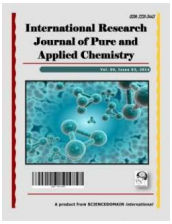

International Research Journal of Pure & Applied Chemistry
11(1): 1-15, 2016, Article no.IRJPAC.22863
ISSN: 2231-3443, NLM ID: 101647669

SCIENCEDOMAIN *international*
www.sciencedomain.org

In-silico Discovery and Simulated Selection of Multi-target Anti-HIV-1 Inhibitors

Emmanuel Israel Edache[1*], Hambali Umar Hambali[2], David Ebuka Arthur[1], Adedirin Oluwaseye[3] and Onoyima Christian Chinweuba[4]

[1]Department of Chemistry, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.
[2]Department of Chemical Engineering, Ahmadu Bello University, Zaria, Kaduna State, Nigeria.
[3]Chemistry Advance Laboratory, Sheda Science and Technology Complex (SHESTCO), P.M.B. 186, Garki, Abuja, Federal Capital Territory, Nigeria.
[4]Department of Chemistry, Nigeria Police Academy, Wudil, Kano State, Nigeria.

Authors' contributions

This work was carried out in collaboration between all authors. Authors EIE, DEA and AO designed the study and wrote the protocol. Authors EIE, DEA and AO preformed the statistical analysis, managed the literature search and wrote the first draft of the manuscript with assistance from authors HUH and OCC. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/IRJPAC/2016/22863
Editor(s):
(1) Hao-Yang Wang, Department of Analytical, Shanghai Institute of Organic Chemistry, Shanghai Mass Spectrometry Center, China.
Reviewers:
(1) Hazem Mohammed Ebraheem Shaheen, Damanhour University, Damanhour, Egypt.
(2) Gyula Oros, Plant Protection Institute of the Centre for Agricultural Research of the Hungarian Academy of Sciences, Budapest, Hungary.
(3) Rajeev Singh, University of Delhi, New Delhi, India.
Complete Peer review History: http://sciencedomain.org/review-history/12951

Original Research Article

Received 2nd November 2015
Accepted 1st December 2015
Published 12th January 2016

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

**Corresponding author: E-mail: inalegwu334real@yahoo.com;*

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 demonstrated that the predictive capacity of the model was attractive, and it can be utilized for outlining comparable gathering of anti-HIV compounds.

Keywords: Multi-target; QSAR; HIV-1 inhibitors; GFA; DFT; applicability domain.

1. 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 resistance 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 have 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 [11,12] 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. 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 lm (micromoles) were converted into negative logarithmic dose in moles (pIC50) for mt-QSAR Analysis (Table 1).

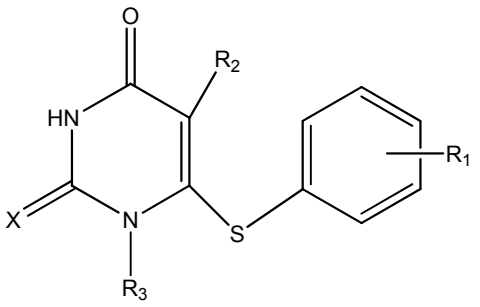

Fig. 1. Compound 1-106

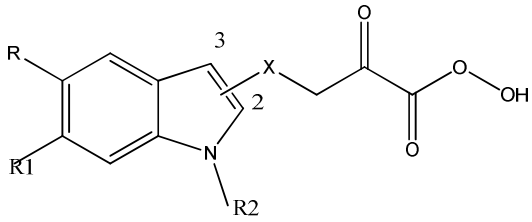

Fig. 2. Compound 107-117

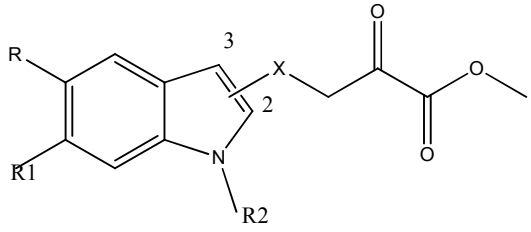

Fig. 3. Compound 118-122

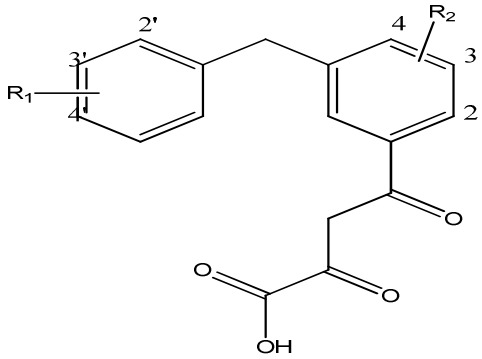

Fig. 4. Compound 123-126

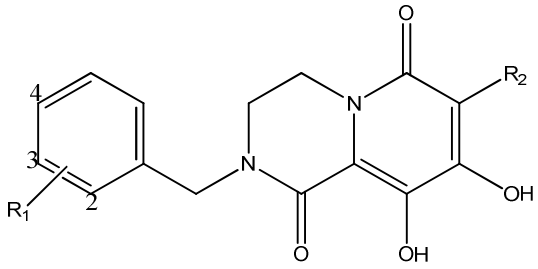

Fig. 5. Compound 127-135

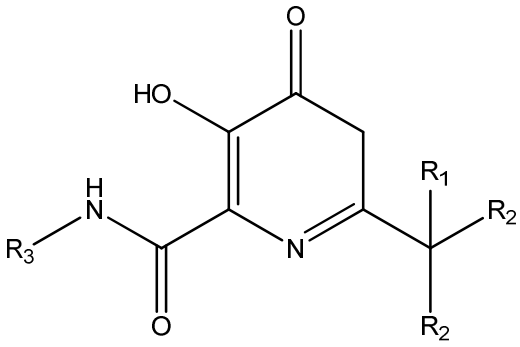

Fig. 6. Compound 136-150

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.

Table 1. Biological activities of the training and test set

No	R1	R2	R3	X	PIC50	Predicted PIC50	Residuals
1*	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.000	4.6797	0.3203
2	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.140	4.7319	0.4081
3	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.100	5.3778	-0.2778
4	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.890	6.1860	-0.2960
5	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.590	5.9066	0.6834
6	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.660	4.8120	-0.1520
7	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.090	3.7714	0.3186
8	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.470	4.8074	-0.3374
9	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.000	5.2473	-0.2473
10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.240	5.0500	0.1900
11	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.890	4.7662	0.1238
12	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.480	4.5106	0.9694
13	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.350	4.9716	-0.6216
14	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.920	4.9853	-0.0653
15	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.570	5.0988	0.4712
16	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.590	4.7311	0.8589
17	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.720	4.8542	-0.1342
18	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.850	4.4880	-0.6380
19	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.150	3.8658	0.2842
20	H	Et	$CH_2OCH_2CH_2OH$	O	6.920	5.4241	1.4959
21	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.200	6.5751	0.6249
22	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.890	7.0244	0.8656
23	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.570	8.1755	0.3945
24	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.850	7.3039	0.5461
25	H	Me	$CH_2OCH_2CH_2OH$	O	5.150	4.3064	0.8436
26	H	I	$CH_2OCH_2CH_2OH$	O	5.440	5.2250	0.2150
27	H	$CH=CPH_2$	$CH_2OCH_2CH_2OH$	O	6.070	5.9337	0.1363
28	4-F	Me	$CH_2OCH_2CH_2OH$	O	3.600	4.2571	-0.6571
29	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.600	3.9485	-0.3485
30	4-OH	Me	$CH_2OCH_2CH_2OH$	O	3.560	3.9728	-0.4128
31	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	3.510	4.5501	-1.0401
32	H	COOMe	$CH_2OCH_2CH_2OH$	O	5.180	5.4159	-0.2359
33	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4.740	4.6154	0.1246
34	H	SPh	$CH_2OCH_2CH_2OH$	O	4.840	5.9415	-1.1015
35*	H	CCH	$CH_2OCH_2CH_2OH$	O	4.740	3.5914	1.1486
36*	H	CCPh	$CH_2OCH_2CH_2OH$	O	5.470	4.3007	1.1693
37	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	4.920	5.3874	-0.4674
38	H	COPh	$CH_2OCH_2CH_2OH$	O	4.890	5.3635	-0.4735
39	H	CCMe	$CH_2OCH_2CH_2OH$	O	4.720	4.3327	0.3873
40	H	F	$CH_2OCH_2CH_2OH$	O	4.000	3.4761	0.5239
41	H	Cl	$CH_2OCH_2CH_2OH$	O	4.520	4.0879	0.4321
42*	H	Br	$CH_2OCH_2CH_2OH$	O	4.700	4.7191	-0.0191
43	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.890	3.7650	0.1250
44	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	3.530	4.5804	-1.0504
45*	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.720	4.3142	-0.5942
46	4-CN	Me	$CH_2OCH_2CH_2OH$	O	3.600	4.3391	-0.7391
47	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	3.600	4.5607	-0.9607
48	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	3.960	4.4301	-0.4701
49*	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	3.450	3.7322	-0.2822

No	R1	R2	R3	X	PIC50	Predicted PIC50	Residuals
50	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	3.600	3.8286	-0.2286
51	H	Pr	$CH_2OCH_2CH_2OH$	O	5.470	5.2652	0.2048
52	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.660	4.0097	-0.3497
53	H	$CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.690	4.7291	0.9609
54*	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.220	4.7125	0.5075
55	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.370	5.0784	-0.7084
56*	H	Me	$CH_2OCH_2CH_2OAc$	O	5.170	5.6595	-0.4895
57	H	Et	CH_2OCH_2Me	O	7.720	6.7095	1.0105
58	H	Et	CH_2CH_2Ph	O	8.230	6.8269	1.4031
59	3,5-Cl_2	Et	CH_2CH_2Me	O	8.130	8.6967	-0.5667
60	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$	O	4.460	5.320	-0.8600
61	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	4.700	5.3307	-0.6307
62	H	Me	H	O	3.600	3.0964	0.5036
63	H	Me	Me	O	3.820	4.9059	-1.0859
64*	H	c-Pr	CH_2OCH_2Me	O	7.000	6.7082	0.2918
65*	H	Et	CH_2O-i-Pr	O	6.470	6.9214	-0.4514
66	H	Et	CH_2O-c-Hex	O	5.400	6.4185	-1.0185
67	H	Et	CH_2OCH_2-c-Hex	O	6.350	5.9712	0.3788
68	H	Et	$CH_2OCH_2CH_2Ph$	O	7.020	6.7969	0.2231
69*	H	Me	CH_2OMe	O	5.680	5.6838	-0.0038
70	H	Me	CH_2OBu	O	5.330	5.5944	-0.2644
71	H	Me	Et	O	5.660	5.8524	-0.1924
72*	H	Me	Bu	O	5.920	5.6360	0.2840
73	H	i-Pr	CH_2OCH_2Me	O	7.990	7.8584	0.1316
74*	H	i-Pr	CH_2OCH_2Ph	O	8.510	7.8725	0.6375
75	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.550	8.3215	0.2285
76	3,5-Me_2	Et	CH_2OCH_2Me	O	8.240	8.3075	-0.0675
77	H	Me	$CH_2OCH_2CH_2OMe$	O	5.060	5.37286	-0.3129
78*	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.120	5.8451	-0.7251
79	H	Me	CH_2OCH_2Me	O	6.480	5.5904	0.8896
80	H	Me	$CH_2OCH_2CH_2Cl$	O	5.820	5.3906	0.4294
81*	H	Me	$CH_2OCH_2CH_2N_3$	O	5.240	5.3620	-0.1220
82	H	Me	$CH_2OCH_2CH_2F$	O	5.960	5.2894	0.6706
83	H	Me	$CH_2OCH_2CH_2Me$	O	5.480	5.5944	-0.1144
84*	H	Me	CH_2OCH_2Ph	O	7.060	5.6046	1.4554
85	H	Et	CH_2OCH_2Me	S	7.580	7.0212	0.5588
86*	H	i-Pr	CH_2OCH_2Me	S	7.890	8.1703	-0.2803
87	H	i-Pr	CH_2OCH_2Ph	S	8.140	8.1835	-0.0435
88	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.890	8.9027	-1.0127
89*	H	Et	CH_2O-i-Pr	S	6.660	7.2336	-0.5736
90	H	Et	CH_2O-c-Hex	S	5.790	6.7307	-0.9407
91	H	Et	CH_2OCH_2-c-Hex	S	6.450	6.2829	0.1671
92	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.920	6.6230	1.2970
93	H	Et	$CH_2OCH_2CH_2Ph$	S	7.040	7.1085	-0.0686
94	H	c-Pr	CH_2OCH_2Me	S	7.020	7.02	-7.1E-15
95*	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.660	6.1276	0.5324
96	H	Pr	$CH_2OCH_2CH_2OH$	S	5.000	5.4871	-0.4871
97*	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.300	8.4001	-0.1001
98*	H	Et	CH_2OCH_2Ph	S	8.090	7.0345	1.0555
99*	3,5-Me	Et	CH_2OCH_2Ph	S	8.140	8.6326	-0.4926
100	3,5-Me_2	Et	CH_2OCH_2Me	S	8.300	8.6192	-0.3192
101*	H	Et	$CH_2OCH_2CH_2OH$	S	6.960	5.6447	1.3153
102*	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.230	6.7974	0.4326
103	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.110	7.2472	0.8628
104*	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.370	7.5269	-0.1569
105*	H	Me	$CH_2OCH_2CH_2OH$	S	6.010	4.5251	1.4849
106*	H	$CH_2CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.600	5.0171	0.5829

No.	R	R1	R2	X	PIC50	Predicted PIC50	Residuals
107	H	H	CH_3	2-CO	0.7780	0.1358	0.6422
108		OCH_2O	CH_3	2-CO	0.3010	0.2200	0.2200
109	H	H	CH_2CH_3	2-CO	0.2040	0.4190	-0.2150
110		OCH_2O	CH_2CH_3	2-CO	0.6990	0.5137	0.5137
111	H	H	Bn	2-CO	0.0000	0.5915	-0.5915
112		OCH_2O	Bn	2-CO	0.3010	0.6869	0.6869
113	H	H	CH_3	3-CO	0.3010	0.3503	-0.0493
114		OCH_2O	CH_3	3-CO	0.4770	0.3874	0.3874
115	H	H	CH_2CH_3	3-CO	0.4770	0.4725	0.0045
116		OCH_2O	CH_2CH_3	3-CO	0.4770	0.4725	0.5209
117*	H	H	Bn	3-CO	0.0000	0.8146	-0.8146

No.	R	R1	R2	X	PIC50	Predicted PIC50	Residuals
118	H	H	CH_3	2-CO	1.6530	1.2915	0.3615
119*		OCH_2O	CH_3	2-CO	1.6990	1.3757	0.3233
120		OCH_2O	CH_2CH_3	2-CO	1.8130	1.6693	0.1437
121		OCH_2O	CH_3	3-CO	1.7780	1.5432	0.2348
122	H	H	CH_2CH_3	3-CO	1.8416	1.6283	-0.2133

No.	R1	R2	PIC50	Predicted PIC50	Residuals
123	4'-Cl	-	0.000	0.2665	-0.2665
124	3'-F	-	0.602	0.6929	-0.0909
125	-	4-OCH_3	0.824	1.1063	-0.2823
126	-	3-OCH_3	0.854	1.1517	-0.2977

No.	R1	R2	PIC50	Predicted PIC50	Residuals
127	4-F	-	1.000	2.3987	-1.3987
128	H	-	0.638	1.8732	-1.2352
129	2-Cl	-	0.432	2.1258	-1.6938
130	3-Cl	-	1.398	0.1628	1.2352
131	4-Cl	-	0.420	1.6481	-1.2281
132	4-F, 3-Cl	-	1.398	1.5282	-0.1302
133	4-F	CN	1.699	0.3459	1.3531
134	4-F	Br	1.523	1.1098	0.4132
135	4-F	I	1.699	1.6384	0.0606

No.	R1	R2	R3	PIC50	Predicted PIC50	Residuals
136	$NHCOCH_3$	CH_3	4-fluorotoluene	2.1555	1.5778	0.5772
137	NH-SO_2-CH_3	CH_3	4-fluorotoluene	2.097	1.9982	0.0988
138	NHCO-$N(CH3)_2$	CH_3	4-fluorotoluene	1.745	1.6208	0.1242
139	$NHSO2-N(CH3)_2$	CH_3	4-fluorotoluene	1.921	1.4660	0.4550
140	NHCOCO-$N(CH3)_2$	CH_3	4-fluorotoluene	2.000	0.9059	1.0941
141	NHCOCO-OCH_3	CH_3	4-fluorotoluene	1.824	1.3640	0.4700
142	NHCOCO-OH	CH_3	4-fluorotoluene	2.398	1.9205	0.4775
143	N(CH3)COCO-$N(CH3)_2$	CH_3	4-fluorotoluene	1.824	2.0287	-0.2047
144*	NHCO-pyridine	CH_3	4-fluorotoluene	1.699	1.4853	0.2137
145	NHCO-pyridazine	CH_3	4-fluorotoluene	1.824	2.0946	-0.2706
146	NHCO-pyrimidine	CH_3	4-fluorotoluene	2.155	1.6141	0.5409
147	NHCO-oxazole	CH_3	4-fluorotoluene	2.155	1.6986	0.4564
148	NHCO-thiazole	CH_3	4-fluorotoluene	2.097	2.4271	-0.3301
149	NHCO-1H imidazole	CH_3	4-fluorotoluene	2.222	2.4303	-0.2083
150	NHCO-1,3,4-oxadiazole	CH_3	4-fluorotoluene	1.824	2.7315	-0.9076

***indicates the compounds considered in the test set*

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:

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

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^2_{adj}), were R^2_{adj} is defined by;

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

R^2_{adj} 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^2_{adj} value diminishes if an added variable to the equation does not reduce the unexplained variance [24]. Subsequently, R^2_{adj} 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^2_{test} , was calculated using the following equation [26]:

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

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^2m by the following equation [26]:

$$r^2_m = r^2 \times \left(1 - \sqrt{r^2 - r^2_0}\right) \quad (6)$$

Where r^2_0 is the square correlation coefficient between experimental and predicted values of the test set compounds with intercept set to zero. The value of $r^2_{m(test)}$ should be greater than 0.5 for an acceptable model. The concept of r^2_m 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 $r^2_{m(overall)}$ 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]:

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

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^2_p, [30] which penalizes the model R^2 for the difference between squared mean correlation coefficient (R^2_r) of randomized models and squared correlation coefficient (R^2) of the nonrandomized model. The R^2_p parameter was calculated by the following equation:

$$R^2_p = R^2 \times \sqrt{R^2 - R^2_r} \quad (8)$$

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

Note that r^2_mvalues 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^2_m does not consider the number of predictor variables used. To solve this problem, another parameter $r^2_{m(overall)}(adjusted)$may be calculated in a manner similar to the adjusted R^2 [26]:

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

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. RESULTS 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:

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

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

Model 2:

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

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

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

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 = 192.75) 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 Figs. 7 and 8. The R^2_{train} and Q^2_{cv} 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^2_{train} and Q^2_{cv} 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 $R^2_{train} - Q^2_{cv} = 0.0264$ 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 ($R^2_{yrand} = 0.0658$ and $Q^2_{yrand} = -0.1002$, no chance correlation) [33]. The values of R^2_r and R^2 were determined, which were then used for calculating the value of R^2_p. Models with R^2_p values greater than 0.5 are considered statistically robust. If the value of R^2_p 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^2_p 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.

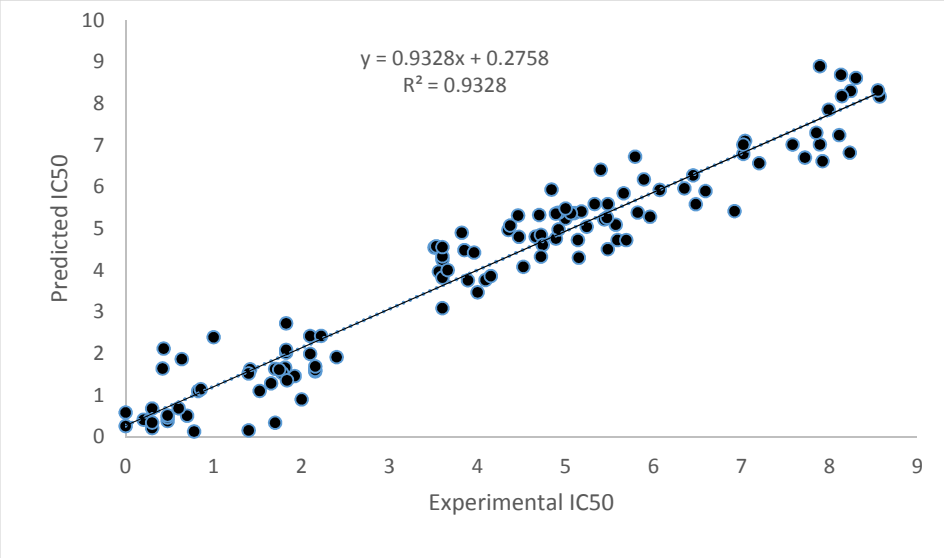

Fig. 7. The calculated PIC50 versus the experimental PIC50 for training set

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 (Tables 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 (nddssS) 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 IC_{50} 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 pIC_{50} values was 0.6955, which reveals good predictability. The estimated correlation coefficients between experimental and predicted pIC_{50} values with intercept (r^2_{testo}) 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.

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 Xi 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 (±3δ) 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.

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

Parameters	Model 1	Model 2
$r^2_{m(LOO)}$	0.8899	0.8988
$r^2_{m(test)}$	0.8381	0.8686
$r^2_{m(overall)}$	0.8919	0.9033
$r^2_{m(overall)}(adjusted)$	0.8841	0.8963
$RMSEP$	0.6955	0.6534
R^2_{pred}	0.9261	0.9348
R^2_p	0.9008	0.8992
Q^2_{f2}	0.8800	0.8941
$\lvert r^2_{testo} - r'^2_{testo} \rvert$	0.002	0.0014
$r^2_{test} - r'^2_{test0} / r^2_{test}$	0.002	0.0032
$r^2_{test} - r^2_{test0} / r^2_{test}$	0.0042	0.0017
k	1.033	1.0313
k'	0.9562	0.9592

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].

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*

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 Valence chain, order 3	SCH3 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 Count of atom-type E-State: >S== Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 5	nHBd nddssS MinHBint5

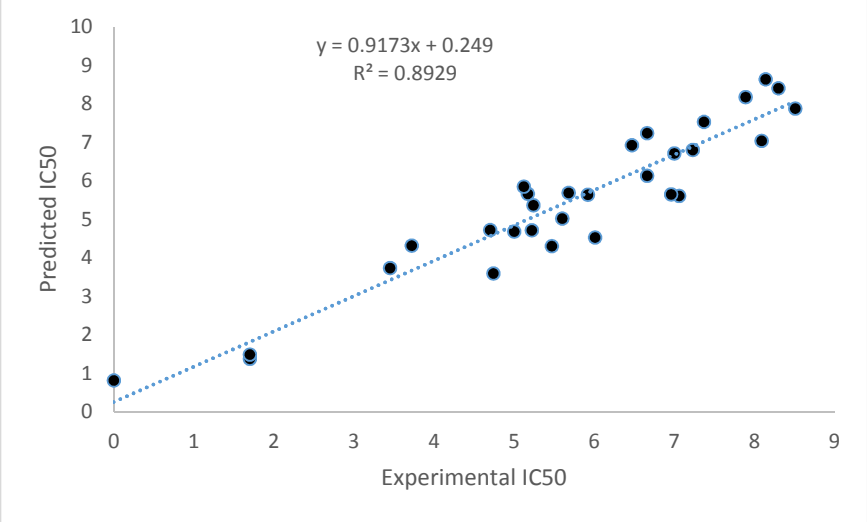

Fig. 8. The calculated PIC50 versus the experimental PIC50 for the test set

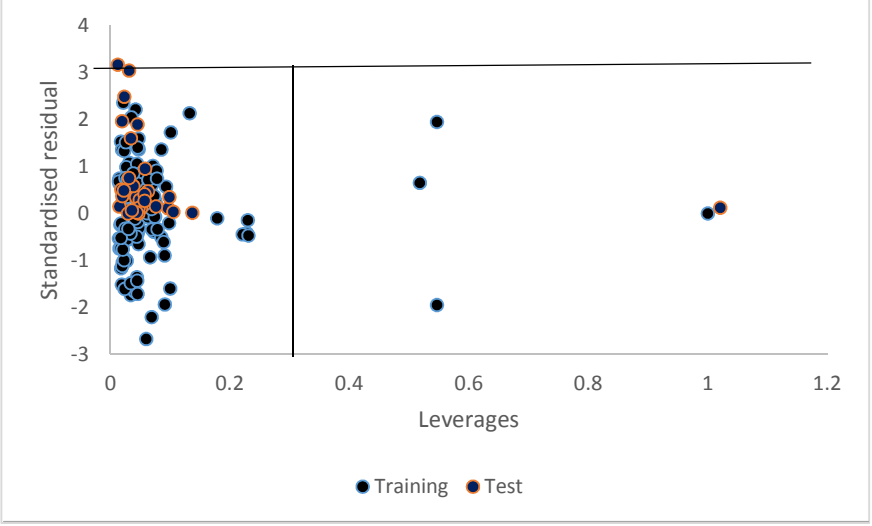

Fig. 9. The William plot, the plot for the standardized residuals versus the leverage value

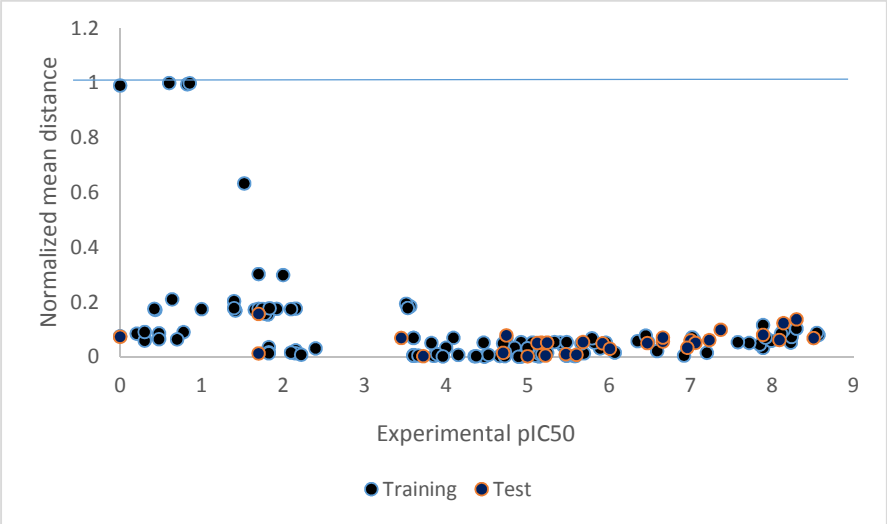

Fig. 10. Euclidean distance norms, the plot for normalized mean distance versus experimental pIC50

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.

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

Where MFj represents the mean effect for the considered descriptor j, βj is the coefficient of the descriptor j, dij 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. 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.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Mehellou Y, De Clercq E. Twenty-six years of anti- HIV drug discovery: Where do we stand and where do we go? Journal of Medicinal Chemistry. 2010;53:521-5.
2. Pomerantz RJ, Horn DL. Twenty years of therapy for HIV-1 infection. Nature Medicine. 2003;9(7):867-873.
3. Dyatkina NB, Roberts CD, Keicher JD, Dai Y, Nadherny JP, et al. Minor groove DNA binders as antimicrobial agents. 1. Pyrrole tetraamides are potent antibacterials against vancomycin resistant *Enteroccoci* and methicillin resistant *Staphylococcus aureus*. Journal Medicinal Chemistry. 2002;45:805-817.

4. Loveday C. Nucleoside reverse transcriptase inhibitor resistance. Journal of Acquire Immune Deficiency Syndrome. 2001;26(Suppl1):S25-S33.
5. Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. British Medical Journal. 2001; 323(7311):487.
6. Volberding PL. Introduction to acquire immune deficiency syndrome. Journal of International Perspectives on Antiretroviral Resistance. 2001;26(Suppl1):S1-S2.
7. Cooley LA, Lewin SR. HIV-1 cell entry and advances in viral entry inhibitor therapy. Journal of Clinical Virology. 2003;26(2): 121-32.
8. Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, Zimmer S, Young MP, Jenkins JL, Glick M, Glen RC, Bender A. From in silico target prediction to multi-target drug design: Current databases, methods and applications. Journal of Proteomics. 2011;74(12):2554–2574.
9. Ma XH, Shi Z, Tan C, Jiang Y, Go ML, Low BC, Chen YZ. In-Silico approaches to multi-target drug discovery. Pharmaceutical Research. 2010;27(5): 739–749.
10. Ajmani S, Kulkarni SA. Application of GQSAR for scaffold hopping and lead optimization in multitarget inhibitors. Molecular Informatics. 2012;31(6–7):473–490.
11. Labute P. A widely applicable set of descriptors. Journal of Molecular Graphic and Modelling. 2000;18(4-5):464-77.
12. Xu J, Stevenson J. Drug-like index: A new approach to measure drug-like compounds and their diversity. Journal Chemical Information and Computer Science. 2000; 40:1177-87.
13. Prado-Prado FJ, González-Díaz H, de la Vega OM, Ubeira FM, Chou KC. 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. 2008; 16(11):5871-80.
14. Todeschini R, Consonni V. Handbook of molecular descriptors. Wiley-VCH, Weinheim; 2000.
15. González-Díaz H, Prado-Prado FJ. Unified QSAR and network-based computational chemistry approach to antimicrobials, part 1: Multispecies activity models for antifungals. Journal of Computational Chemistry. 2008;29(4):656–667.
16. Enrique Molina, Humberto González Díaz, Maykel Pérez González, Elismary Rodríguez, Eugenio Uriarte. Designing antibacterial compounds through a topological substructural approach. Journal of Chemical Information and Computer Science. 2004;44(2):515–521.
17. Edache EI, Uzairu A, Abechi SE. 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. 2015;4:374-390.
18. Luco JM, Ferretti FH. 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. 1997;37:392-401.
19. Wavefunction, Inc. Spartan'14, version 1.1.2, Irvine, California, USA; 2013.
20. Becke AD. A new mixing of Hartree-Fock and local density-functional theories. Journal of Chemical Physics. 1993;98: 1372－1377.
21. Petersson GA, Bennett A, Tensfeldt TG, Al-Laham MA, Shirley WA, Mantzaris J. 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. 1988;89: 2193－2218..
22. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 2011;32(7): 1466-1474.
23. Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. Journal of Chemical Information and Computer Science. 1994;34:854-866.
24. Hansch C, Sammes PG, Taylor JB. Comprehensive medicinal chemistry: The rational design, mechanistic study & therapeutic application of chemical compounds. Pergamon, New York. 1990;6: 1.
25. Golbraikh A, Tropsha A. Beware of q2! Journal Molecular Graphic and Modelling. 2002;20:269-276.

26. Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR and Combinatorial Science. 2008;27:302-313.
27. Pogliani L. Structure-property relationships of amino acids and some dipeptides. Journal of Amino Acids. 1994;6:141-153.
28. Pogliani L. Modeling with special descriptors derived from a medium-sized set of connectivity indices. Journal of Physical Chemistry. 1996;100:18065-18077.
29. Verma RP, Hansch C. QSAR modeling of taxane analogues against colon cancer. European Journal of Medicinal Chemistry. 2010;45:1470-1477.
30. Roy K, Paul S. Exploring 2D and 3D QSARs of 2,4-diphenyl-1,3-oxazolines for ovicidal activity against *Tetranychus urticae.* QSAR and Combinatorial Science. 2008;28:406-425.
31. Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT. Carbonic anhydrase inhibitors: The first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorganic and Medicinal Chemistry Letter. 2004;14:3283-3290.
32. Shapiro S, Guggenheim B. Inhibition of oral bacteria by phenolic compounds: Part 1. QSAR analysis using molecular connectivity. Quantitative Structure Activity Relationship. 1998;17:327-337.
33. Kiralj R, Ferreira MMC. Basic validation procedures for regression models in QSAR and QSPR studies: Theory and Application. Journal of Brazilian Chemical Society. 2009;20(4):770-787.
34. Deswal S, Roy N. Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors. European Journal of Medicinal Chemistry. 2006;41:1339-1346.
35. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, Van De Sandt JJM, Tong W, Veith G, Yang C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. Alternative to Laboratory Animal. 2005;33:155-173.
36. OECD, 2007. Guidance document on the validation of (Quantitative) Structure–Activity Relationships [(Q)SARs] models.
Available:http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)2S
37. Minovski N, Zuperl S, Drgan V, Novi M. 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. 2013;759:28–42.
38. Hoaglin DC, Welsch RE. The hat matrix in regression and ANOVA. The American Statistician. 1978;32(1):17-22.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/12951