

Original Research Articles

Multi-target In-silico study in the Analysis and Design of HIV-1 inhibitors

Abstract

The multi-target quantitative structure-activity relationship (mt-QSAR) study of human immunodeficiency virus (HIV-1) inhibitors was addressed by applying a modest, hitherto active linear regression model based on the Genetic function approximation. QSAR studies were performed on two datasets of HIV-1 inhibitors targeted on integrase and reverse transcriptase, respectively. By using the genetic function approximation method, the collaboration among different set of inhibitors was exploited and an efficient multi-target QSAR modeling for HIV-1 inhibitors was obtained. The predictive quality of the mt-QSAR models was tested for an external set of 30 compounds, randomly chosen out of 150 compounds. The linear regression model based on the Genetic function approximation with eight selected descriptors was obtained. The accuracy of the proposed model is illustrated using the following evaluation techniques: cross-validation, validation through an external test set, applicability domain, and Y-randomization. We accordingly propose a quantitative model, and we interpret the activity of the compounds relying on the multivariate statistical analysis. This study shows that the prediction results were in excellent agreement with the experimental value.

Keywords: multi-target, QSAR, HIV-1 inhibitors, GFA, DFT, applicability domain.

1.0 Introduction

The handling of the acquired immunodeficiency syndrome (AIDS) is the utmost challenging worldwide medical problem. So far, there is no realistic cure for HIV/AIDS. “Highly Active Antiretroviral Therapy” (HAART) is recommended for the treatment of HIV [1]. HAART is an aggressive treatment of HIV where the combination of different antiviral drugs is used to suppress HIV replication and the progression of the disease [2]. Most of the current strategies for treating AIDS depend on inhibiting HIV-1 reverse transcriptase enzyme. Multi-drug resistant is one of the major immediate threats to human health today [3], trends in the incidence of HIV together with the development of multi-drug and extensively drug resistant strains of HIV raises the need to intensify the search for more efficient drugs to combat this disease. The majority of existing therapy methods has targeted the viral replication at reverse transcriptase (RT), integrase and protease enzyme [4,5]. However, the emergence of drug resistance has been observed [6], therefore, new therapeutic agents are still needed. Recently, a new class of therapeutic agents has focused on inhibiting HIV entry into cells, CD4 binding, co-receptor binding and membrane fusion such as T-20 [7].

The multi-target drug design method is an encouraging way to complement the current single-target process and an embarrassment of studies address the problem of target prediction [8] and multi-target structure-activity models [9, 10]. The multi-target drug prediction is a current research topic in the field of drug design. Despite the positive results of the studies mentioned above, the considered models were still trained for each target separately. In this study, the multi-target QSAR study of HIV-1 inhibitors was addressed by applying a simple, yet effective linear regression model based on genetic function approximation, which is recently presented in machine learning community. QSAR studies were performed on two datasets of HIV-1 inhibitors

targeted on integrase and reverse transcriptase, by using the GFA method, the collaboration among different set of inhibitors was exploited and an efficient multi-target QSAR modeling for HIV-1 inhibitors was obtained. The general descriptor features and drug-like features for compound description were ranked according to their jointly importance in multi-target QSAR modeling respectively, which will offer useful hints for the design of novel multi-target HIV-1 inhibitors with increasing likelihood of successful therapies of HIV.

Computer-aided drug design techniques may play a very important role. These techniques are based on multi-target Quantitative Structure-Activity Relationship (mt-QSAR) studies. It means that they are models linking the structure of drugs with the biological activity against different targets [13]. This kind of study may also be useful in a Multi-Objective Optimization of desired properties or activity of drugs against different targets. There are over 5000 descriptors that may be comprehensive and used to solve these problem [14]. QSAR studies reported up-to-date are based on descriptors and databases of structurally parent compounds relevant to only one viral species. Subsequently, the researcher interested in predicting, for example, the antiviral activity for a given series of compounds, has to develop as many QSAR equations as combinations of compound families versus viral species have to be predicted. Therefore, it is of major interest the development of a single unified equation explaining the antiviral activity of structurally heterogeneous series of compounds against as many viral species as possible [15]. In fact, other mt-QSAR approaches, with demonstrated usefulness, have been introduced recently in Medicinal Chemistry [16]. The results of this study will go a long way to authenticate the claims by QSAR expert and will as well enrich the database on 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine, indole β -diketo acid, diketo acid and carboxamide derivatives with anti-HIV-1 activity that can be used in drug discovery with the development of rational/QSAR tools for decision support in anti-HIV therapy.

2.0 Materials and Methods

Our study was performed on two kinds of HIV target datasets conformed from a far-reaching literature review, which consisted of inhibitors with their binding affinities on HIV integrase and reverse transcriptase. These inhibitors are correspondingly referred as; integrase inhibitors, which inhibit the proviral DNA to insert into the host cell genome, and non-nucleoside reverse transcriptase inhibitors (NNRTI), which inhibit the virus by preventing the copying of its genomic DNA into proviral DNA for incorporation into the host cell DNA. The dataset containing 150 compounds with well-defined activity [17, 18], was selected for QSAR study. The biological activity data in the form of IC₅₀ and EC₅₀ (molar concentration of the drug leading to 50% inhibition of enzyme) value in μ M (micromoles) were converted into negative logarithmic dose in moles (pIC₅₀) for mt-QSAR Analysis (Table 1).

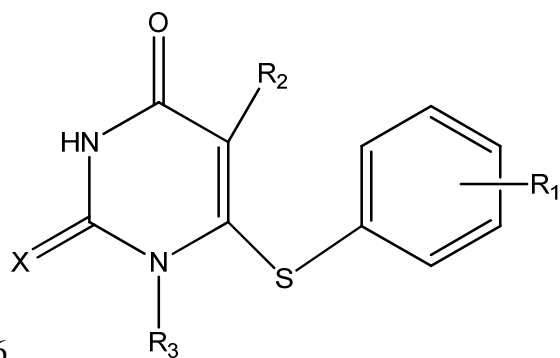


Figure 1: Compound 1-106

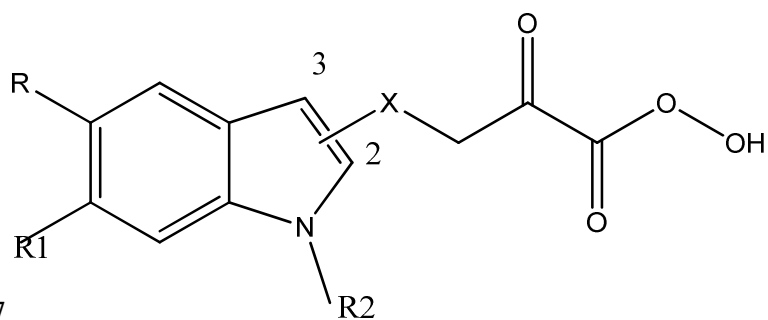


Figure 2: Compound 107-117

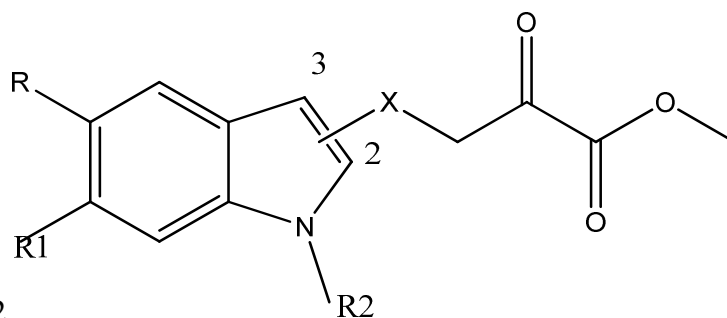


Figure 3: Compound 118-122

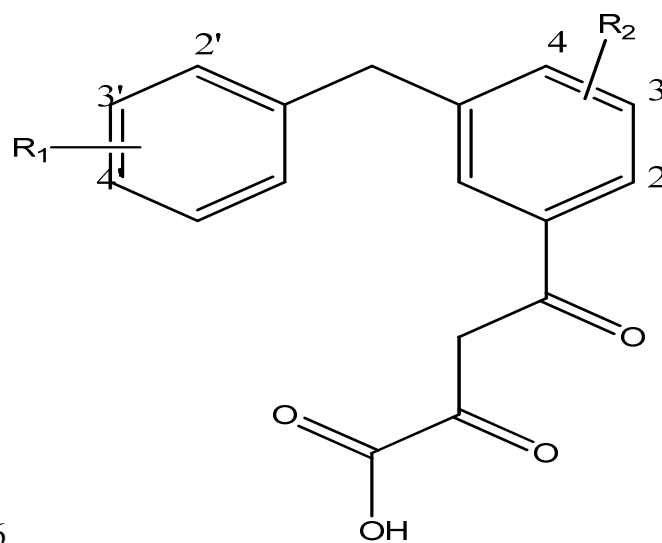


Figure 4: Compound 123-126

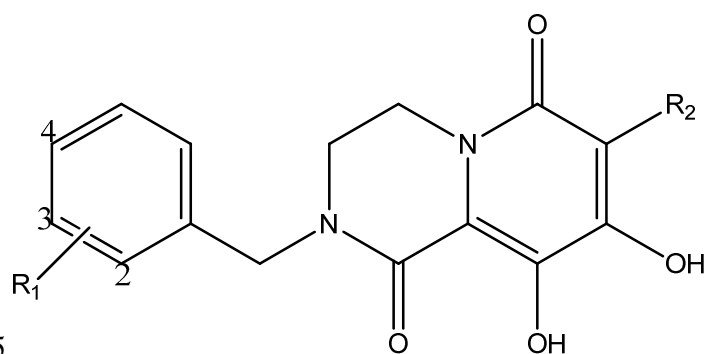


Figure 5: Compound 127-135

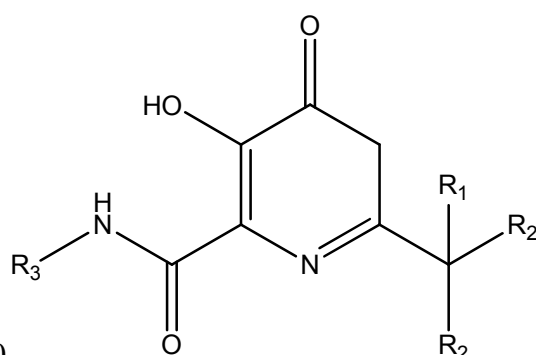


Figure 6: Compound 136-150

Table 1: Biological activities of the training and test set

No	R1	R2	R3	X	PIC50	Predicted PIC50	Residuals
1*	3-CN	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.000	4.679717	0.320283
2	3-COMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.140	4.731931	0.408069
3	3-COOMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.100	5.377841	-0.27784
4	3,5-Cl ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.890	6.185995	-0.29599
5	3,5-Me ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	6.590	5.906561	0.683439
6	3-OMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.660	4.811973	-0.15197
7	3-OH	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.090	3.771356	0.318644

8	3-NO ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.470	4.807413	-0.33741
9	3-I	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.000	5.247345	-0.24734
10	3-Br	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.240	5.049991	0.190009
11	3-Cl	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.890	4.76618	0.12382
12	3-F	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.480	4.510582	0.969418
13	3-CF ₃	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.350	4.971573	-0.62157
14	3-t-Bu	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.920	4.985295	-0.0653
15	3-Et	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.570	5.098838	0.471162
16	3-Me	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.590	4.73109	0.85891
17	2-OMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.720	4.854167	-0.13417
18	2-NO ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.850	4.488034	-0.63803
19	2-Me	Me	CH ₂ OCH ₂ CH ₂ OH	O	4.150	3.865752	0.284248
20	H	Et	CH ₂ OCH ₂ CH ₂ OH	O	6.920	5.424086	1.495914
21	H	i-Pr	CH ₂ OCH ₂ CH ₂ OH	O	7.200	6.575076	0.624924
22	3,5-Me ₂	Et	CH ₂ OCH ₂ CH ₂ OH	O	7.890	7.024387	0.865613
23	3,5-Me ₂	i-Pr	CH ₂ OCH ₂ CH ₂ OH	O	8.570	8.175459	0.394541
24	3,5-Cl ₂	Et	CH ₂ OCH ₂ CH ₂ OH	O	7.850	7.303889	0.546111
25	H	Me	CH ₂ OCH ₂ CH ₂ OH	O	5.150	4.306412	0.843588

26	H	I	CH ₂ OCH ₂ CH ₂ OH	O	5.440	5.22504	0.21496
27	H	CH=CPH ₂	CH ₂ OCH ₂ CH ₂ OH	O	6.070	5.93367	0.13633
28	4-F	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.600	4.257113	-0.65711
29	4-Cl	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.600	3.948451	-0.34845
30	4-OH	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.560	3.972844	-0.41284
31	3- CONH ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.510	4.550109	-1.04011
32	H	COOMe	CH ₂ OCH ₂ CH ₂ OH	O	5.180	5.415909	-0.23591
33	H	CONHPh	CH ₂ OCH ₂ CH ₂ OH	O	4.740	4.615358	0.124642
34	H	SPh	CH ₂ OCH ₂ CH ₂ OH	O	4.840	5.941528	-1.10153
35*	H	CCH	CH ₂ OCH ₂ CH ₂ OH	O	4.740	3.591406	1.148594
36*	H	CCPh	CH ₂ OCH ₂ CH ₂ OH	O	5.470	4.300706	1.169294
37	H	COCHMe ₂	CH ₂ OCH ₂ CH ₂ OH	O	4.920	5.387378	-0.46738
38	H	COPh	CH ₂ OCH ₂ CH ₂ OH	O	4.890	5.363464	-0.47346
39	H	CCMe	CH ₂ OCH ₂ CH ₂ OH	O	4.720	4.332714	0.387286
40	H	F	CH ₂ OCH ₂ CH ₂ OH	O	4.000	3.47608	0.52392
41	H	Cl	CH ₂ OCH ₂ CH ₂ OH	O	4.520	4.087923	0.432077
42*	H	Br	CH ₂ OCH ₂ CH ₂ OH	O	4.700	4.719097	-0.0191

43	2-Cl	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.890	3.764952	0.125048
44	3- CH ₂ OH	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.530	4.580372	-1.05037
45*	4-NO ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.720	4.314184	-0.59418
46	4-CN	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.600	4.33906	-0.73906
47	4-OMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.600	4.560737	-0.96074
48	4-COMe	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.960	4.430108	-0.47011
49*	4-COOH	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.450	3.732212	-0.28221
50	3-NH ₂	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.600	3.828573	-0.22857
51	H	Pr	CH ₂ OCH ₂ CH ₂ OH	O	5.470	5.26521	0.20479
52	4-Me	Me	CH ₂ OCH ₂ CH ₂ OH	O	3.660	4.009684	-0.34968
53	H	CH=CH ₂	CH ₂ OCH ₂ CH ₂ OH	O	5.690	4.729078	0.960922
54*	H	CH=CHPh	CH ₂ OCH ₂ CH ₂ OH	O	5.220	4.712519	0.507481
55	H	CH ₂ Ph	CH ₂ OCH ₂ CH ₂ OH	O	4.370	5.078417	-0.70842
56*	H	Me	CH ₂ OCH ₂ CH ₂ OAc	O	5.170	5.659474	-0.48947
57	H	Et	CH ₂ OCH ₂ Me	O	7.720	6.709459	1.010541
58	H	Et	CH ₂ CH ₂ Ph	O	8.230	6.826881	1.403119
59	3,5-Cl ₂	Et	CH ₂ CH ₂ Me	O	8.130	8.696731	-0.56673

60	H	Me	CH ₂ OCH ₂ CH ₂ OC ₅ H ₁₁	O	4.460	5.319991	-0.85999
61	H	Me	CH ₂ OCH ₂ CH ₂ OCH ₂ Ph	O	4.700	5.330732	-0.63073
62	H	Me	H	O	3.600	3.096354	0.503646
63	H	Me	Me	O	3.820	4.905933	-1.08593
64*	H	c-Pr	CH ₂ OCH ₂ Me	O	7.000	6.708195	0.291805
65*	H	Et	CH ₂ O-i-Pr	O	6.470	6.92136	-0.45136
66	H	Et	CH ₂ O-c-Hex	O	5.400	6.418495	-1.0185
67	H	Et	CH ₂ OCH ₂ -c-Hex	O	6.350	5.971166	0.378834
68	H	Et	CH ₂ OCH ₂ CH ₂ Ph	O	7.020	6.796898	0.223102
69*	H	Me	CH ₂ OMe	O	5.680	5.683756	-0.00376
70	H	Me	CH ₂ OBu	O	5.330	5.594366	-0.26437
71	H	Me	Et	O	5.660	5.852396	-0.1924
72*	H	Me	Bu	O	5.920	5.636009	0.283991
73	H	i-Pr	CH ₂ OCH ₂ Me	O	7.990	7.858364	0.131636
74*	H	i-Pr	CH ₂ OCH ₂ Ph	O	8.510	7.872493	0.637507
75	3,5-Me ₂	Et	CH ₂ OCH ₂ Ph	O	8.550	8.321546	0.228454
76	3,5-Me ₂	Et	CH ₂ OCH ₂ Me	O	8.240	8.307485	-0.06748
77	H	Me	CH ₂ OCH ₂ CH ₂ OMe	O	5.060	5.372858	-0.31286

78*	H	Me	CH ₂ OCH ₂ CH ₂ OCOPh	O	5.120	5.845098	-0.7251
79	H	Me	CH ₂ OCH ₂ Me	O	6.480	5.590438	0.889562
80	H	Me	CH ₂ OCH ₂ CH ₂ Cl	O	5.820	5.39062	0.42938
81*	H	Me	CH ₂ OCH ₂ CH ₂ N ₃	O	5.240	5.361999	-0.122
82	H	Me	CH ₂ OCH ₂ CH ₂ F	O	5.960	5.289403	0.670597
83	H	Me	CH ₂ OCH ₂ CH ₂ Me	O	5.480	5.594396	-0.1144
84*	H	Me	CH ₂ OCH ₂ Ph	O	7.060	5.604552	1.455448
85	H	Et	CH ₂ OCH ₂ Me	S	7.580	7.021186	0.558814
86*	H	i-Pr	CH ₂ OCH ₂ Me	S	7.890	8.1703	-0.2803
87	H	i-Pr	CH ₂ OCH ₂ Ph	S	8.140	8.183494	-0.04349
88	3,5-Cl ₂	Et	CH ₂ OCH ₂ Me	S	7.890	8.902717	-1.01272
89*	H	Et	CH ₂ O-i-Pr	S	6.660	7.233555	-0.57356
90	H	Et	CH ₂ O-c-Hex	S	5.790	6.73068	-0.94068
91	H	Et	CH ₂ OCH ₂ -c-Hex	S	6.450	6.282895	0.167105
92	H	Et	CH ₂ OCH ₂ C ₆ H ₄ (4-Cl)	S	7.920	6.622962	1.297038
93	H	Et	CH ₂ OCH ₂ CH ₂ Ph	S	7.040	7.108549	-0.06855
94	H	c-Pr	CH ₂ OCH ₂ Me	S	7.020	7.02	-7.1E-15
95*	3,5-Me ₂	Me	CH ₂ OCH ₂ CH ₂ OH	S	6.660	6.127552	0.532448

96	H	Pr	CH ₂ OCH ₂ CH ₂ OH	S	5.000	5.487097	-0.4871
97*	3,5-Me ₂	i-Pr	CH ₂ OCH ₂ CH ₂ OH	S	8.300	8.400069	-0.10007
98*	H	Et	CH ₂ OCH ₂ Ph	S	8.090	7.034523	1.055477
99*	3,5-Me	Et	CH ₂ OCH ₂ Ph	S	8.140	8.632577	-0.49258
100	3,5-Me ₂	Et	CH ₂ OCH ₂ Me	S	8.300	8.619221	-0.31922
101*	H	Et	CH ₂ OCH ₂ CH ₂ OH	S	6.960	5.644658	1.315342
102*	H	i-Pr	CH ₂ OCH ₂ CH ₂ OH	S	7.230	6.797403	0.432597
103	3,5-Me ₂	Et	CH ₂ OCH ₂ CH ₂ OH	S	8.110	7.247153	0.862847
104*	3,5-Cl ₂	Et	CH ₂ OCH ₂ CH ₂ OH	S	7.370	7.526883	-0.15688
105*	H	Me	CH ₂ OCH ₂ CH ₂ OH	S	6.010	4.525119	1.484881
106*	H	CH ₂ CH=CH ₂	CH ₂ OCH ₂ CH ₂ OH	O	5.600	5.017075	0.582925

100

No.	R	R1	R2	X	PIC50	Predicted PIC50	Residuals
107	H	H	CH ₃	2-CO	0.7780	0.135809	0.642191
108	OCH ₂ O		CH ₃	2-CO	0.3010	0.220012	0.220012
109	H	H	CH ₂ CH ₃	2-CO	0.2040	0.41899	-0.21499
110	OCH ₂ O		CH ₂ CH ₃	2-CO	0.6990	0.513655	0.513655
111	H	H	Bn	2-CO	0.0000	0.591473	-0.59147
112	OCH ₂ O		Bn	2-CO	0.3010	0.686876	0.686876
113	H	H	CH ₃	3-CO	0.3010	0.35025	-0.04925

114	OCH ₂ O		CH ₃	3-CO	0.4770	0.387358	0.387358
115	H	H	CH ₂ CH ₃	3-CO	0.4770	0.472458	0.004542
116	OCH ₂ O		CH ₂ CH ₃	3-CO	0.4770	0.472458	0.520857
117*	H	H	Bn	3-CO	0.0000	0.814558	-0.81456

101

No.	R	R1	R2	X	PIC50	Predicted PIC50	Residuals
118	H	H	CH ₃	2-CO	1.6530	1.291472	0.361528
119*	OCH ₂ O		CH ₃	2-CO	1.6990	1.375696	0.323304
120	OCH ₂ O		CH ₂ CH ₃	2-CO	1.8130	1.669318	0.143682
121	OCH ₂ O		CH ₃	3-CO	1.7780	1.543228	0.234772
122	H	H	CH ₂ CH ₃	3-CO	1.628327	1.628327	-0.21333

102

No.	R1	R2	PIC50	Predicted PIC50	Residuals
123	4'-Cl	-	0.000	0.266536	-0.26654
124	3'-F	-	0.602	0.692933	-0.09093
125	-	4-OCH ₃	0.824	1.106334	-0.28233
126	-	3-OCH ₃	0.854	1.15166	-0.29766

103

No.	R1	R2	PIC50	Predicted PIC50	Residuals
127	4-F	-	1.000	2.398653	-1.39865
128	H	-	0.638	1.87317	-1.23517
129	2-Cl	-	0.432	2.125824	-1.69382
130	3-Cl	-	1.398	0.16283	1.23517
131	4-Cl	-	0.420	1.648111	-1.22811

132	4-F, 3-Cl	-	1.398	1.528234	-0.13023
133	4-F	CN	1.699	0.345914	1.353086
134	4-F	Br	1.523	1.109773	0.413227
135	4-F	I	1.699	1.638357	0.060643

104

No.	R1	R2	R3	PIC50	Predicted PIC50	Residuals
136	NHCOCH ₃	CH ₃	4-fluorotoluene	2.1555	1.577784	0.577216
137	NH-SO ₂ -CH ₃	CH ₃	4-fluorotoluene	2.097	1.998185	0.098815
138	NHCO-N(CH ₃) ₂	CH ₃	4-fluorotoluene	1.745	1.620846	0.124154
139	NHSO ₂ -N(CH ₃) ₂	CH ₃	4-fluorotoluene	1.921	1.465967	0.455033
140	NHCOCO-N(CH ₃) ₂	CH ₃	4-fluorotoluene	2.000	0.905856	1.094144
141	NHCOCO-OCH ₃	CH ₃	4-fluorotoluene	1.824	1.363955	0.470045
142	NHCOCO-OH	CH ₃	4-fluorotoluene	2.398	1.920462	0.477538
143	N(CH ₃)COCO- N(CH ₃) ₂	CH ₃	4-fluorotoluene	1.824	2.028732	-0.20473
144*	NHCO-pyridine	CH ₃	4-fluorotoluene	1.699	1.485309	0.213691
145	NHCO-pyridazine	CH ₃	4-fluorotoluene	1.824	2.094595	-0.27059
146	NHCO-pyrimidine	CH ₃	4-fluorotoluene	2.155	1.614115	0.540885
147	NHCO-oxazole	CH ₃	4-fluorotoluene	2.155	1.698617	0.456383
148	NHCO-thiazole	CH ₃	4-fluorotoluene	2.097	2.427144	-0.33014
149	NHCO-1H imidazole	CH ₃	4-fluorotoluene	2.222	2.430314	-0.20831
150	NHCO-1,3,4-	CH ₃	4-fluorotoluene	1.824	2.731546	-0.90755

	oxadiazole					
--	------------	--	--	--	--	--

*indicates the compounds considered in the test set

2.1 Molecular modeling and generation of molecular descriptors

The dual core personal computer equipped with the operating system Windows seven was used for making calculations of this work. Structure of all the compounds was drawn using ChemDraw Ultra module of the program and transferred to Spartan'14 (2013) version 1.1.2 [19] module to create the three-dimensional (3D) structure. These structures were then subjected to energy minimization using molecular mechanics (MMFF). Energy minimized molecules were subjected to optimization via DFT (density function theory) method with B3LYP function [20] and 6-311G* basic set [21]. These methods have become popular in recent years because they can reach similar precision to other methods in less time and less cost from the computational point of view. The geometry optimization of the lowest energy structure was carried out without any symmetry constraints were also transferred to PaDEL-Descriptor [22] version 2.18 and were subjected to re-optimization (with the MMFF94 force field). Most stable structure for each compound was generated and used for calculating various physicochemical parameters used for the statistical analysis.

2.2. Variable selection and model generation

Even though many molecular descriptors are available, only a subclass of them is statistically important in terms of correlation with biological activity. Therefore, it is very important to address the variable selection method for deriving the best QSAR model. GFA [23] approach were adopted to select the best possible variables as well as for the generation of QSAR models.

2.2.1. Genetic Function Approximation method

GFA [23] approach is a search method to find approximate solutions to optimization and search problems. GFA is conceived from

(i) Genetic algorithm and

(ii) Friedman's multivariate adaptive regression splines (MARS) algorithm.

The following steps were performed:

(i) Initial population of equations were generated by random number of descriptors,

(ii) Pairs from the population of equations were chosen at random, crossovers were performed and posterity equations were generated,

(iii) The fitness of each posterity equation was assessed by lack of fit (LOF) score that automatically penalizes models with too many features. A distinctive feature of GFA is that it generates a population of equations rather than a single equation as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors. By examining these models, additional information can be obtained. For example, the frequency of use of a particular descriptor in the population of equations may indicate how relevant the descriptor is to the prediction of activity. The fitness function, i.e., lack-of-fit is calculated by

$$LOF = \frac{LSE}{\left(1 - \frac{c+dp}{m}\right)^2} \quad (1)$$

Where c is the number of basis functions, d is the smoothing parameter, m is the number of samples in the training set, LSE is the least square error and p is the total number of features contained in all basis functions. Material Studio version 7.0 was used for GFA.

2.3 Validation of the QSAR model

The predictive capability of the QSAR equation was determined using the leave-one-out cross-validation method. The cross-validation regression coefficient (Q_{cv}^2) was calculated by the following equation:

$$Q_{cv}^2 = 1 - \frac{\sum(Y_{exp} - Y_{pred})^2}{\sum(Y_{exp} - \bar{Y}_{exp})^2} \quad (2)$$

Where Y_{pred} , Y_{exp} , and $\bar{Y}_{exp}$ are the predicted, experimental, and mean values of experimental activity, respectively. Also, the accuracy of the prediction of the QSAR equation was validated by F value, and R^2 . The R^2 value can be generally increased by adding the additional predictor variables to the model, even if the added variable does not contribute to the reduction of the unexplained variance of the dependent variable. Therefore, the R^2 usage requires special attention. For this reason, it is better to use another statistical parameter, called the adjusted R^2 (R_{adj}^2), where R_{adj}^2 is defined by;

$$R_{adj}^2 = 1 - (1 - R^2) \frac{N-1}{N-P-1} \quad (3)$$

R_{adj}^2 is interpreted similarly to the R^2 value, considering the number of degrees of freedom also. It is adjusted by dividing the residual sum of squares and total sum of squares by their respective degrees of freedom. The R_{adj}^2 value diminishes if an added variable to the equation does not reduce the unexplained variance [24]. Subsequently, R_{adj}^2 is used to compare models with different numbers of predictor variables.

A large F indicates that the model fit is not a chance occurrence. It has been shown that a high value of statistical characteristics is not necessary for the proof of a highly predictive model [25, 26]. Hence, to evaluate the predictive ability of our QSAR model, we used the method described by Golbraikh and Tropsha [25] and Roy and Roy [26]. The values of the correlation coefficient of predicted and actual activities and the correlation coefficient for regressions through the origin (predicted vs. actual activities and vice versa) were calculated using the regression of analysis Tool-pak option of Excel, and other parameters were calculated as reported by the above authors [25, 26]. The determination coefficient in prediction, Q_{test}^2 , was calculated using the following equation [26]:

$$Q_{test}^2 = 1 - \frac{\sum(Y_{pred_{test}} - Y_{test})^2}{\sum(Y_{test} - \bar{Y}_{training})^2} \quad (4)$$

where $Y_{pred_{test}}$ and Y_{test} are the predicted value based on the QSAR equation (model response) and experimental activity values, respectively, of the external test set compounds. $\bar{Y}_{training}$ is the mean activity value of the training set compounds. Quality factor (Q) is calculated as;

$$Q = \frac{R}{SEE} \quad (5)$$

Where R is variance and SEE is the standard error of estimate. Over fitting and chance correlation, due to excess number of predictor variables can be detected by Q value [27, 28]. Positive value of this QSAR model suggests its high predictive power and lack of over fitting [29].

Further evaluation of the predictive ability of the QSAR model for the external test set compounds was done by determining the value of r_m^2 by the following equation [26]:

$$r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2}) \quad (6)$$

where r_0^2 is the square correlation coefficient between experimental and predicted values of the test set compounds with intercept set to zero. The value of $r_{m(test)}^2$ should be greater than 0.5 for an acceptable model. The concept of r_m^2 was not only applied to test set prediction, but it can as well be applied for training set if one considers the correlation between observed and leave-one out predicted values of the training set compounds [26]. Moreover, this can be used for the whole set considering Leave-one-out predicted values for the training set and predicted values of the test set compounds [23]. The $r_{m(overall)}^2$ statistic may be used for selection of the best predictive models from among comparable models. The values of k and k' , slopes of the regression line of the predicted activity versus actual activity and vice versa, were calculated using the following equations [25]:

$$k = \frac{\sum Y_i \bar{Y}_i}{\sum \bar{Y}_i^2} \text{ and } k' = \frac{\sum Y_i \bar{Y}_i}{\sum Y_i^2} \quad (7)$$

where Y_i and $\bar{Y}_i$ are the predicted and experimental activities, respectively.

Further statistical significance of the relationship between activity and the descriptors was checked by randomization test (Y-randomization) of the models. The Y column entries were scrambled and new QSAR models were developed using same set of variables as present in the un-randomized model. We have used a parameter, R_p^2 , [30] which penalizes the model R^2 for the difference between squared mean correlation coefficient (R_r^2) of randomized models and squared correlation coefficient (R^2) of the nonrandomized model. The R_p^2 parameter was calculated by the following equation:

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2} \quad (8)$$

This parameter, R_p^2 , ensures that the models so developed are not obtained by chance. We have assumed that the value of R_p^2 should be greater than 0.5 for an acceptable model.

Note that r_m^2 values do not take into account the number of predictor variables included in a model. When different models, having different number of predictor variables are compared then it may be very difficult to determine which one is the best model as r_m^2 does not consider the

number of predictor variables used. To solve this problem, another parameter $r_{m(overall)}^2(adjusted)$ may be calculated in a manner similar to the adjusted R^2 [26]:

$$r_{m(overall)}^2(adjusted) = \frac{(N-1) \times r_{m(overall)}^2 - P}{N-P-1} \quad (9)$$

Where N is the total number of compounds and P is the number of predictor variables.

To check the intercorrelation of descriptors, variance inflation factor (VIF) analysis was performed. The VIF value is calculated from:

$$VIF = \frac{1}{1-R^2} \quad (10)$$

Where R^2 is the multiple correlation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. If the VIF value is larger than 10, information of descriptors can be hidden by correlation of descriptors [31, 32].

3.0 RESULT AND DISCUSSION

The 150 active compounds with their biological activity were randomly divided into a training set of 120 compounds and a test set of 30 compounds. With the wide range of difference between the experimental values and the large diversity in the structures, the combined data set of 120 molecules and 30 molecules is ideal as a training and test set, as both sets do not suffer from bias due to the similarity of the structures. The various molecular descriptors (885 in total) as described in PaDEL-Descriptors version 2.18 [22] were calculated initially. By applying a missing value test, a zero test, a correlation test with a cutoff value of 0.0001, and a multicollinearity test with a cutoff value of 0.80, we have discarded the most likely parameters, resulting in 172 parameters. Further additional parameters were discarded by applying the GFA, and finally 8 parameters were selected for the development of the QSAR equation. As the squared correlation coefficient, R^2 , can be easily increased by the number of terms in the QSAR equation, we took the cross-validation correlation coefficient, Q_{cv}^2 , as the limiting factor for a number of descriptors to be used in the final model. It was observed that the Q_{cv}^2 value increased until the number of descriptors in the equation reached 8. So, the number of descriptors was restricted to 8 in the final QSAR model. The best significant relationship for the activity has been realized to be;

Model 1:

$$PIC50 = 3.08172(+/-0.52134) + 3.09393(+/-0.52028) ATSc3 + 28.05602(+/-3.2988) SCH - 3 - 34.91622(+/-3.30279) VCH - 7 - 13.37449(+/-1.35611) VC - 5 + 3.67308(+/-0.18491) VPC - 5 - 0.76154(+/-0.08721) nHBd - 2.48376(+/-0.49227) nddssS - 0.24588(+/-0.04857) minHBint5. \quad (11)$$

$$N = 120, LOF = 1.7850, R = 0.9658, R_{train}^2 = 0.9329, R_{adj}^2 = 0.9280, Q_{cv}^2 = 0.9065, F - test = 192.7484, SEE = 0.6557, Q = 1.4729, PRESS = 66.4307, SDEP = 0.7440, r_{test}^2 = 0.8929, r_0^2 = 0.8891, r_0'^2 = 0.8911$$

Model 2:

$$PIC50 = 3.35843(+/-0.50132) + 3.45359(+/-0.5738) ATSc3 + 26.16473(+/-3.20745) SCH - 3 - 32.26124(+/-3.02354) VCH - 7 - 13.26959(+/-1.37737) VC - 5 + 3.57781(+/-0.18586) VPC - 5 - 0.43222(+/-0.08688) nHBd - 0.18283(+/-0.03768) maxHBint5 + 0.56005(+/-0.10593) gmin. \quad (12)$$

$$N = 120, LOF = 1.8483, R = 0.9648, R_{train}^2 = 0.9305, R_{adj}^2 = 0.9255, Q_{cv}^2 = 0.9108, F - test = 185.6753, SEE = 0.6672, Q = 1.4460, PRESS = 63.3978, SDEP = 0.7269, r_{test}^2 = 0.9039, r_{test0}^2 = 0.9024, r_{test0}'^2 = 0.901$$

In the equations, the figures in the parentheses are the standard errors of the regression coefficients.

Table 2: Predicted values of the test set (external cross-validation) and results of statistical parameters

Parameters	Model 1	Model 2
$r_{m(LOO)}^2$	0.8899	0.8988
$r_{m(test)}^2$	0.8381	0.8686
$r_{m(overall)}^2$	0.8919	0.9033
$r_{m(overall)}^2(adjusted)$	0.8841	0.8963
$RMSEP$	0.6955	0.6534
R_{pred}^2	0.9261	0.9348
R_p^2	0.9008	0.8992
Q_{f2}^2	0.8800	0.8941
$ r_{test0}^2 - r_{test0}'^2 $	0.002	0.0014
$r_{test}^2 - r_{test0}'^2 / r_{test}^2$	0.002	0.0032
$r_{test}^2 - r_{test0}^2 / r_{test}^2$	0.0042	0.0017
k	1.033	1.0313
k'	0.9562	0.9592

where N is the number of compounds in the training set, R_{train}^2 is the squared correlation coefficient, SEE is the estimated standard deviation about the regression line, R_{adj}^2 is the square of the adjusted correlation coefficient for degrees of freedom, F test is the measure of variance that compares two models to see if the more complex model is more reliable than the less complex one (the model is supposed to be good if the F test is above a threshold value), and Q_{cv}^2 is the square of the correlation coefficient of the cross-validation using the leave-one-out cross-validation technique. The QSAR model developed in this study was statistically ($R_{train}^2 = 0.9329$, $Q_{cv}^2 = 0.9065$, F test = 192.7484) best fitted and consequently was used for prediction of activities (pIC_{50}) of training and test sets of molecules. The relationships between predicted (both training and test) activities and the corresponding experimental activities are shown in **Figures 7** and **8**. The R_{train}^2 and Q_{cv}^2 values of 0.9329 and 0.9065, respectively, of the model corroborate with the criteria for a QSAR model to be highly predictive [25]. The difference between R_{train}^2 and Q_{cv}^2 never exceed 0.3. A large difference suggests the following: presence of outliers, over-fitted model, and presence of irrelevant variables in data [29] as such $R_{train}^2 - Q_{cv}^2 = 0.0264$ which is less than 0.3. The standard error of estimate for the model was 0.6557, which is an indicator of the robustness of the fit and suggested that the predicted pIC_{50} based on model is reliable. The developed model was further validated by a randomization technique ($R_{yrand}^2 = 0.0658$ and $Q_{yrand}^2 = -0.1002$, no chance correlation) [33]. The values of R_r^2 and R^2 were determined, which were then used for calculating the value of R_p^2 . Models with R_p^2 values greater than 0.5 are considered statistically robust. If the value of R_p^2 is less than 0.5, then it may be concluded that the outcome of the model is merely by chance, and it is not at all well predictive for truly external data sets. In this data set, values of R_p^2 for all the 100 models were well above the stipulated value of 0.5 (Table 2). Therefore, it can be concluded that besides being robust, the model developed is well predictive.

Table 3: Specification of entered descriptors in GFA method

Descriptors	*t Stat	**P-value	***VIF	****MF
ATSc3	5.911163	3.81E-08	2.852	0.4582
SCH-3	5.946609	3.23E-08	2.0917	0.0658
VCH-7	8.50492	9.55E-14	4.4264	-3.9335
VC-5	-10.5717	1.72E-18	1.9016	-2.0945
VPC-5	-9.86242	7.44E-17	2.6669	8.1554
nHBd	19.86406	2.43E-38	2.2706	-1.3983
nddssS	-8.7327	2.9E-14	1.1085	-0.0404
minHBint5	-5.04555	1.78E-06	1.8683	-0.2127

*t-stat was introduced to compare under the confidence level 95%, **p-value was introduced to compare under the confidence level 95%, ***Variation inflation factor, ****Mean effect

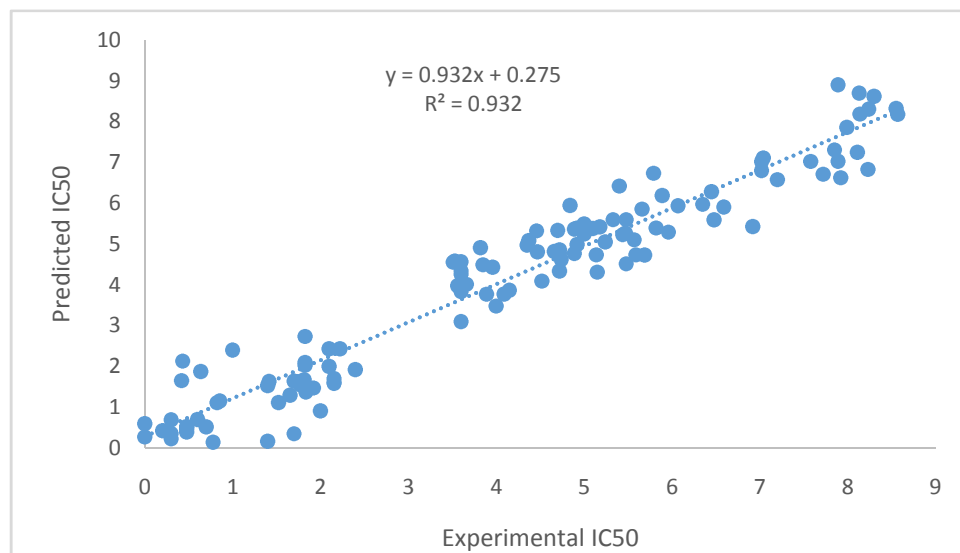


Figure 7: The calculated PIC50 versus the experimental PIC50 for training set

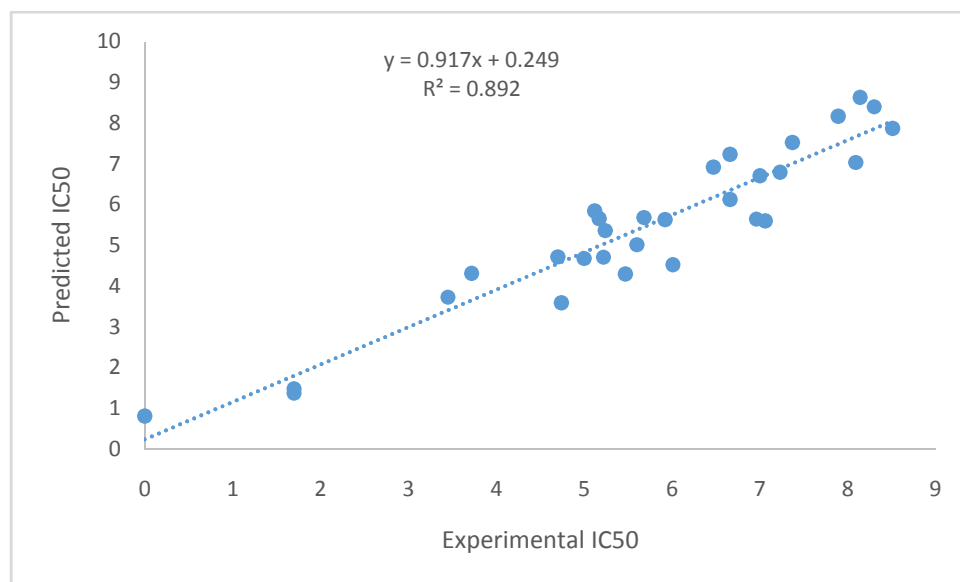


Figure 8: The calculated PIC50 versus the experimental PIC50 for the test set

The inter-correlation of the descriptors used in the QSAR model was very low (below 0.8), which is in conformity to the study that, for a statistically significant model, it is necessary that the descriptors involved in the equation should not be inter-correlated with each other [34]. To further check the intercorrelation of descriptors, VIF analysis was performed. In this model, the VIF values of these descriptors are (Table 3 and 4) 2.852 (ATSc3), 2.0917(SCH-3), 4.4264 (VCH-7), 1.9016 (VC-5), 2.6669 (VPC-5) 2.2706 (nHBd) 1.1085 (noddssS) and 1.8683 (minHBint5) (Table 3), which are less than the threshold value of 10 [31, 32]. Satisfied with the robustness of the QSAR model developed using the training set, we have applied the QSAR model to an external data set constituting the test set. As the experimental values of IC₅₀ for these inhibitors are already available, this set of molecules provides an excellent data set for testing the

prediction power of the QSAR model for new compounds. Table 1 represents the predicted pIC₅₀ values of the test set based on model (1). The overall root mean square error (RMSE) between the experimental and predicted pIC₅₀ values was 0.6955, which reveals good predictability. The estimated correlation coefficients between experimental and predicted pIC₅₀ values with intercept (r^2_{test0}) and without intercept (r^2_{test}) were 0.8891 and 0.8929, respectively. The value of $[(r^2_{test} - r^2_{test0})/r^2_{test}] = (0.8929 - 0.8891)/0.8929 = 0.0042$, which is less than 0.1 stipulated value [25] and therefore validates the usefulness of the QSAR model for predicting the biological activity of the external data set. Also, the values of k and k' were 1.033 and 0.9562, which are well within the specified ranges of 0.85 and 1.15 [25]. The values of $r^2_{m(LOO)} = 0.8899$, $R^2_{pred} = 0.9261$, $r^2_{m(test)} = 0.8381$, and $r^2_{m(overall)} = 0.8919$ were found to be in the acceptable range [26], thereby indicating the good external predictability of the QSAR model.

Table 4: Physical-Chemical meanings of the descriptors used in the developed mt-QSAR model

Descriptor	Definition	Symbol
Autocorrelation Descriptor Charge	ATS autocorrelation descriptor, weighted by charges	ATSc3
ChiChain Descriptor	Simple chain, order 3	SCH3
	Valence chain, order 3	VCH-7
ChiCluster Descriptor	Valence cluster, order 5	VC-5
ChiPathCluster Descriptor	Valence path cluster, order 5	VPC-5
Electrotopological State Atom Type Descriptor	Count of E-States for (strong) Hydrogen Bond donors	nHBd
	Count of atom-type E-State: >S==	nddssS
	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 5	MinHBint5

The Williams plot, the plot of the standardized residuals versus the leverage, was exploited to picture the applicability domain (AD) [35, 36]. Leverage indicates a compound's distance from the centroid of X. The leverage of a compound in the original variable space is defined as:

$$h_i = X_i^T (X^T X)^{-1} X_i \quad (13)$$

where X_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) is defined as:

$$h^* = \frac{3(p+1)}{N} \quad (14)$$

Where N is the number of training compounds, p is the number of predictor variables. From the Williams plot (Fig. 9), it is obvious that one compounds in the test set fall inside the domain (No. 63) of the GFA-MLR model (the warning leverage limit is 0.225). There are only four compounds (No. 94, 128, 130 and No. 134 in the training set) which have the leverage higher than the warning h^* value, thus they can be regarded as structural outliers. Fortunately, in this case the data predicted by the model are good for compound numbers 7 and 24, thus they are “good leverage” compounds. For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units ($\pm 3\delta$) except compound number 105. Consequently compound 105 can be as outlier. Because this compound is one of the test set compounds, there is no need to remove this compound from the data set.

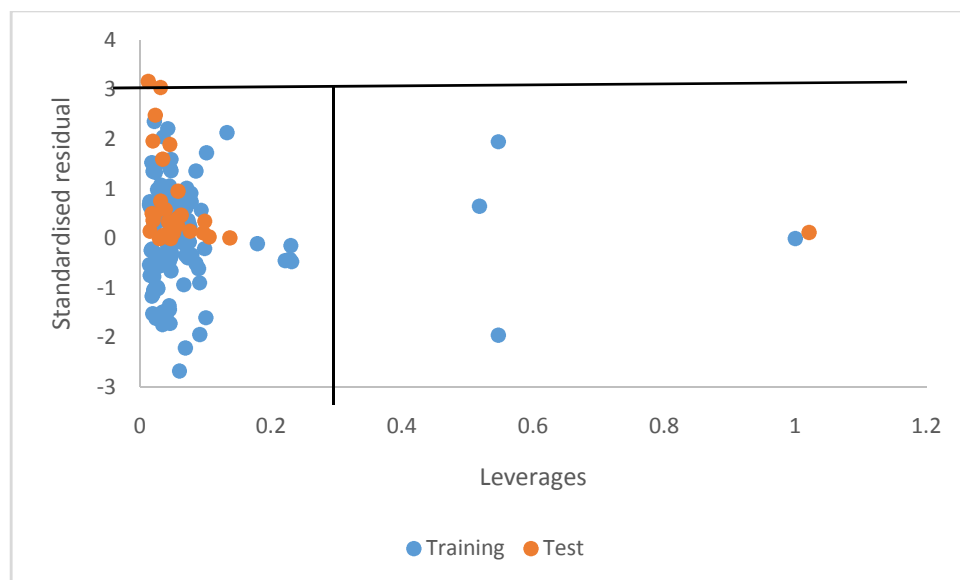


Figure 9: The William plot, the plot for the standardized residuals versus the leverage value

Euclidean based applicability domain, It is based on distance scores calculated by the Euclidean distance norms. First and foremost, normalized mean distance score for training set compounds are calculated and these values ranges from 0 to 1(0=least diverse, 1=most diverse training set compound). Then normalized mean distance score for test set are calculated, and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain. This can also be checked by plotting a ‘Scatter plot’ (normalized mean distance vs. respective activity) including both training and test set. If the test set compounds are inside the domain/area covered by training set compounds that means these compounds are inside the applicability domain otherwise not. From the plot (Fig. 10), all the test set compound are inside the domain of the training set [37].

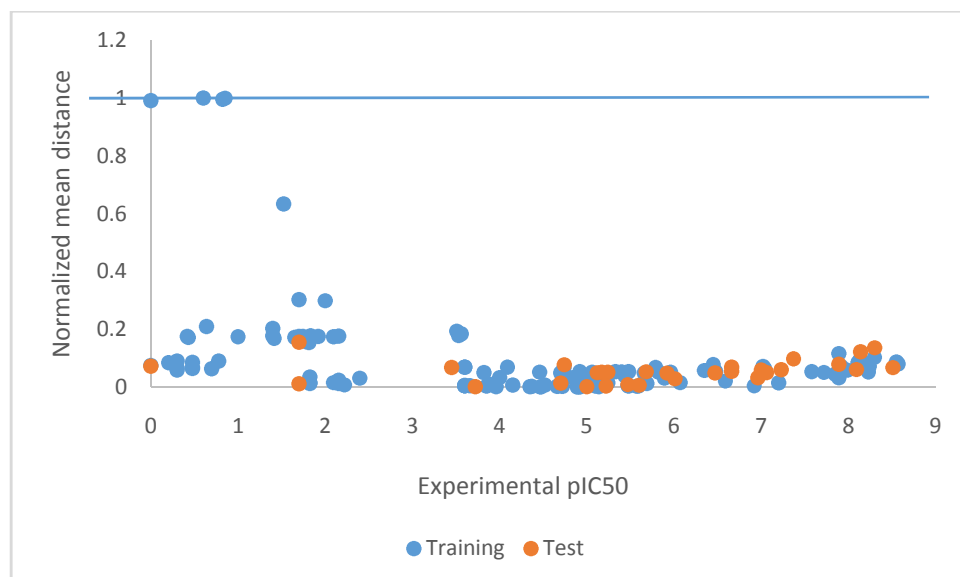


Figure 10: Euclidean distance norms, the plot for normalized mean distance versus experimental pIC50

To examine the relative importance, as well as the contribution of each descriptor in the model, the value of the mean effect (MF) [37, 38] was calculated for each descriptor. This calculation was performed using the following equation.

$$MF_j = \frac{\beta_j \sum_{i=1}^n d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \quad (15)$$

Where MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule and eventually, m is the descriptors number for the model. The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign (+, -) indicates the variation direction in the values of the activities as a result of the increase or decrease in the descriptor values. The mean effect values are shown in Table 3. All descriptors were calculated for the sorts. The activity is assumed to be highly dependent upon the ATSc3, SCH3 and VPC-5. In the model, a student's t-test was performed at a confidence level of 95% to confirm the significance of each descriptor. All the p-values (fig. 3) of the descriptors were less than 0.05, indicating that the selected descriptors were statistically significant at the 95% level.

4.0 Conclusion

In this article, a QSAR study of 150 molecules showing HIV-1 inhibitor activity was performed based on the theoretical molecular descriptors calculated by the PaDEL-Descriptors software. The built model was assessed comprehensively (internal and external validation) and all the validations indicated that the QSAR model built was robust and satisfactory and that the selected descriptors could account for the structural features responsible for the HIV-1 inhibitors. The QSAR model developed in this study can provide a useful tool to predict the activity of new compounds and also to design new compounds with high anti-HIV-1 inhibitor activity.

References:

- [1] Y. Mehellou, E. De Clercq, "Twenty-six years of anti- HIV drug discovery: where do we stand and where do we go?" *Journal of Medicinal Chemistry*, vol. 53, pp.521-5, 2010.
- [2] R.J. Pomerantz and D.L. Horn, "Twenty Years of Therapy for HIV-1 Infection. *Nature medicine*, vol. 9, issue 7, pp867-873, 2003.
- [3] N.B. Dyatkina, C.D. Roberts, J.D. Keicher, Y. Dai and J.P. Nadherny *et al.*, "Minor groove DNA binders as antimicrobial agents. 1. Pyrrole tetraamides are potent antibacterials against vancomycin resistant *Enterococci* and methicillin resistant *Staphylococcus aureus*," *Journal Medicinal Chemistry*, 45: 805-817, 2002.
- [4] C. Loveday, "Nucleoside reverse transcriptase inhibitor resistance," *Journal of Acquire Immune Deficiency Syndrome*. 26 Suppl1; S25-S33. 2001.
- [5] J.J. Deeks, "Systematic reviews of evaluations of diagnostic and screening tests," *British Medical Journal.*; 323(7311) pp. 487. 2001.
- [6] P.L. Volberding, "Introduction to Acquire Immune Deficiency Syndrome," *Journal of international perspectives on antiretroviral resistance*, 26 (Suppl1); S1-S2. 2001.
- [7] L.A. Cooley, S.R. Lewin, "HIV-1 cell entry and advances in viral entry inhibitor therapy." *Journal of Clinical Virology*. 26(2), pp. 121-32, 2003.
- [8] A. Koutsoukas, B. Simms, J. Kirchmair, P.J. Bond, A.V. Whitmore, S. Zimmer, M.P. Young, J.L. Jenkins, M. Glick, R.C. Glen, A. Bender, "From in silico target prediction to multi-target drug design: current databases, methods and applications," *Journal of Proteomics*, 74(12), pp. 2554–2574, 2011.
- [9] X.H. Ma, Z. Shi, C. Tan, Y. Jiang, M.L. Go, B.C. Low, Y.Z. Chen, "In-Silico approaches to multi-target drug discovery," *Pharmaceutical Research*, 27(5), pp.739–749, 2010.
- [10] S. Ajmani, S.A. Kulkarni, "Application of QQSAR for scaffold hopping and lead optimization in multitarget inhibitors," *Molecular Informatics*, 31(6–7), pp. 473–490, 2012.
- [11] P. Labute, "A widely applicable set of descriptors," *Journal of Molecular Graphic and Modelling*, 18(4-5): pp. 464-77, 2000.
- [12] J. Xu, J. Stevenson, "Drug-like index: A new approach to measure drug-like compounds and their diversity," *Journal Chemical Information and Computer Science*, vol. 40, pp. 1177-87, 2000.
- [13] F.J. Prado-Prado, H. González-Díaz, O.M. de la Vega, F.M. Ubeira, K.C. Chou, "Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds," *Bioorganic and Medicinal Chemistry*, vol. 16(11):5871-80, 2008.

- [14] R. Todeschini, V. Consonni, (2000) Handbook of Molecular Descriptors. Wiley-VCH, Weinheim.
- [15] H. González-Díaz and F.J. Prado-Prado, “Unified QSAR and network-based computational chemistry approach to antimicrobials, part 1: Multispecies activity models for antifungals,” *Journal of Computational Chemistry*, vol. 29, Issue 4, pp. 656–667, 2008.
- [16] Enrique Molina , Humberto González Díaz , Maykel Pérez González , Elismary Rodríguez , and Eugenio Uriarte Designing Antibacterial Compounds through a Topological Substructural Approach. *Journal of Chemical Information and Computer Science*, vol. 44 (2), pp. 515–521, 2004
- [17] E.I. Edache, A. Uzairu, S.E. Abechi, “Multivariate QSAR Study of Indole β - Diketo Acid, Diketo Acid and Carboxamide Derivatives as Potent Anti-HIV Agents,” *International Journal of Innovative Research and Development*, vol. 4, 374-390, 2015.
- [18] J.M. Luco and F.H. Ferretti, “QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives,” *Journal of Chemical Information and Computer Sciences*, vol. 37, pp. 392-401, 1997.
- [19] Wavefunction, 2013, Inc. Spartan’14, version 1.1.2, Irvine, California, USA
- [20] A.D. Becke, “A new mixing of Hartree-Fock and local density-functional theories,” *Journal of Chemical Physics*, vol. 98, pp. 1372–1377, 1993.
- [21] G.A. Petersson, A. Bennett, T.G. Tensfeldt, M.A Al-Laham, W.A. Shirley, J. Mantzaris, “A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements,” *Journal of Chemical Physics*, vol. 89, pp. 2193–2218. 1988.
- [22] C.W. Yap, “PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints,” *Journal of Computational Chemistry*, vol. 32(7). Pp.1466-1474. 2011.
- [23] D. Rogers, A.J. Hopfinger. “Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships,” *Journal of Chemical Information and Computer Science*, 34, 854-866, 1994.
- [24] C. Hansch, J. Taylor, P. Sammes, Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds, Pergamon, New York, 6:1, 1990.
- [25] A. Golbraikh, A. Tropsha, “Beware of q²!” *Journal Molecular Graphic and Modelling*, vol. 20, pp. 269-276, 2002.
- [26] P.P. Roy, K. Roy, “On some aspects of variable selection for partial least squares regression models,” *QSAR and Combinatorial Science*, vol. 27, pp. 302-313, 2008.

- [27] L. Pogliani, "Structure-property relationships of amino acids and some dipeptides". *Journal of Amino Acids*, vol. 6, pp.141-153, 1994.
- [28] L. Pogliani, "Modeling with special descriptors derived from a medium-sized set of connectivity indices," *Journal of physical chemistry*, vol. 100, pp. 18065-18077, 1996.
- [29] R.P. Verma, C. Hansch, "QSAR modeling of taxane analogues against colon cancer," *European Journal of Medicinal Chemistry*, vol.45. pp. 1470-1477, 2010.
- [30] K. Roy, S. Paul, "Exploring 2D and 3D QSARs of 2,4-diphenyl-1,3-oxazolines for ovicidal activity against *Tetranychus urticae*," *QSAR and Combinatorial Science*, vol. 28, pp. 406-425, 2008
- [31] M. Jaiswal, P.V. Khadikar, A. Scozzafava, C.T. Supuran, "Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides." *Bioorganic and Medicinal Chemistry Letter*, vol. 14, pp. 3283-3290, 2004.
- [32] S. Shapiro, B. Guggenheim, "Inhibition of oral bacteria by phenolic compounds: Part 1. QSAR analysis using molecular connectivity," *Quantitative Structure Activity Relationship*, vol. 17, pp. 327-337, 1998.
- [33] R. Kiralj and M.M.C. Ferreira, "Basic validation procedures for Regression Models in QSAR and QSPR Studies: Theory and Application," *Journal of Brazilian Chemical Society*, vol. 20(4), pp. 770-787, 2009.
- [34] S. Deswal, N. Roy, "Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors," *European Journal of Medicinal Chemistry*, vol. 41: pp. 1339-1346, 2006.
- [35] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. Van De Sandt, W. Tong, G. Veith, C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships," *Alternative to Laboratory Animal*, vol. 33, 155-173, 2005.
- [36] OECD, 2007. Guidance document on the validation of (Quantitative)Structure– Activity Relationships[(Q)SARs] models. Available online at [/http://applied.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono\(2007\)2S](http://applied.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)2S).
- [37] N. Minovski, S. Zuperl, V. Drgan, M. Novi, "Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study," *Analytica Chimica Acta*, vol. 759, pp. 28– 42, 2013
- [38] D.C. Hoaglin and R.E. Welsch. The Hat Matrix in Regression and ANOVA. *The American Statistician*, vol. 32, No. 1, pp. 17-22, 1978.