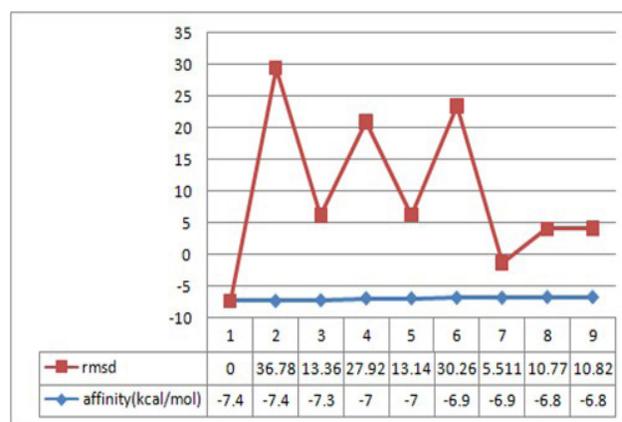
has been provided in Figure 15 and 16. In all the above points the steric descriptor with negative coefficients shows bulky groups are unfavorable at this place and the presence of bulky groups would reduce the anti-cancer activity of these compounds. The above model is validated by predicting the biological activities of the test molecules, as indicated in Table 5.

### Binding mode analysis by molecular docking

The intermolecular interaction between ligand and targeted receptor PI3K $\alpha$  was confirmed by docking study



**Figure 17: The docking study of most active compound 1 with protein binding site.**



**Figure 18: The conformations ranked based on the RMSD values.**

through Auto Dock Vina software. Here, the active compound 1 binds with different sites on PI3K $\alpha$  which provided information for further structural optimization. The binding mode analysis revealed that compound 16 interacts closely with the receptor site and bind with different amino acid residues like ARG-281 and ARG-852 of PI3K $\alpha$  enzyme through hydrogen bonds (Figure 17). The conformations were graded depending on the RMSD values, are described in Figure 18 for most active compound. Lowest RMSD value has 5.511 Å with -6.9

**Table 8. Comparative observed and predicted activities (LOO) of thiazolidine-2,4-dione derivatives by best G-QSAR and 3D models**

Compounds	Exp. pIC <sub>50</sub> (M)	G-QSAR		G-QSAR		3D-QSAR	
		(PLSR)		(MLR)		(kNN)	
		SWFB	GA	SWFB	GA	SWFB	GA
1_opt.mol2	6.44	6.16	6.39	6.19	6.35	6.07	5.91
2_opt.mol2	6.06	5.89	6.18	5.90	5.88	6.26	5.84
3_opt.mol2	5.64	5.72	5.82	5.65	5.90	5.40	5.71
4_opt.mol2	5.38	5.51	5.82	5.42	5.87	5.54	5.83
5_opt.mol2	5.62	5.33	5.62	5.32	5.42	5.30	5.71
6_opt.mol2	5.49	5.50	5.62	5.50	5.40	5.36	5.74
7_opt.mol2	5.55	5.38	5.69	5.31	5.87	5.76	5.65
8_opt.mol2	5.43	5.51	5.69	5.50	5.35	5.30	5.62
9_opt.mol2	5.10	5.21	5.29	5.16	5.36	5.56	5.55
10_opt.mol2	5.43	5.93	5.82	5.95	5.92	5.51	5.82
11_opt.mol2	6.08	5.97	5.82	5.97	5.78	5.94	5.69
12_opt.mol2	6.29	6.08	5.82	6.10	5.91	5.89	5.78
13_opt.mol2	6.03	6.08	5.82	6.10	5.85	6.25	6.11
14_opt.mol2	6.17	6.06	5.62	6.09	5.91	5.95	6.00
15_opt.mol2	6.16	6.07	5.62	6.09	5.91	5.89	6.01
16_opt.mol2	5.85	5.97	5.69	6.00	5.86	5.89	6.07
17_opt.mol2	5.85	5.96	5.69	5.97	5.85	5.98	5.80
18_opt.mol2	6.04	6.06	5.29	6.08	5.87	5.89	5.80
19_opt.mol2	5.92	6.04	5.82	6.06	5.92	5.95	6.05
20_opt.mol2	5.52	6.09	5.82	6.11	5.76	5.89	5.90
21_opt.mol2	4.22	4.04	4.32	4.07	4.25	4.28	4.27
22_opt.mol2	4.28	3.86	4.19	3.83	4.18	4.22	4.25
23_opt.mol2	4.24	4.17	4.19	4.18	4.18	4.25	4.65
24_opt.mol2	4.22	4.44	4.12	4.48	4.23	4.25	4.27

kcal/mol binding affinity proved good inhibitory activity against PI3K $\alpha$ .

## CONCLUSION

The present research represents an effort to recognize the necessary structural requirements of TZD derivatives to be potential anticancer agents. From QSAR analysis, ten best models were generated among which any model may be employed for the creation of new molecules as anticancer agents. Descriptors generated in 2D-QSAR equation describe the imperative role of TZD scaffold for anticancer activity of compounds. The 2D-QSAR model signifies that the descriptors possess high statistically significance and agreeable with a high correlation coefficient and reliable certainty. These computational models have generated a variety of descriptors like MomInertiaZ, kappa2, XKMos-

tHydrophilic, Quadropole2, XAMostHydrophilic, chi3Cluster, T\_2\_O\_7, T\_2\_Br\_7, T\_2\_O\_3, T\_T\_Br\_7, T\_N\_N\_7 contribute to biological activity. The negative coefficient value of descriptors count on the biological activity indicated that a lower value leads to better anticancer effect. The descriptors having a positive coefficient value implies that a higher value leads to better anticancer activity. The outcome of G-QSAR generated diverse descriptors like R<sub>1</sub>-XAMostHydrophobic, R<sub>2</sub>-MomInertiaZ, R<sub>1</sub>-+vePotentialSurfaceArea, R<sub>2</sub>-Epsilon3, R<sub>1</sub>-polarizabilityAHC, R<sub>1</sub>-T\_C\_O\_6, R1-SKMostHydrophilic and R<sub>2</sub>-XAAverage, contribute to the biological activity. G-QSAR studies revealed that presence of hydrophobic group at R<sub>1</sub> position and a more electropositive atom(s) at R<sub>2</sub> position increase the anticancer activity. The results obtained from 3D-QSAR studies were used to optimize the steric, electrostatic, and hydrophobic requirements in the region

of the TZD scaffold for enhancing the anticancer activity. 3D-QSAR studies suggested that substitution of less bulky and electronegative groups around R<sub>1</sub> and R<sub>2</sub> increase anticancer activity. The results of present work may be useful for (medicinal) chemists in understanding the relationship of biological activity with that of physicochemical parameters associated with the structure which will be helpful to select the suitable substituent for the development of more potent, effective and selective anticancer agents.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge India Council of Medical Research (ICMR), New Delhi, for providing Senior Research Fellowship (SRF) to Mr. Vivek Asati (SRF No. 45/28/2013-PHA/BMS). The authors also acknowledge Scripps Research Institute for Auto Dock Vina and V-life Sciences Technologies Pvt. Ltd. for providing the software.

## CONFLICT OF INTEREST

Authors have no conflict of interest with the content of this article.

## ABBREVIATIONS USED

**TZD:** Thiazolidinedione; **SWFB:** Step-wise Forward Backward; **GA:** Genetic Algorithm; **G-QSAR:** Group-QSAR; **MFA:** Molecular field analysis; **SBDD:** Structure-based Drug Design.

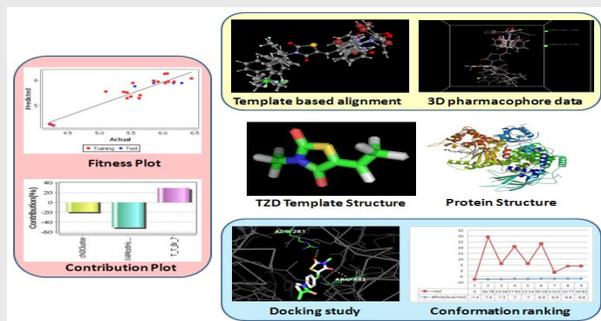
## REFERENCES

- Asati V, Mahapatra DK, Bharti SK. Thiazolidine-2, 4-diones as multi-targeted scaffold in medicinal chemistry: potential anticancer agents. *Eur J Med Chem.* 2014;87:814-33. <https://doi.org/10.1016/j.ejmech.2014.10.025> ; PMID:25440883.
- Jain AK, Vaidya A, Ravichandran V, Kashaw SK, Agrawal RK. Recent developments and biological activities of thiazolidinone derivatives: a review. *Bioorg Med Chem.* 2012;20(11):3378-95. <https://doi.org/10.1016/j.bmc.2012.03.069> ; PMID:22546204.
- Knight SD, Adams ND, Burgess JL, Chaudhari AM, Darcy MG, Donatelli CA, Luengo JI, Newlander KA, Parrish CA, Ridgers LH, Sarpong MA. Discovery of GSK2126458, a highly potent inhibitor of PI3K and the mammalian target of rapamycin. *ACS Med Chem Lett.* 2010;1(1):39-43. <https://doi.org/10.1021/ml900028r> ; PMID:24900173 PMID:PMC4007793.
- Azizmohammadi M, Khoobi M, Ramazani A, Emami S, Zarrin A, Firuzi O, Miri R, Shafiee A. 2H-chromene derivatives bearing thiazolidine-2, 4-dione, rhodanine or hydantoin moieties as potential anticancer agents. *Eur J Med Chem.* 2013;59:15-22. <https://doi.org/10.1016/j.ejmech.2012.10.044> ; PMID:23202485.
- Asati V, Bharti SK. QSAR studies for some thiazolidine-2, 4-dione derivatives as PIM-2 kinase inhibitors. *Med Chem Res* 2016;25(7):1329-39. <https://doi.org/10.1007/s00044-016-1577-z>.
- Patil V, Tilekar K, Mehendale-Munj S, Mohan R, Ramaa CS. Synthesis and primary cytotoxicity evaluation of new 5-benzylidene-2, 4-thiazolidinedione

- derivatives. *Eur J Med Chem.* 2010;45(10):4539-44. <https://doi.org/10.1016/j.ejmech.2010.07.014> ; PMID:20667627.
- Penthala NR, Yerramreddy TR, Crooks PA. Synthesis and *in vitro* screening of novel N-benzyl aplysinopsin analogs as potential anticancer agents. *Bioorg Med Chem Lett.* 2011;21(5):1411-3. <https://doi.org/10.1016/j.bmcl.2011.01.020> ; PMID:21295476 PMID:PMC4400819.
- Shiau CW, Yang CC, Kulp SK, Chen KF, Chen CS, Huang JW, Chen CS. Thiazolidenediones mediate apoptosis in prostate cancer cells in part through inhibition of Bcl-xL/Bcl-2 functions independently of PPAR $\gamma$ . *Cancer Res.* 2005;65(4):1561-9. <https://doi.org/10.1158/0008-5472.CAN-04-1677> ; PMID:15735046.
- Xia Z, Knaak C, Ma J, Beharry ZM, McInnes C, Wang W, Kraft AS, Smith CD. Synthesis and evaluation of novel inhibitors of Pim-1 and Pim-2 protein kinases. *J Med Chem.* 2008;52(1):74-86. <https://doi.org/10.1021/jm800937p> ; PMID:19072652.
- Barros FW, Silva TG, da Rocha Pitta MG, Bezerra DP, Costa-Lotufo LV, de Moraes MO, Pessoa C, de Moura MA, de Abreu FC, de Lima MD, Galdino SL. Synthesis and cytotoxic activity of new acridine-thiazolidine derivatives. *Bioorg Med Chem.* 2012;20(11):3533-9. <https://doi.org/10.1016/j.bmc.2012.04.007> ; PMID:22546208.
- Romagnoli R, Baraldi PG, Salvador MK, Camacho ME, Balzarini J, Bermejo J, Estévez F. Anticancer activity of novel hybrid molecules containing 5-benzylidene thiazolidine-2, 4-dione. *Eur J Med Chem.* 2013;63:544-57. <https://doi.org/10.1016/j.ejmech.2013.02.030> ; PMID:23537942.
- V Life Sciences Technology Pvt. Ltd. Pune-411045.
- Halgren TA. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J Comp Chem.* 1996;17(5-6):553-86. [https://doi.org/10.1002/\(SICI\)1096-987X\(199604\)17:6<553::AID-JCC3>3.3.CO;2-H](https://doi.org/10.1002/(SICI)1096-987X(199604)17:6<553::AID-JCC3>3.3.CO;2-H) ; [https://doi.org/10.1002/\(SICI\)1096-987X\(199604\)17:5/6<553::AID-JCC3>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<553::AID-JCC3>3.0.CO;2-T).
- Ghosh P, Bagchi MC. QSAR modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. *Curr Med Chem.* 2009;16(30):4032-48. <https://doi.org/10.2174/092986709789352303> ; PMID:19747124.
- Sahu NNK, Shahi S, Sharma MC, Kohli DV. QSAR studies on imidazopyridazine derivatives as PfPK7 inhibitors. *Mol Simul.* 2011;37(9):752-65. <https://doi.org/10.1080/08927022.2010.547050>.
- Baumann K. An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. *J Chem Inf Comp Sci.* 2002;42(1):26-35. <https://doi.org/10.1021/ci990070t>.
- Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers.* 2000;5(4):231-43. <https://doi.org/10.1023/A:1021372108686>.
- Ajmani S, Jadhav K, Kulkarni SA. Group-Based QSAR (G-QSAR): Mitigating Interpretation Challenges in QSAR. *QSAR Comb Sci.* 2009;28(1):36-51. <https://doi.org/10.1002/qsar.200810063>.
- Ajmani S, Agrawal A, Kulkarni SA. A comprehensive structure-activity analysis of protein kinase B-alpha (Akt1) inhibitors. *J Mol Graph Model.* 2010;28(7):683-94. <https://doi.org/10.1016/j.jmgs.2010.01.007> ; PMID:20153226.
- Ajmani S, Jadhav K, Kulkarni SA. Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J Chem Inf Model.* 2006;46(1):24-31. <https://doi.org/10.1021/ci0501286> ; PMID:16426036.
- Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J Med Chem.* 2003;46(14):3013-20. <https://doi.org/10.1021/jm020491t> ; PMID:12825940.
- Hasegawa K, Kimura T, Funatsu K. GA strategy for variable selection in QSAR studies: enhancement of comparative molecular binding energy analysis by GA-based PLS method. *QSAR.* 1999;18(3):262-72. [https://doi.org/10.1002/\(sici\)1521-3838\(199907\)18:3<262::aid-qsar262>3.0.co;2-s](https://doi.org/10.1002/(sici)1521-3838(199907)18:3<262::aid-qsar262>3.0.co;2-s).
- Sharaf MA, Illman DL, Kowalski BR. *Chemometrics.* John Wiley & Sons; 1986. PMID:3726980.
- Holland JH. Genetic algorithms. *Sci Am.* 1992;267(1):66-72. <https://doi.org/10.1038/scientificamerican0792-66>.
- Wold S. PLS for multivariate linear modelling. In: van de Waterbeemd H, ed. *QSAR: chemometric methods in molecular design.* Weinheim: Wiley-VCH 1995;195-218. <http://vina.scripps.edu/> (Accessed on Oct. 2, 2016)..

26. Trott O, Olson AJ. Auto Dock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comp Chem*. 2010;31(2):455-61. PMID:19499576 PMCID:PMC3041641.
27. Seeliger D, de Groot BL. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J Comput Aided Mol Des*. 2010;24(5):417-22. <https://doi.org/10.1007/s10822-010-9352-6>; PMID:20401516 PMCID:PMC2881210.
28. Chang MW, Ayeni C, Breuer S, Torbett BE. Virtual screening for HIV protease inhibitors: a comparison of AutoDock 4 and Vina. *PLoS One*. 2010;5(8):e11955. <https://doi.org/10.1371/journal.pone.0011955>; PMID:20694138 PMCID:PMC2915912.
29. Österberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins Struct Funct Bioinform*. 2002;46(1):34-40. <https://doi.org/10.1002/prot.10028> ; PMID:11746701.
30. Stigliani JL, Bernardes-Génisson V, Bernadou J, Pratiel G. Cross-docking study on InhA inhibitors: a combination of Autodock Vina and PM6-DH2 simulations to retrieve bio-active conformations. *Org Biomol Chem*. 2012;10(31):6341-9. <https://doi.org/10.1039/c2ob25602a> ; PMID:22751934

## PICTORIAL ABSTRACT



## SUMMARY

- SWFB and GA strategies have been performed for various QSAR models.
- The 3D-QSAR studies were performed by k-nearest neighbor molecular field analysis.
- A docking study revealed the binding orientations of the inhibitors.
- The study revealed the potential of thiazolidine-2,4-dione derivatives as promising antitumor candidates.

## About Authors



**Vivek Asati:** He is currently working as an ICMR-SRF under Dr. Sanjay Kumar Bharti since 2013 in Guru Ghasidas Vishwavidyalaya Bilaspur, Chhattisgarh. He had completed B. Pharm. and M. Pharm. from Department of Pharmaceutical Sciences, Dr. Hari Singh Gour Sagar University, Sagar, Madhya Pradesh. He also has a teaching experience of 3 years in nationally reputed colleges. His current research includes QSAR studies and synthesis of heterocyclic anti-cancer compounds. He has published research papers, review articles, and book chapters in internationally reputed journals. He has presented research outcomes at several national and international platforms.



**Sanjay Kumar Bharti:** He had completed B. Pharm. from IT-BHU, Varanasi in 2003, M. Pharm. from RGPV, Bhopal in 2004 and Ph.D. from IIT-BHU, Varanasi in 2011. He has working experience in reputed organizations like Win-Medicare Pvt. Ltd., Meerut as Chemist and NIPER, Hajipur as Lecturer. He has published research papers and review articles in various reputed journals, along with several book chapters for internationally reputed Presses. He has also authored and edited books at both national and international levels. Presently, he is working as Assistant Professor in Guru Ghasidas Vishwavidyalaya (A Central University), Bilaspur, Chhattisgarh, India.



**Debarshi Kar Mahapatra:** He is currently working as Assistant Professor in DBCOP, Nagpur where he teaches medicinal and computational chemistry at both undergraduate and post-graduate levels. He has mentored few students in their dissertation and carried out various national research projects. His current area of interest involves computer assisted rational designing and synthesis of low molecular weight ligands. He has authored a book entitled Drug Design and edited a book entitled Handbook of Research in Medicinal Chemistry. He has presented research papers at several international platforms for which he received several awards. He is a member of a number of professional and scientific societies as well as serving the position of reviewer and editorial board member of various international journals.

**Cite this article:** Asati V, Bharti SK, Rathore A, Mahapatra DK. SWFB and GA strategies for variable selection in QSAR studies for the validation of thiazolidine-2,4-dione derivatives as promising antitumor candidates. *Indian J of Pharmaceutical Education and Research*. 2017;51(3):436-51.