

QSAR, molecular docking and protein ligand interaction fingerprint studies of N-phenyl dichloroacetamide derivatives as anticancer agents

Masood Fereidoonzhad¹, Zeinab Faghih², Elham Jokar¹, Ayyub Mojaddami², Zahra Rezaei², Mehdi Khoshneviszadeh^{2,3,*}

¹Department of Medicinal Chemistry, School of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

²Department of Medicinal Chemistry, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.

³Medicinal and Natural Products Chemistry Research Center, Shiraz University of Medical Sciences, Shiraz, Iran.

Abstract

Dichloroacetate (DCA) is a pyruvate mimetic compound that stimulates the activity of the enzyme pyruvate dehydrogenase (PDH) through inhibition of the enzyme pyruvate dehydrogenase kinases (PDK1-4). DCA works by turning on the apoptosis which is suppressed in tumor cells, hence letting them die on their own. Here, in this paper a series of DCA analogues were applied to 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), simple Free-Wilson analysis (FWA) and partial least squared combined with genetic algorithm for variable selection (GA-PLS), were conducted to make relations between structural features and cytotoxic activities of a variety of DCA derivatives. The best multiple linear regression equation was obtained from genetic algorithms partial least squares which predicted 91% of variances. On the basis of the produced model, an *in silico*-screening study was also employed and new potent lead compounds based on new structural patterns were suggested. Docking studies of these compounds were also investigated and promising results were obtained. The docking results were also conducted to protein ligand interaction fingerprints (PLIF) studies, using self-organizing map (SOM) in order to evaluate the predictive ability in suggesting new potent compounds and some compounds were introduced as a good candidate for synthesis.

Keywords: *In silico* screening, Molecular docking, N-phenyl dichloroacetamide, Protein ligand interaction fingerprints, QSAR.

1. Introduction

Recently the tumor metabolism and the Warburg effect have attracted scientists' interest in the fields of mitochondrial function and oncogenic regulation of metabolism (1). Some metabolic pathways such as programmed cell death, that play a great role in tumor growth are being introduced as novel targets for anticancer drug development (2, 3). To the best of our knowledge apoptosis

and the mechanisms evolved by tumor cells to refrain from engagement in cell death, are complicated processes. Pyruvate dehydrogenase complex (PDC), is one of the major regulators of mitochondrial function. The activity of PDC is regulated by reversible phosphorylation of three serine residues on the E1 α subunit. PDH kinases (PDK) phosphorylate these sites. There are four known isoforms of PDKs (PDK1-4) that are distributed in a different manner in tissues. PDKs are novel therapeutic targets in the treatment of cancer (2, 4).

Dichloroacetate (DCA) is a lactate-lowering drug which has been in use for many years to

Corresponding Author: Mehdi Khoshneviszadeh, Department of Medicinal Chemistry, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran.
Email: khoshnevim@sums.ac.ir

treat various diseases such as lactic acidosis and inherited defects in mitochondrial metabolism (5). In 2007 it was discovered that DCA induces the death of human lung, breast and brain cancer cells that were embedded into rats, while being non-toxic to healthy cells (6). Current studies show that sodium dichloroacetate (DCA) can selectively promote mitochondria-regulated apoptosis, depolarizing the hyperpolarized inner mitochondrial membrane potential to normal levels, inhibit tumor growth and reduce proliferation by shifting the glucose metabolism in cancer cells from anaerobic to aerobic glycolysis (2, 4).

Expressing biological activity quantitatively is of great importance in the field of medicinal chemistry. Moreover, having expressed structures or physicochemical properties by numbers, a mathematical relationship can be found between the two. The mathematical expression, if carefully validated can then be used to predict the modeled response of other chemical structures. In the field of drug design and medicinal chemistry the QSAR information have great importance (7). There are different variable selection methods such as multiple linear regression (MLR), principal component or factor analysis (PCA/FA), genetic algorithm, and so on available for QSAR studies (8).

Here, QSAR studies of a series of N-phenyl dichloroacetamide derivatives with great cytotoxic activity against different cancer cell lines, which have been recently designed and synthesized by Y. Yang *et al.* (9) have been explored. Recently the cytotoxic activity of some of these compounds against different cell lines such as human lung (NCI-H460), colon (HCA-7) and endometrial (MCF-7) cancer cell lines was also evaluated (10). Among different QSAR models, the best multiple linear regression equation was obtained from GA-PLS models which was a linear seven-parameter model. Thereafter, a virtual screening study was employed to determine novel biologically active patterns by insertion, deletion and substitution of different substitutes of the primary molecules. The results of this study led to the identification of novel structures, which are potent anticancer agents according to the QSAR model. It should also be mentioned that docking and PLIF studies of these compound were also carried out and promising re-

sults were obtained. There was a good correlation between the results of docking and QSAR studies.

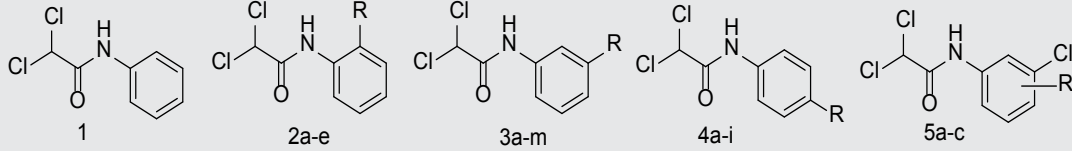
2. Materials and Methods

2.1. Data set

The biological data used in this paper are the cytotoxic activity of a series of N-phenyl dichloroacetamide derivatives on a human nonsmall cell lung cancer cell line (A549), which were designed, synthesized and evaluated for their ability to induce apoptosis by Yang *et al.* (9). The structural features and biological activities of these compounds are listed in Table 1. The biological data were converted to logarithmic scale (pIC_{50}) and then used for subsequent QSAR analysis as dependent variables.

2.2. Molecular descriptors

The two dimensional structures of the ligands were drawn using ACD chemsketch software. Then the ligands were subjected to minimization procedures by means of an in house TCL script using Hyperchem (Version 8, Hypercube Inc., Gainesville, FL, USA). Each ligand was optimized with different minimization methods such as commonly used molecular mechanics method (MM+) and then quantum based semiempirical method (AM1) by using Hyperchem package. The Z-matrices of the structures were constructed by the software and then transferred to the Gaussian 98 program (11). HyperChem, Gaussian 98 and Dragon softwares (12) were used for calculation of molecular descriptors. Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies and molecular dipole moment were calculated by Gaussian 98. Quantum chemical indices of hardness ($\eta=0.5$ (HOMO+LUMO)); softness ($S=1/\eta$); electronegativity ($\chi=-0.5$ (HOMO-LUMO)); and electrophilicity ($\omega=\chi^2/2\eta$) were calculated according to the equations proposed by Thanikaivelan *et al.* (13). Some chemical parameters including molar volume (V), molecular surface area (SA), hydrophobicity (logP), hydration energy (HE) and molecular polarizability were calculated using Hyperchem software. Dragon calculated different topological, geometrical, charge, empirical and constitutional descriptors for each molecule. 2D autocorrelations,

Table 1. Chemical structure of the N-phenyl-2,2-dichloroacetamide analogues used in this study and their experimental and cross-validated predicted activity (by GA-PLS) for cytotoxic activity as well as their docking binding energies.


Name	R	Exp.pIC ₅₀	Pred. pIC ₅₀ ^a	Binding Energy (kcal/mol)
1	-	3.88605	4.08685	-4.75
2a	F	4.05784	4.03370	-5.25
2b	Cl	4.29973	4.14783	-5.42
2c	Br	4.43191	4.22958	-5.56
2d	NO ₂	3.7043	3.90310	-4.51
2e	NHCOCHCl ₂	3.88292	4.21780	-4.73
3a	CH ₃	4.38668	4.30985	-5.54
3b	Cl	4.83032	4.85377	-5.78
3c	Br	5.1079	5.11930	-5.89
3d	I	5.32239	5.31882	-6.12
3e	CN	4.17698	4.91989	-5.31
3f	C ≡ CH	4.95	4.93897	-5.81
3g	NO ₂	4.74376	4.65637	-5.66
3h	OCH ₃	4.58871	4.61296	-5.51
3i	CF ₃	4.91435	4.62387	-5.82
3j	OCF ₃	4.82246	4.67821	-5.73
3k	SCF ₃	4.8517	4.98226	-5.77
3l	SO ₂ CF ₃	5.18508	5.16325	-5.87
3m	SO ₂ Ph(m-NHCOCHCl ₂)	5.22988	4.57566	-6.03
4a	CH ₃	4.31051	4.28693	-5.57
4b	F	4.16488	4.27381	-5.3
4c	Cl	4.691	4.60795	-5.59
4d	Br	4.86201	4.79344	-5.69
4e	I	4.9017	4.94087	-5.67
4f	NO ₂	4.46319	4.74464	-5.37
4g	OCH ₃	4.09216	4.15534	-5.27
4h	SCF ₃	4.976336	4.76654	-5.91
4i	SO ₂ CF ₃	5.070581	5.06627	-5.72
5a	4Cl	5.29328	5.19766	-5.66
5b	5Cl	5.35654	5.40682	-6.05
5c	6Cl	4.90728	4.85008	-5.54

^aCross-validated prediction by GAPLS.

aromaticity indices, atom-centered fragments and functional groups were also calculated by dragon.

In the case of docking procedure, each ligand was optimized with different minimization

MM+ then AM1 using HyperChem 8. The output structures were thereafter converted to PDBQT using MGLtools 1.5.6 (14). The three dimensional crystal structure of pyruvate dehydrogenase kinase 2 (PDB ID:2BU8) 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 modeller 9.12 (16). An in house application (MODEL-FACE) was used for generation of python script and running modeller software. Subsequently, the enzymes were converted to PDBQT and gasteiger partial charges were added using MGLTOOLS 1.5.6.

2.3. Model development

Four different regression methods were conducted for constructing QSAR equations: (1) simple multiple linear regression with stepwise variable selection (MLR) (2) factor analysis as the data preprocessing step for variable selection (FA-MLR), (3) principal component regression analysis (PCRA), and (4) genetic algorithm–partial least squares (GA-PLS). Simple Free-Wilson analysis (FWA) was also carried out. These methods are well substantiated in the QSAR studies, and therefore, these methods are described briefly.

Stepwise regression is a semi-automated process of building a model by successively adding or removing variables, based solely on the t-statistics of their estimated coefficients. In stepwise regression (17), a multiple-term linear equation was constructed step by step. The basic procedures include (i) recognizing a primary model, (ii) iteratively ‘stepping’, that is, repetitively changing the model at the prior step by adding or removing a predictor variable in accordance with the ‘stepping criteria’ (in our case, probability of $F=0.05$ for inclusion; probability of $F=0.1$ for leaving out for the forward selection method), and (iii) terminating the search when stepping is no longer possible given the stepping criteria, or when a known maximum number of steps have been obtained. Particularly, at each step, for determining which one will contribute most to the equation, all variables are reviewed for evaluation (17). The variable will then be applied in the model, and the process

starts again. A limitation of the stepwise regression search approach is that it assumes there is a single ‘best’ subset of X variables and searches to identify it. There is often no unique ‘best’ subset, and whole possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to explore whether some other subsets of X variables might be better (18). Here in this study, MLR with stepwise selection and elimination of variables was applied for developing QSAR models using SPSS software (version 21; SPSS Inc., IBM, Chicago, IL, USA). Using MATLAB 2015 software (version 8.5; Math work Inc., Natick, MA, USA), the resulted models were validated by leave-one-out cross-validation procedure to check their predictivity and robustness.

In the FA-MLR method, although classical approach of multiple regression technique was applied as the final statistical tool for developing QSAR relation, factor analysis (FA) (8, 17) was used as the data-preprocessing step to identify the important predictor variables contributing to the response variable and to avoid collinearities among them. In a typical factor analysis procedure, standardizing the data matrix then correlation matrix is constructed. An eigenvalue problem is then solved and the factor pattern can be acquired from the corresponding eigenvectors (characteristic vector). The principal objectives of factor analysis (FA) are to display multidimensional data in a space of lower dimensionality with minimum loss of information (explaining >95% of the variance of the data matrix) and to extract the basic features behind the data with ultimate goal of interpretation or prediction. Factor analysis was done on the data set, containing biological activity and all descriptor variables, which were to be considered. The factors were extracted by principal component method and then rotated by VARIMAX rotation (19).

Along with FA-MLR, PCRA was also tried for the data set. In this method (8, 17), factor scores obtained from FA are used as the predictor variables. PCRA has a benefit that collinearities among X variables are not a disturbing factor and that the number of variables included in the analysis may exceed the number of observations (20).

While the main purpose of FA-MLR is to identify relevant descriptors, in PCRA all descriptors are supposed to be important.

Genetic algorithms (GA) generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Partial least square (PLS) is a generalization of regression, that can handle data with forcefully correlated and or numerous X variables (21). It gives reduced solution, which is statistically more robust and reliable than MLR. The linear PLS model finds 'new variables' (latent variables or X scores) that are linear combination of the original variables. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross-validation is a practical and credible method of testing this significance (22). Application of PLS thus allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables. Usually, PLS is applied in combination with cross-validation to obtain the optimum number of components (17, 23, 24). In the GA-PLS procedure, in addition to the best set of descriptors, the optimum number of concealed variables must be determined. Here, for each subset of descriptors (i.e., for each chromosome of the GA), a PLS model was developed separately and therefore the number of latent variables was optimized. The PLS regression method was applied based on the NIPALS-based algorithm existing in the chemometrics toolbox of MATLAB software. Leave-one-out cross-validation procedure was used to obtain the optimum number of factors based on the Haaland and Thomas F -ration criterion (17, 25). The MATLAB PLS toolbox developed by eigenvector company was used for PLS and GA modeling. All calculations were run on a core i7 personal computer (CPU at 6 MB) with Windows 7 as the operating system.

2.4. Model validation

Statistical parameters including correlation coefficient (R^2), standard error of regression (SE), and variance ratio (F) at specified degrees of freedom were used for validating the goodness-of-

fit of the resulted QSAR models. The generated QSAR equations were also validated by leave-one-out cross-validation correlation coefficient (Q^2), root mean square error of cross-validation (RMSE_{cv}) and cross validation cross validation (C_{vcv}). According to Tropsha *et al.* (26) the predictive ability of a QSAR model should be tested on an external set of data that has not been taken into account during the process of developing the model. Therefore, as it was shown in Table 1, an external test set, composed of 6 randomly selected molecules (for example 3a, 3d, 3g, 3m, 4f and 5a) was applied to determine the overall prediction ability of the resulted models. It should be emphasized that we carried out each QSAR model with more than 3 test sets and the best equation was considered as the best model.

2.5. Applicability domain

One of the great uses of a QSAR model is based on its precise prediction ability for new compounds. A model validation is just within its training domain, and new compounds must be appraised as belonging to the domain before the model is applied. The applicability domain is appraised by the leverage values for each compound. A Williams plot (the plot of standardized residuals versus leverage values (h)) can then be used for an immediate and simple graphical detection of both response outliers (Y outliers) and structurally influential chemicals (X outliers) in our model. In this graph, the applicability domain is established inside a squared area within $\pm x$ (standard deviations) and a leverage threshold h^* . The threshold h^* is generally fixed at $3(k+1)/n$ (k is the number of model parameters and n is the number of training set compounds), whereas x is normally equal to 2 or 3. Prediction must be considered unreliable for compounds with a high leverage value ($h > h^*$). From the other point of view, when the leverage value of a compound is lower than the threshold value, the probability of agreement between observed and predicted values is as high as that for the training set compounds (7, 27).

2.6. Docking procedure

The docking simulations were carried out by means of an in house batch script (DOCK-

FACE) for automatic running of AutoDock 4.2 (28) in a parallel mode, using all system resources. In all experiments Genetic algorithm search method was used to find the best pose of each ligand in the active site of the target enzyme. Random orientations of the conformations were generated after translating the center of the ligand to a specified position within the receptor active site, and making a series of rotamers. This process was recursively repeated until the desired number of low-energy orientations was obtained. No attempt was made to minimize the ligand-receptor complex (rigid docking). Ligands were submitted to 100 independent genetic algorithm (GA) runs for search. For Lamarckian GA method, 150 population size, a maximum number of 2,500,000 energy evaluations and 27,000 maximum generations were used. A grid of 50, 50, and 50 points in x-, y-, and z-direction, respectively, for PDK2 receptor (2BU8) with grid spacing of 0.375 Å was built, centered on the catalytic site of the receptors. No. of points

in x, y and z was 51, 44 and 82 respectively.

2.7. Protein ligand interaction fingerprint (PLIF)

In order to perform PLIF studies on docking results, the poses of docking were extracted from dlG files using an in house vb.net application (preAuposSOM) (29). The resulted pdbqts and the receptor were converted to mol2 using OpenBabel 2.3.1. The resulted mol2 files were submitted to AuposSOM 2.1 web server (30-32). Two training phases with 1000 iterations were set in the self-organizing map settings of AuposSOM conf files. Other parameters of the software remained as default. The output files were subjected to Dendroscope 3.2.10 for visualization of the results (33, 34).

3. Results and Discussion

In this paper, we executed a detailed QSAR study using a combination of chemical, electronic, substituent constant, and Free-Wilson analysis to explore structural parameters affecting cyto-

Table 2. The results of different QSAR model analysis with different type of dependent variables.

Eq.no.	Equation	n ^a	R ² c	Q ²	Rmscv	Cvcv	F	SE	R ² p
1) MLR	pIC ₅₀ =-2.019 MATS5p(±0.237)-0.043 DipX(±0.039)+0.261 nCaR(±0.044)+1.238 ASP(±0.20)+0.688 IC2(±0.188)+1.196 ATS7e(±0.366)-0.166(±0.930)	24	0.94	0.91	0.128	2.68	60.4	0.11	0.672
2) FA-MLR	pIC ₅₀ =-0.742 MATS5p(±0.371)-0.237 DipX(±0.039)+0.206 logP (±0.043)+0.407 ATS7v (±0.060)+0.700 MATS7p (±0.296)+0.65 X4Av(±0.104)+3.305(±0.206)	24	0.91	0.87	0.168	3.79	24.9	0.29	0.697
3) PCR	pIC ₅₀ =0.270 FAC3(±0.038)+0.170 FAC2(±0.038)+0.139 FAC11(±0.038)+0.132 FAC12(±0.038)-0.111 FAC14(±0.038)-0.105 FAC9 (±0.038)+0.098 FAC1(±0.038)-0.088 FAC13(±0.038)+4.660(±0.037)	24	0.94	0.90	0.157	3.38	36.5	0.25	0.695
4) GA-PLS	pIC ₅₀ =-0.209 DipX(±0.025)+0.085 HE(±0.085)+1.303 MATS7v(±0.166)-0.361 C-040(±0.066)+2.140 ATS7v(±0.213)+1.137 ATS2e(±0.379)+3.073(±0.465)	24	0.98	0.94	0.106	2.30	117	0.07	0.91

^aNumber of molecules of training set used to derive the QSAR models.

toxic activity of N-phenyl dichloroacetamide derivatives. Among the different chemometric tools available for modeling the relationship between the biological activity and molecular descriptors, four methods (i.e., stepwise MLR, FA-MLR, PCRA, and GA-PLS) were applied and compared here. FWA were also performed. A comparison between stepwise FA-MLR and MLR will indicate which variable selection method (stepwise or FA) is well suited for MLR, whereas a comparison between FA-MLR and PCRA reveals for modeling of the studied biological activities, using original descriptors selected, based on factor loading or using the factor scores calculated based on all calculated descriptors, results in a more suitable model. Eventually, GA-PLS, which is assumed to produce the most useful model, was employed, and its results were compared with the other employed models.

3.1. MLR modeling

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. As there are 31 molecules in the training set and according to the rule of thumb (the ratio of 5:1 for molecule/variable/Toplis ratio), MLR models with maximum number of variables of 6 were selected. Statistical parameters such as correlation coefficient (R^2), correlation coefficient (R^2 test set) of test set, standard error of regression (SE), and variance ratio (F) at specified degrees of freedom, leave-one-out cross-validation correlation coefficient (Q^2), cross validation cross validation (Cvcv) and root mean square error of cross-validation (RMScv) were used for validating the

goodness-of-fit of the resulted MLR equations. As it was shown in Table 2, Equation 1 was selected as the best equation in the MLR model because of its greatest statistical parameters. The selected variables demonstrate that quantum (DipX), topological (IC2), geometrical (ASP), 2D autocorrelations (MATS5p, ATS7e), and functional (nCaR) descriptors affect the cytotoxic activity of the studied compounds.

A small difference between the conventional and cross-validate correlation coefficients of the different MLR equations (Table 3) reveals that none of the models are overfitted, 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. The correlation coefficient (r^2) matrix for the descriptors used in MLR Equation 1 (as the best equation in this series) shows that no significant correlation exists between pairs of descriptors (Tables 3).

3.2. Free-Wilson analysis

The simple Free-Wilson analysis (FWA) (35) was selected in this article to show which substituents on phenyl ring contribute to cytotoxic activity and which ones detract from activity. As it is shown in Table 1, the selected compounds have a phenyl ring with different substituents on different position of the ring. The important substituents such as OMe, F, Br, Cl, NO₂, CH₃, CN, Cl, CF₃, OCF₃, SCF₃, SO₂CF₃ and some other substituents were used in this calculation. Therefore, the descriptors data matrix built for the FWA has 31 rows (i.e., number of molecules) and 29 columns. The elements of the descriptor data matrix are 1 or 0, which indicate the presence or absence of a given substituent on a specified position in a molecule,

Table 3. Correlation coefficient (R^2) matrix for descriptors represented in multiple linear regression Equation 1.

	IC2	MATS5p	nCaR	DipX	ATS7e	ASP
IC2	1	-0.243	0.094	-0.127	0.148	-0.313
MATS5p		1	-0.463	-0.043	0.149	0.318
nCaR			1	0.156	-0.059	-0.094
DipX				1	-0.460	-0.440
ATS7e					1	-0.076
ASP						1

respectively. The following multi-linear equation was found between the activity data (y) and the Free-Wilson type descriptors data matrix:

$$pIC_{50} = -0.925I_{2-NO_2} (\pm 0.422) + 0.467I_{3-Cl} (\pm 0.223) + 4.630 (\pm 0.081) \quad (\text{Eq. 5})$$

N=31, R²=0.832, SE=0.414, F=19.605

These equations indicate that cytotoxic activity of studied compounds are directly affected by the presence of nitro on position 2 and chloro on position 3 of phenyl ring. While 3-chloro substitu-

tion showed positive effects on the activity of the molecules, 2-nitro substitution represented negative effects on the cytotoxic activity. This should be explained that substitution of the 3rd position of the phenyl ring results in higher activity and vice versa.

3.3. FA-MLR and PCRA

It was discovered that seven factors could explain the data matrix to the extent of 95.4%,

Table 4. Factor loadings of some significant descriptors after VARIMAX rotation.

Descriptor	factor1	factor2	factor3	factor9	factor11	factor13	factor14	Communalities	PIC ₅₀
pIC ₅₀	0.21	0.365	0.58	-0.225	0.298	-0.189	-0.06	0.927	
v1	0.93	0.2	0.1	-0.08	0.146	0.041	0.064	0.989	0.48
HE	-0.5	-0.2	0.28	-0.16	-0.097	-0.05	0.023	0.934	0.15
logp	0.07	0.4	0.46	-0.16	0.097	-0.08	-0.001	0.980	0.71
mass	0.8	0.4	0.46	0	-0.063	0.048	-0.001	0.998	0.54
MW	0.8	0.4	0.46	0	-0.063	0.048	-0.018	0.998	0.54
AMW	-0.05	0.2	0.93	0.06	-0.093	0.004	0.015	0.998	0.57
Ss	0.64	0.7	-0.17	0.02	0.044	0.033	0.046	0.998	0.3
nAB	0.73	-0.1	0.03	-0.06	0.042	-0.24	-0.011	0.967	0.36
nN	0.45	-0.1	-0.18	0.04	0.012	0.33	-0.026	0.963	-0.3
X1A	-0.5	-0.8	-0.08	0.18	0.03	0.073	-0.153	0.992	-0.5
X4Av	0.23	0.1	0.73	-0.09	-0.075	-0.19	-0.008	0.974	0.63
IC2	0.18	0.4	-0.05	-0.14	0.708	-0.07	-0.444	0.91	0.47
ATS7v	0.01	-0.1	0.78	-0.18	0.149	0.025	0.127	0.953	0.66
ATS6e	-0.18	0.9	0.07	0.12	-0.104	-0.02	-0.189	0.955	0.18
ATS7e	-0.15	0.8	0.08	-0.13	0.051	0.04	-0.173	0.915	0.37
MATS7v	-0.01	0.3	0.16	0.02	0.597	0.193	-0.659	0.740	0.54
MATS4e	-0.1	-0.1	0.09	-0.28	0.166	0.19	0.052	0.867	0.27
MATS3p	-0.04	0.2	0.3	0.81	-0.039	-0.03	0.311	0.933	0.08
MATS5p	-0.23	-0.1	-0.75	0.08	-0.222	0.023	-0.335	0.933	-0.7
MATS7p	0.13	0.3	0.23	-0.25	0.524	0.067	-0.21	0.912	0.64
GATS3v	-0.04	0	0.16	-0.88	0.068	-0.11	-0.463	0.957	0.34
GATS5v	0.12	0	0.48	-0.01	0.254	-0.01	0.509	0.904	0.55
GATS4e	0.08	0.2	-0.01	0.24	-0.156	-0.33	-0.103	0.892	-0.1
GATS8e	-0.33	-0.2	-0.01	-0.21	0.015	-0.1	-0.022	0.645	-0.1
GATS2p	0.1	0.4	-0.67	0.17	-0.015	0.024	0.556	0.970	-0.3
GATS4p	-0.1	-0	0.5	0.23	0.15	-0.05	0.051	0.932	0.28
HOMT	0.73	-0.1	0.05	-0.08	0.069	-0.25	-0.014	0.970	0.37
J3D	-0.02	0.4	-0.47	-0.01	-0.104	0.629	0.016	0.948	-0.3
MAXDN	0.34	0.9	-0.15	-0.06	0.102	-0.05	-0.043	0.985	0.39
v27	0.69	-0.2	-0.05	0.14	0.127	0.425	0.024	0.834	-0.2
v33	-0.15	-0	-0.23	0.35	-0.19	-0.47	0.177	0.627	-0.3
DipX	-0.02	-0.6	0.01	0.02	-0.065	0.148	-0.06	0.916	-0.5

from the factor analysis on the data matrix consisting of the pIC_{50} and calculated molecular descriptors. Table 4 shows that the biological activity is highly loaded with factors 2 and especially 3. The highest loading values for factor 2 are associated with ATS6e, ATS7e, MAXDN, X1A, Ss and DipX descriptors whereas AMW, X4Av, MATS5p, and GATS2p are the highly loaded descriptors of factor 3. Table 4 revealed that, factors 2 and 3 are moderately loaded with cytotoxic activity. Interestingly, the former possessed the highest loadings from to-

pological (X1A), geometrical (MAXDN), constitutional (Ss), 2D autocorrelations (ATS6e, ATS7e) and quantum (DipX) descriptors, whereas the latter is containing the information from topological (X4Av), constitutional (AMW) and 2D autocorrelation (MATS5p, GATS2p) descriptors. As it was shown in Equation 3, the highly loaded descriptors of factors 1, 2, 3, 9, 11, 13 and 14 (instead of applying the pool of all calculated descriptors), can be considered as the source of molecular descriptors for QSAR model building. So, the probability

Table 5. Definitions of molecular descriptors present in the models.

Descriptor Type	Descriptors	Brief description
2D autocorrelations	ATS7v	Broto-Moreau autocorrelation of a topological structure-lag7/weighted by atomic van der Waals volumes
	MATS7v	Moran autocorrelation - lag7/weighted by atomic van der Waals volumes
	ATS2e	Broto-Moreau autocorrelation of a topological structure-lag2/weighted by atomic Sanderson electronegativities
	ATS7e	Broto-Moreau autocorrelation of a topological structure-lag7/weighted by atomic Sanderson electronegativities
	MATS3p	Moran autocorrelation-lag3/weighted by atomic polarizabilities
	MATS5p	Moran autocorrelation-lag5/weighted by atomic polarizabilities
	MATS7p	Moran autocorrelation-lag7/weighted by atomic polarizabilities
	GATS3v	Geary autocorrelation-lag3/weighted by atomic van der Waals volumes
	GATS6v	Geary autocorrelation-lag6/weighted by atomic van der Waals volumes
	GATS2p	Geary autocorrelation-lag2/weighted by atomic polarizabilities
Chemical descriptors	logP	the logarithm of its partition coefficient between n-octanol and water
	HE	Hydration Energy
Connectivity indices	X1A	average connectivity index chi-1
	X4Av	average valence connectivity index chi-4
Functional group	nCaR	number of substituted aromatic C(sp ²)
Geometrical descriptors	J3D	3D-Balaban index
	ASP	asphericity
Topological descriptors	T(N..F)	sum of topological distances between N..F
	IC2	information content index (neighborhood symmetry of 2-order)
Atom-centred fragments	F-084	F attached to C1(sp ²)
	C-040	R-C(=X)-X/R-C#X/X=C=X
Quantum	(DipX	Molecular dipole moment at X-direction

Table 6. Structural modification of N-phenyl dichloroacetamide derivatives and their predicted activities.

Name	R	pIC ₅₀ pred	leverage	Binding Energy (kcal/mol)
1a	I	5.32163	0.14206	-6.07
1b	C ≡ CH	4.83645	0.03573	-5.75
1c	CF ₃	4.51794	0.00431	-5.62
1d	OCF ₃	4.49525	0.01466	-5.59
1e	SCF ₃	4.73234	0.06899	-5.61
1f	SO ₂ CF ₃	5.12789	0.13082	-5.87
1g	SO ₂ CF ₃	5.15017	0.15233	-5.54
1h	I	4.93311	0.04201	-5.78
1i	4Cl	5.19182	0.09354	-5.89
1j	5Cl	5.33568	0.15471	-6.12
2n	3-pyridine	4.42361	0.01447	-5.41
2o	4-pyridine	4.08172	0.18136	-5.21
2p	4-imidazole	3.11165	0.35164	-5.66
2q	2-thiazole	3.50977	0.20467	-4.89
2r	2-benzothiazole	4.92476	0.10824	-5.82
2s	2-oxazole	3.36374	0.31652	-4.93
2t	2-benzoxazole	5.02848	0.06733	-5.97
2u	2-benzoimidazole	4.83826	0.12422	-5.85
2v	4-(t-butyl)-2-imidazole	4.65586	0.07917	-5.63
2w	3-methyl,3-pyridine	4.63209	0.0005	-5.59
3n	I	5.24612	0.1607	-6.07
3o	C ≡ CH	4.96457	0.19515	-5.89
3p	CF ₃	4.38738	0.03099	-5.49
3q	OCF ₃	4.66007	0.23308	-5.57
3r	SCF ₃	5.07760	0.29115	-5.87
3s	SO ₂ CF ₃	5.09804	0.22186	-5.92
3t	SO ₂ CF ₃	4.53719	0.10288	-5.48
3u	I	4.88779	0.04451	-5.76
3v	4Cl	5.06404	0.10435	-5.92
3w	5Cl	5.35769	0.15984	-6.15

of obtaining chance models is decreased (36).

The subsequent MLR equation using highly loaded descriptors is shown in Equation 2, Table 2.

3.4. PCRA

When factor scores were used as the predictor parameters in a multiple regression equation (instead of their highly loaded descriptors), a predictive QSAR model with factor scores of 1, 2, 3,

9, 11, 13 and 14 as input variables, was obtained (Equation 3). This equation shows statistical quantities similar to those obtained by the FA-MLR method (Table 2). However, it shows slightly higher calibration and lower cross-validation statistics with respect to Equation 2. This shows a sign of overfitting since the factors considered in Equation 3 have information from irrelevant descriptors too. Considering this information in modeling, it may apparently increase the model variances (i.e., R^2) but they are not useful for prediction. On the other hand, the advantage of the QSAR model obtained by PCRA is that the factors that appear in the MLR Equation 6 are orthogonal. The regression coefficients calculated for such variables are more stable and thus are closer to the real values. In addition, from the factor scores used, significance of the original variables for modeling the activity, can be obtained. Factor score 1 indicates the importance of constitutional (MW, Ss, nAB), aromaticity indices (HOMT) and atom-centered fragment (C-040) descriptors. The factor score 2 indicates the importance of (X1A), geometrical (MAXDN), constitutional (Ss), 2D autocorrelations (ATS6e, ATS7e) and quantum (DipX) descriptors, and factor score 3 signifies the importance of topological (X4Av), constitutional (AMW) and 2D autocorrelation (MATS5p, GATS2p) descriptors. The factor score 9 reveals the importance of the 2D autocorrelation parameters (MATS3p, GATS3v). The factor score 11 signifies the importance of topological (IC2), and 2D autocorrelation (MATS7v, MATS7p) descriptors. The factor score 13 indicates the importance of only geometrical (J3D) descriptors and finally, the factor score 14 shows the importance of 2D autocorrelation (MATS7v) descriptors.

3.5. GA-PLS

In PLS analysis, having decomposed the descriptors data matrix to orthogonal matrices, the scores are constrained to have inner relationship with the dependent variables. Hence; similar to PCRA, the multicollinearity problem in the descriptors is omitted by PLS analysis. Genetic algorithm was applied to find the more useful set of descriptors in PLS modeling. So, many different GA-PLS runs were done using different initial set of populations. As it is shown in Table 2, in

Equation 4 (the best equation in GA-PLS model because of its greatest statistical parameters) a combination of quantum (DipX), 2D autocorrelations (MATS7v, ATS2e, ATS7v), atom-centered fragments (C-040) and chemical (HE) descriptors have been selected by GA-PLS to account for the cytotoxic activity of N-phenyl dichloroacetamide derivatives. The resulted GA-PLS model possessed very high statistical quality parameters (i.e., $R^2=0.98$ and $Q^2=0.94$). The predictive ability of the model was measured by application to 10 external test set molecules. The squared correlation coefficient for prediction was 0.91, and standard error of prediction was 0.202.

The brief description of the descriptors used by QSAR models are summarized in Table 5.

3.6. GA-PLS

In silico research in medicine is thought to have the potential to speed the rate of discovery, predicting and identifying new biologically active compounds while reducing the need for expensive lab work and clinical trials. One way to attain this is by generating and screening drug candidates more effectively. On the other hand, the *in silico* procedure, minimizes the time and cost associated with identifying new leads (37, 38).

A virtual screening was applied by deletion, insertion and substitution of different substitutes on the parent molecules and the effects of the structural modifications on the biological activity were investigated. Then, the domain application of QSAR model was determined to apply the model for screening new compounds. The applicability domain (AD) of QSAR model was used to verify the prediction reliability, to identify the troublesome compounds and to predict the compounds with acceptable activity that falls within this domain.

The important descriptors selected by GA-PLS model (chosen as the best model because of its greatest statistical parameters compared to the others) could be used for designing new active compounds. Analyzing the model applicability domain (AD) in the Williams plot (Figure 1) of the GA-PLS model based on the whole data set, showed that none of the compounds were identified as an obvious outlier for the cytotoxic activ-

ity, if the limit of normal values for the Y outliers (response outliers) was set as 2.5 times of the standard deviation units. As it is understood, none of the compounds have leverage (h) values greater than the threshold leverages (h^*). The warning leverage (h^*), was found to be 0.87 for the developed QSAR model. The compounds that had a standardized residual more than three times of the standard deviation units, were considered to be outliers. For both the training set and prediction set, the presented model matches the high quality parameters with good fitting power and the capability of assessing external data. Moreover, almost all of the compounds were within the applicability domain of the proposed model and were evaluated accurately while chemicals with a leverage value higher than h^* were considered to be influential or high leverage chemicals (17, 25).

Next, the *in silico* screening was used to design new compounds with potential cytotoxic activity according to the developed QSAR model and was validated by the developed GA-PLS model. So, the compounds in Table 1 with $IC_{50} < 12.5 \mu m$ were selected as template due to their good cytotoxic activity. Then, the *in silico* screening was applied by substituting different bioisosteric groups in the NH and the phenyl ring; the results of this investigation are summarized in Table 6.

The model tolerated various heterocyclic ring substituents in replacement of phenyl ring and bioisosteric changes of NH groups by CH_2 and oxygen groups, considering the fact that all of the

studied derivatives were within the applicability domain. Among different designated molecules, the compound 1a, 1g, 1i, 1j, 3n, 3w showed the best activity ($pIC_{50} > 5.15$). Thus, in order to clarify the relation between the activities of the compounds with different functional groups, this compound was chosen for more structural modifications. As it was shown in Table 6, esteric derivatives of DCA have good potential for becoming an anticancer agent. Finally, this result confirms the reliability of the models and shows that with the aim of the QSAR model and use of *in silico* screening, it is possible to identify new synthetic compounds for drug discovery.

The proposed QSAR models have all conditions to be considered as predictive models. Firstly, all have correlation coefficient of cross-validation (Q^2) larger than 0.5 and of prediction (r^2) higher than 0.6. Thus, according to great statistics, GA-PLS can be considered as the most predictive model. According to the cross-validation results, all models have $Q^2 > 0.7$ and can be considered as predictive models. To have a consideration on the cross-validated prediction results, the predicted activity data are plotted against the experimental activities in Figure 2. It should be mentioned that the least scattering of data was obtained from GA-PLS.

Table 2 shows that none of the proposed QSAR models were obtained by chance and the GA-PLS model, because of its greatest statistical parameters, is the best predictive model.

3.7. Docking Studies

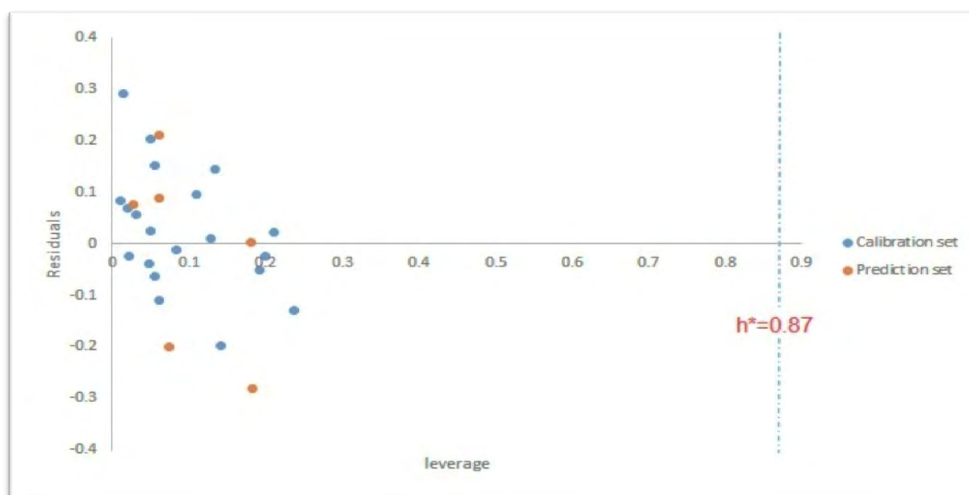


Figure 1. Williams plot for the training set and external prediction set for cytotoxic activity of N-arylphenyl-2,2-dichloroacetamide analogues.

Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays a great role in the rational design of drugs. DCA stimulates the activity of the enzyme PDH through inhibition of the enzyme PDKs. The crystal structure of PDK2 in complex with DCA has been acquired, and it shows that DCA indwells the pyruvate binding site in the N-terminal regulatory (R) domain (2).

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. The docking simulations were carried out by means of an in house batch script (DOCKFACE) for automatic running of AutoDock 4.2 in a parallel mode, using all system resources. Having completed the docking process, the protein–ligand complex was analyzed to investigate the type of interactions. Top ranked binding energies (kcal/mol) in AutoDock dlG output file, were considered as response in each run. Docking results were supported almost by high cluster populations. The conformation with the lowest binding energy was considered as the best docking result in each case.

As it was shown in Figure 3, there is a good correlation between experimental pIC_{50} and docking binding energy. Hence, our docking protocol can discriminate between the ligand (active) and decoys (non-active). The validated docking protocol was also applied to our designated

ligands. Compounds 1a, 1j, 3n and 3w based on their highest docking binding energy can be a good candidate for synthesis. It should be emphasized that there is a good correlation between the QSAR and docking results.

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. Hydrogen bindings between docked potent agents such as 3g and the PDK receptor (2BU8) were analyzed using Autodock tools program (ADT, Version 1.5.6), ligplotv.4.5.3 (39) and LigandScout 3.12 (40). As it is shown in Figure 4, a hydrogen bond acceptor interaction exists between oxygens of carboxyl group of co-crystal ligand (DCA) and Arg154, Tyr80 in the receptor (Figure 3A). Meanwhile, a hydrogen bond acceptor interaction exists between oxygen of nitro group of 2d and Arg158, Arg 154, Arg 112 in the receptor, there also exists an arene-arene interaction between phenyl group of compound 2d with the imidazole ring of His115 in the receptor (Figure 3B).

It is clear that score analysis of the docking process is not capable of detecting all active compounds due to a bad evaluation of the ligand binding energies and protein ligand interaction fingerprint (PLIF) studies could be used as a more reliable analysis technique (30). This method makes it possible to study the effect of different starting states of the structures on generated poses as well as their corresponding vectors of contacts towards

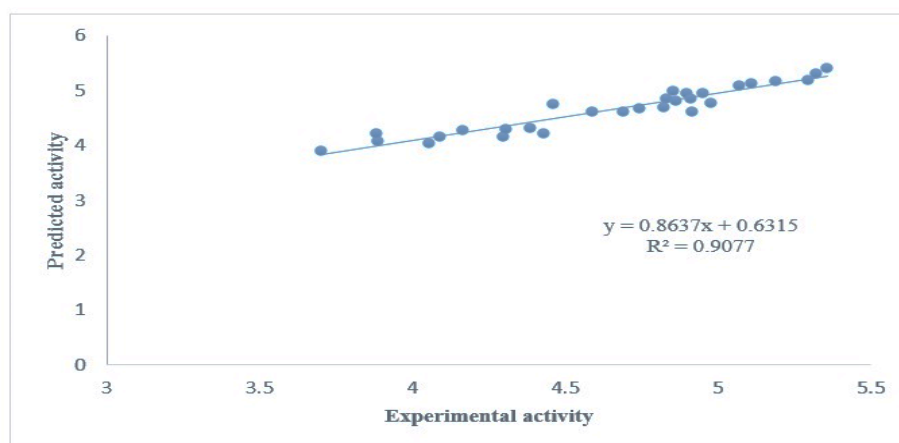


Figure 2. Plots of cross-validated predicted values of activity by GA-PLS against the experimental values.

receptors during docking procedure. For this purpose, the docking of all 31 compounds of the QSAR study as well as our designated compounds (Table 1, 6) were carried out, then all generated poses of the ligands were subjected to AuPosSOM 2.1 to calculate their contact vectors within the receptor binding cavity and appraisal of docking results based on the clustering of ligands by the resemblance of their contacts with target. In this procedure, the contacts between the structures and the protein, comprise of hydrophobic, hydrogen bonding and coulombic interactions. The resulted vectors of contacts are then analyzed using self-organizing map as implemented in AuPosSOM software. The output of self-organizing map is a classification pattern for ligands. For visualization of the results, the output files were subjected to Dendroscope 3.2.10. To the best of our knowledge, ligands in the same subgroup show a similar behavior. Having chosen the compound 5b as the best compound due to its greatest experimental cytotoxic activity, the PLIF results of our designated compounds were compared to this structure. As it was shown in Figure 5, the designated ligands such as 1j, 1i, 1f, 2t, 2v, 3p, 3q and 3s are clustered in the 5b subgroup. Therefore these compounds can be good candidates for synthesis.

4. Conclusion

In this study, four different QSAR modeling methods, MLR, FA-MLR, PCR and GA-PLS as well as FWA were used in the construction of

a QSAR model for cytotoxic activity of N-phenyl dichloroacetamide derivatives and the resulting models were compared. As it was shown in the article, having performed GA before the calibration, a regression model with enhanced predictive power would be obtained. The reliability, accuracy and predictability of the proposed models were evaluated by various criteria, including cross-validation, the root mean square error of prediction (RMSEP), root mean square error of cross-validation (RMSECV), validation through and Y-randomization. It was also shown that the proposed model is a useful aid for reduction of the time and cost of synthesis and biological evaluation of DCA analogues. Moreover, the results confirm that among the applied models, the GA-PLS is superior for the prediction of the pIC_{50} of DCA analogues. The statistical parameters of the four different chemometric methods used in this study are represented in Table 2. 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. To our knowledge, GA-PLS is the best choice for the prediction purpose of QSAR study, and for descriptive purpose it should be better to use the MLR method. The cross-validation statistics reported in Table 2 suggest the higher prediction ability of the GA-PLS model. This can be ascribed to the exploit of a large number of descriptors by GA-PLS compared to the MLR. The study suggests the importance of dipole moment in x-direction (DMX), 2D autocorrelations and hydration energy (HE) of

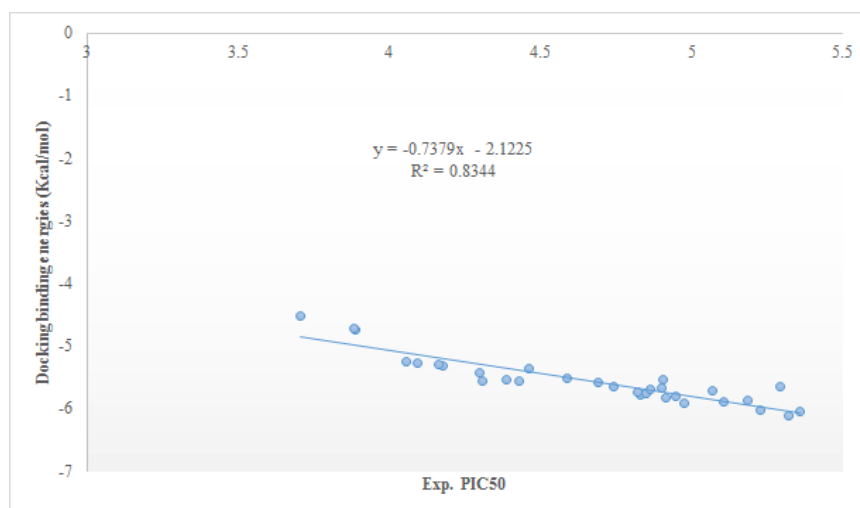


Figure 3. Plots of experimental pIC_{50} values versus docking binding energy.

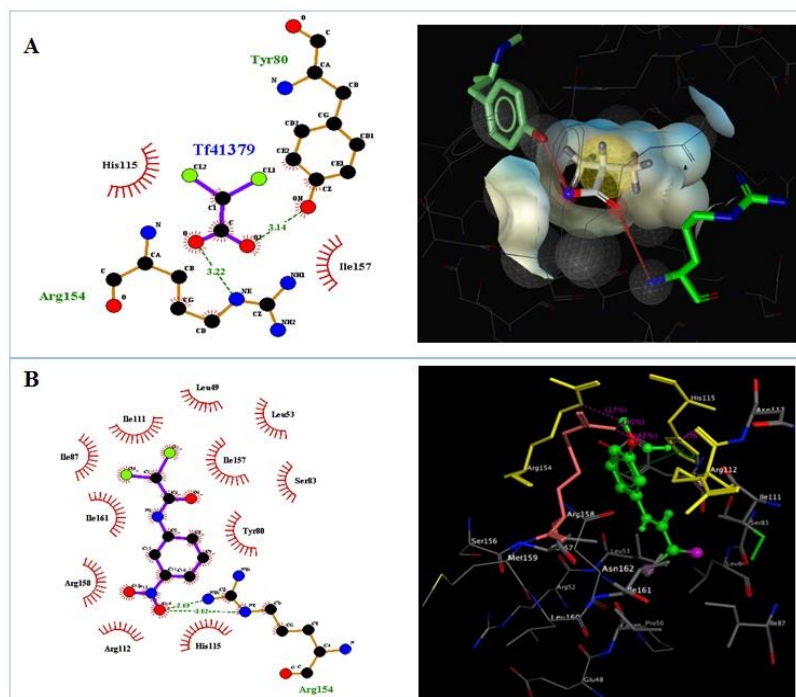


Figure 4. Interactions of A) DCA and B) compound 2d with the residues in the binding site of PDK (2BU8) receptor.

molecules for DCA derivatives' cytotoxic activity. It is clearly understood that 2D autocorrelation descriptors such as MATS7v, ATS7v, ATS2e and quantum chemical parameter (DMX) are important structural parameters that significantly influence the cytotoxic activity. The 2D autocorrelation descriptors depict the topological structure of the compounds, but are more complicated in nature

with respect to the classical topological descriptors. The calculation of these descriptors includes the summations of different autocorrelation functions, corresponding to different structural lags and leads to different autocorrelation vectors, corresponding to the lengths of the substructural fragments. As a result, these descriptors address the topology of the structure or parts thereof in

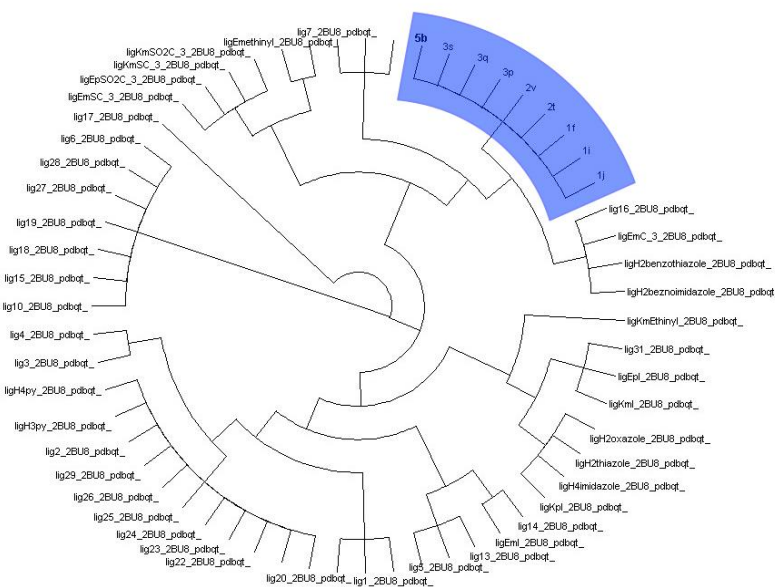


Figure 5. ApusSOM results for poses of docking.

association with a specific physicochemical property. According to the developed QSAR model, in silico screening was applied and new compounds such as 1a, 1g, 1i, 1j, 3n, 3w with potential cytotoxic activity were suggested for synthesis.

The docking study revealed that, there exists an arene-arene interaction between phenyl group of our ligands with imidazole ring of His115 in the receptor and based on the substituents on the phenyl group there might exist a hydrogen bond interaction with Arg158, Arg 154, Arg 112 in the receptor. However, because our biological data is merely cytotoxicity data, not enzyme (PDK) inhibitory data, and the docking estimation of the ligand binding energies is not good enough, no good relation between pIC_{50} and docking energy exists. Therefore the docking results were subjected

to PLIF studies to distinguish active compounds from inactive ones with particular analysis of interatomic contacts between the ligand- receptor complexes. As a result, compounds 1j, 1i, 1f, 2t, 2v, 3p, 3q and 3s are introduced as good candidates for synthesis.

Acknowledgment

The authors would like to thank the department of medicinal chemistry at school of pharmacy, Shiraz University of Medical Sciences for its kind contribution in providing the needed facilities for this work.

Conflict of Interest

None declared.

5. References

- Hans HK, Kim T, Kim E, Park JK, Park S, Joo H, *et al.* The Mitochondrial Warburg Effect: A Cancer Enigma. *IBC*. 2009;1:7.
- Papandreou I, Goliassova T, Denko NC. Anticancer drugs that target metabolism: Is dichloroacetate the new paradigm? *Int J Cancer*. 2011;128:1001-8.
- Stockwin LH, Yu SX, Borgel S, Hancock C, Wolfe TL, Phillips LR, *et al.* Sodium dichloroacetate selectively targets cells with defects in the mitochondrial ETC. *Int J Cancer*. 2010;127:2510-9.
- Kankotia S, Stacpoole PW. Dichloroacetate and cancer: New home for an orphan drug? *Biochim Biophys Acta*. 2014;1846:617-29.
- Abdelmalak M, Lew A, Ramezani R, Shroads AL, Coats BS, Langae T, *et al.* Long-term safety of dichloroacetate in congenital lactic acidosis. *Mol Genet Metab*. 2013;109:139-43.
- Bonnet S, Archer SL, Allalunis-Turner J, Haromy A, Beaulieu C, Thompson R, *et al.* A Mitochondria-K⁺ Channel Axis Is Suppressed in Cancer and Its Normalization Promotes Apoptosis and Inhibits Cancer Growth. *Cancer Cell*. 2007;11:37-51.
- Weaver S, Gleason MP. The importance of the domain of applicability in QSAR modeling. *J Mol Graph Model*. 2008;26:1315-26.
- Khoshneviszadeh M, Edraki N, Miri R, Hemmateenejad B. Exploring QSAR for Substituted 2-Sulfonyl-Phenyl-Indol Derivatives as Potent and Selective COX-2 Inhibitors Using Different Chemometrics Tools. *Chem Biol Drug Des*. 2008;72:564-74.
- Yang Y, Shang P, Cheng C, Wang D, Yang P, Zhang F, *et al.* Novel N-phenyl dichloroacetamide derivatives as anticancer reagents: design, synthesis and biological evaluation. *Eur J Med Chem*. 2010;45:4300-6.
- Fereidoonezhad M, Faghieh Z, Mojadami A, Tabaei S, Rezaei Z. Novel Approach Synthesis, Molecular Docking and Cytotoxic Activity Evaluation of N-phenyl-2, 2-dichloroacetamide Derivatives as Anticancer Agents. *J Sci I R Iran*. 2016;27:39-49.
- Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, *et al.* Gaussian 09. *Gaussian, Inc., Wallingford CT*, 2009.
- Mauri A, Consonni V, PavanM, Todeschini R. RAGON Software: An Easy Approach to Molecular Descriptor Calculations. *MATCH Commun Math Comput Chem*. 2006;56:237-48.
- Thanikaivelan P, Subramanian V, Raghava Rao J, Unni Nair B. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chem Phys Lett*. 2000;323:59-70.
- Morris GM, Huey R, Olson AJ. Using AutoDock for Ligand-Receptor Docking. *Curr Protoc Bioinformatics*. 2008;Chapter 8:Unit 8.14.
- 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. Khoshneviszadeh M, Edraki N, Miri R, Foroumadi A, Hemmateenejad B. QSAR Study of 4-Aryl-4H-Chromenes as a New Series of Apoptosis Inducers Using Different Chemometric Tools. *Chem Biol Drug Des*. 2012;79:442-58.
 18. Leonard JT, Roy K. QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. *Bioorg Med Chem*. 2006;14:1039-46.
 19. Sharaf MA, Illman DL, Kowalski BR. Chemometrics. New York: *Wiley*. 1986;332.
 20. Cho SJ, Hermsmeier MA. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J Chem Inform Comput Sci*. 2002;42:927-36.
 21. Leardi R. Genetic algorithms in chemometrics and chemistry: a review. *J Chemometr*. 2001;15:559-69.
 22. Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN. Quantitative Structure-Antitumor Activity Relationships of Camptothecin Analogues: Cluster Analysis and Genetic Algorithm-Based Studies. *J Med Chem*. 2001;44:3254-63.
 23. Edraki N, Hemmateenejad B, Miri R, Khoshneviszade M. Research article: QSAR Study of Phenoxypyrimidine Derivatives as Potent Inhibitors of p38 Kinase Using different Chemometric Tools. *Chem Biol Drug Des*. 2007;70:530-9.
 24. Miri R, Javidnia K, Mirkhani H, Hemmateenejad B, Sepeher Z, Zalpour M, *et al.* Synthesis, QSAR and Calcium Channel Modulator Activity of New Hexahydroquinoline Derivatives Containing Nitroimidazole. *Chem Biol Drug Des*. 2007;70:329-36.
 25. Asadollahi T, Dadfarnia S, Shabani AMH, Ghasemi JB, Sarkhosh M. QSAR Models for CXCR2 Receptor Antagonists Based on the Genetic Algorithm for Data Preprocessing Prior to Application of the PLS Linear Regression Method and Design of the New Compounds Using In Silico Virtual Screening. *Molecules*. 2011;16:1928-55.
 26. Tropsha A, Gramatica P, Gombar VK. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb Sci*. 2003;22:69-77.
 27. Roy K, Kar S, Das RN. Chapter 7-Validation of QSAR Models. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. *Boston: Academic Press*. 2015;231-89.
 28. Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SW, *et al.* Validation of Molecular Docking Programs for Virtual Screening against Dihydropteroate Synthase. *J Chem Inf Model*. 2009;49:444-60.
 29. Sakhteman A. PreAuposSOM, <https://www.biomedicale.univ-paris5.fr/aupossom/>.
 30. Mantsyzov AB, Bouvier G, Evrard-Todeschi N, Bertho G. Contact-based ligand-clustering approach for the identification of active compounds in virtual screening. *Adv Appl Bioinforma Chem*. 2012;5:61-79.
 31. Bouvier G, Evrard-Todeschi N, Girault JP, Bertho G. Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics*. 2010;26:53-60.
 32. Parameswaran S, Saudagar P, Dubey VK, Patra SM. Discovery of novel anti-leishmanial agents targeting LdLip3 lipase. *J Mol Graph Model*. 2014;49:68-79.
 33. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*. 2012;61:1061-7.
 34. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinf*. 2007;8:460.
 35. Free SM, Wilson JW. A Mathematical Contribution to Structure-Activity Studies. *J Med Chem*. 1964;7:395-9.
 36. Hemmateenejad B. Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based PCR. *J Chemom*. 2004;18:475-85.
 37. Bauch C, Bevan S, Woodhouse H, Dilworth C, Walker P. Predicting *in vivo* phospholipidosis-inducing potential of drugs by a combined high content screening and *in silico* modelling approach. *Toxicol in Vitro*. 2015;29:621-30.
 38. Murakami Y, Hayakawa M, Yano Y, Tanahashi T, Enomoto M, Tamori A, *et al.* Discovering novel direct acting antiviral agents for HBV using

in silico screening. *Biochem Biophys Res Commun.* 2015;456:20-8.

39. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein*

Eng. 1995;8:127-34.

40. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inf Model.* 2005;45:160-9.