International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: http://dx.doi.org/10.7324/IJASRE.2017.32521

Vol.3 (10) November-2017

InsilicoSearch for Molecules with Enhanced Histone Deacetylase Inhibitory Properties as Templates for Novel Anticancer Agents

AbdullahiMoyosore*[1]

[1]*Department of Chemistry,*
Federal College of Education,
Kastina, Nigeria

ABSTRACT

Histone dacetylases (HDACs) are a group of enzymes that remove acetyl groups from histones and regulate expression of tumor suppressor genes making them a promising therapeutic target for treatment of cancer by developing a wide variety of inhibitors. Developing these inhibitors requires accurate understanding of how their molecular structures are link to their respective inhibitory properties. A Genetic Function Approximation based Multi-linear regression Quantitative structure activity relationship modelling was performed on a data set of 29 HDAC inhibitors using Semi-empirical (PM3) computational level of theory. The best QSAR model reveals that FMF, Kier3, n5HeteroRing, globaltopo and Kier1 descriptors have pronounced influence on the HDAC inhibitory properties of the compounds. The validation parameters of the best model are LOF = 0.137, R^2 = 0.933, R^2_{adj} = 0.902, Q^2_{LOO} = 0.841, F-value = 30.239, $R^2_{pred.}$ = 0.6495. The wealth of information provided by this model will undoubtedly be of immense help in the structural modifications of the studied molecules as a guide to discover additional HDAC inhibitors with greater therapeutic utility.

Keywords*: Histone dacetylases, DFT, Kier1, Cancer, QSAR, GFA, PM3*

1.0 INTRODUCTION

Cancer is a malignant growth or tumor resulting from uncontrolled division of cells as a result of genetic and genomic alterations such as amplifications, translocations, deletions, and point mutations [1].Going by the 2016 report of the cancer facts and figures about 1,685,210 new cancer cases are expected to be diagnosed in 2016 [2]. This estimate has been reported to exclude carcinoma in situ (noninvasive cancer) and basal cell or squamous cell skin cancers [2]. It has also been reported that about 595,690 Americans are expected to die of cancer in 2016, which translates to about 1,630 people per day. Cancer is the second most common cause of death in the US, exceeded only by heart disease, and accounts for nearly 1 of every 4 deaths [2].

According to estimates from the International Agency for Research on Cancer (IARC), there were 12.7 million new cancer cases in 2008 worldwide, with economically developing countries having 7.1 million cases[3]. This report reveals that developing nations of the world are not left out as far as exposure to the riskposed by the prevalence of this disease is concern. In view of the morbidity and mortality of this diseaseand its threatening prevalence despite existing treatment for it, it has become necessary to search for newer drug candidates that will curb this disease.

Histone deacetylases (HDACs) simply refers to a group of enzymes that eliminate acetyl groups from histones and regulate expression of tumor suppressor genes. They are implicated in cancer making them a promising therapeutic target for treatment of this disease by developing a wide variety of inhibitors [4]. Inhibitors of HDACs interfere with HDAC activity and regulate biological events, such as cell cycle, differentiation and apoptosis in cancer cells. As a result, HDAC inhibitor-based therapies have gained much attention for cancer treatment [1].

In order to discover new, hopefully more therapeutically efficacious HDAC inhibitors, adequate knowledge of the dominant structural features (molecular descriptors) influencing the observed HDAC inhibitory activities of molecules has become a sine qua non. These descriptors if known can be modified in the molecules to give rise to highly potent inhibitors of this enzyme.

Quantitative Structure Activity Relationship (QSAR)modelling has much to offer in this regard. QSAR establishes the mathematical relationship between physic-chemical properties or biological activities of interest and measurable or computable parameters called molecular descriptors [5]. The fundamental principle underlying QSAR is that the difference in structural properties is responsible for the variations in biological activities of the compounds. It assumes that the potency of a certain biological activity exerted by a series of congeneric compounds is a function of various physicochemical parameters of the compounds. Once statistical analysis shows that certain physico-chemical properties are favorable to the concerned activity, the concerned activity can be optimized by choosing such substituents which would enhance such physicochemical properties[6]. This practice has formed an integral part of computer aided drug design (CADD) as it helps to minimize the trial and error techniques employed in traditional method of drug discovery and development by not using leads that will likely not be successful, minimizes animal usage, reduces time and cost of drug discoveryas well as promoting green chemistry via the reduction of waste and improved efficiency.

More recently, Thangapandian*et al*. [7] performed a Genetic Function Approximation (GFA) based QSAR studies on HDAC8 Inhibitors. The compounds in the data set were subjected to energy minimization using CHARMM force fieldto generate the lowest energy conformation of every compounds. The descriptors were calculated using Dragon program. The two best QSAR models generated have squared correlation coefficient R^2 value of 0.505 and 0.515.

In this work, the selected HDAC8 Inhibitors were optimized with the aid of Density Functional Theory (DFT), a higher and better level of theory than that used by Thangapandian*et al*. [7] Likewise, our validation parameters revealed that our GFA-QSAR models are more robust, stable and reliable.

The aim of this work is two-fold; to harness the dominant structural features responsible for the observed HDAC8 Inhibitory activities of the studied molecules and to build robust and rational GFA based QSAR model for predicting this bioactivity in molecules that fall within the applicability domain of the best QSAR model.

2.0 MATERIALS AND METHODS

In the present study, QSAR studies were performed using Hansch's approach[8]. In Hansch's approach, structural properties of compounds are calculated in terms of different physicochemical parameters and these parameters are correlated with biological activity through equation using regression analysis. The various steps are presented in flowchart in Figure 1

A novel set of 30 HDAC8 Inhibitors were gotten from literature [7]. The general molecular structures of the studied compounds are shown in Table 1. The inhibitory activity values of these compounds were calculated in IC_{50}values which were converted to –logarithmic (-logIC50 or pIC50) scale to be utilized in this study. These operation was performed in order to reduce the dispersion of data set and to get linear response and good data fitting.

Spartan 14 V.1.1.0 program (Spartan 14) was used to get the minimum energy geometry of each molecule in the data set. The optimization was performed using the DFT (B3LYP) and 6-31G* basis set. The lowest energy structure was used for each molecule to calculate their physicochemical properties (molecular descriptors).

Descriptors are the numerical representation of molecular structures. The information about any molecular structure is encoded by descriptors. The molecular descriptors ranging from 0D, 2D and 3D used in this QSAR modelling were calculated using *Padel descriptor* tool kit, *Spartan'14* softwares.

The data set of the 29 compounds was split into 70% training set (20 compounds) and 30% test set (9 compounds). The training set was used to adjust the parameters of the model while the test set was used to evaluate its prediction ability. In the model building stage, the correlations between pIC_{50} of the compounds and the calculated descriptors were obtained via correlation analysis using the Microsoft excel package in Microsoft office 2013. Pearson's correlation matrix was used as a qualitative model, in order to select the suitable descriptors for the GFA regression analysis.

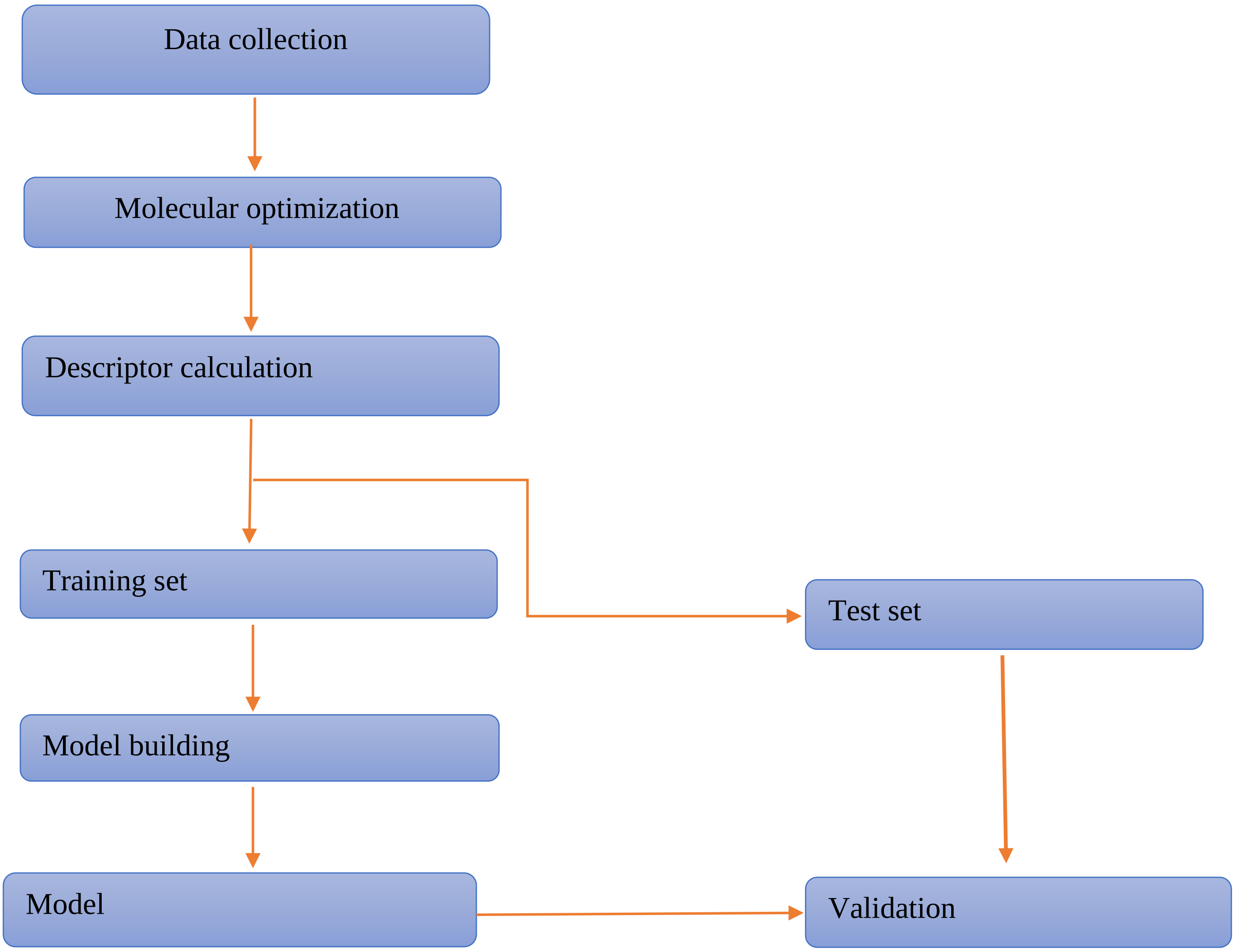
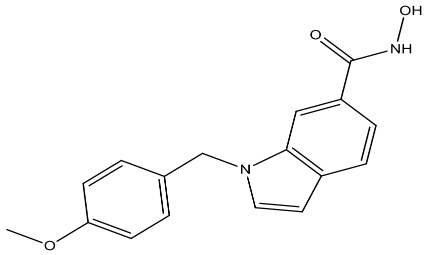
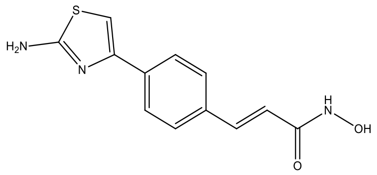
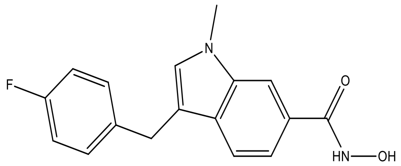
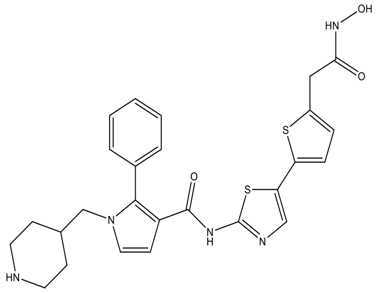
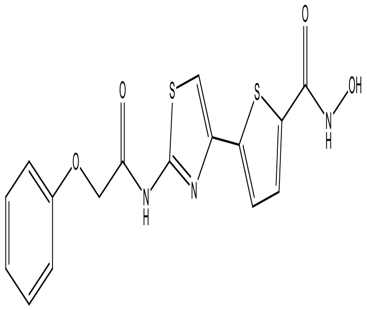
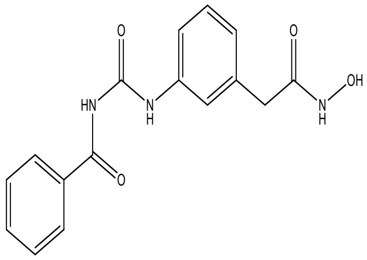
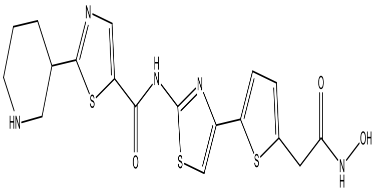
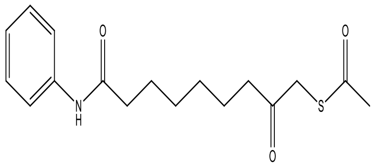
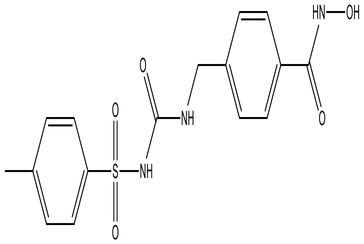
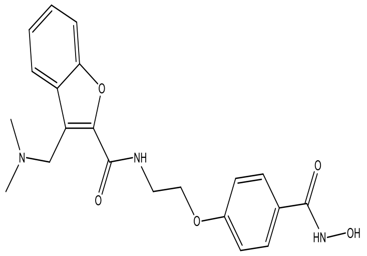
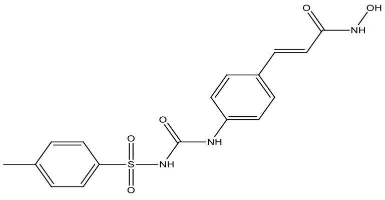
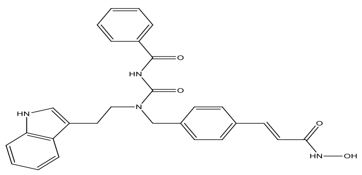
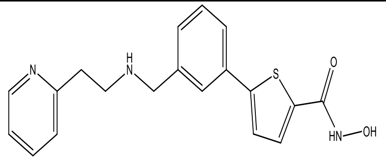
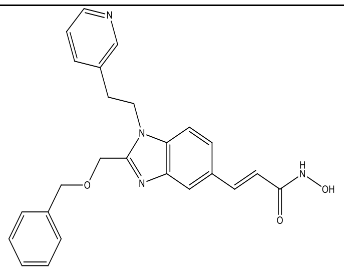
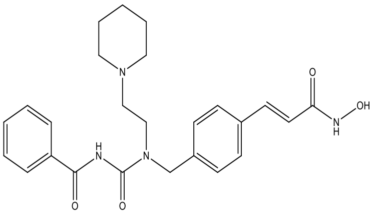
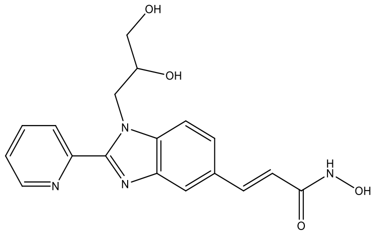
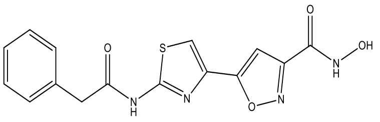
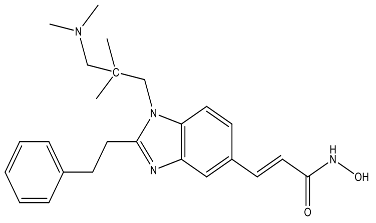
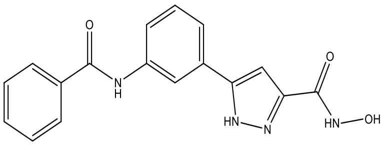
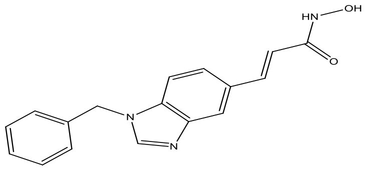
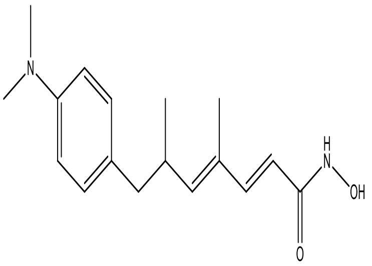
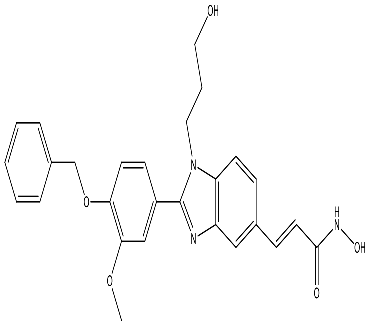
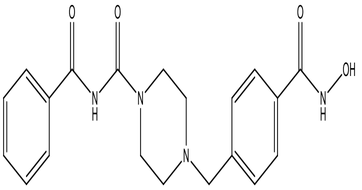
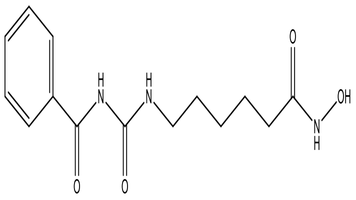
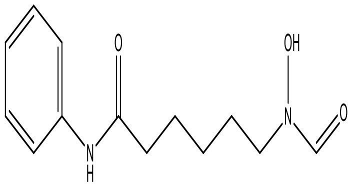
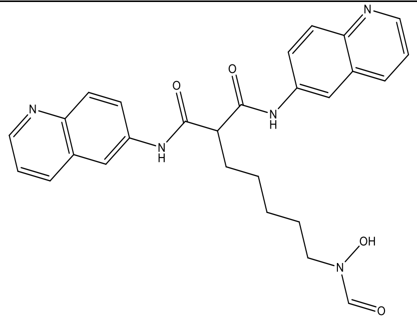
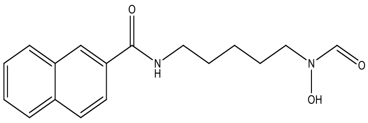
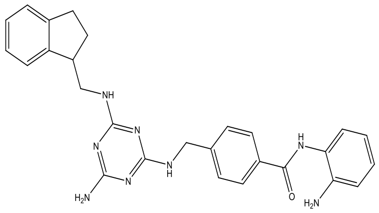
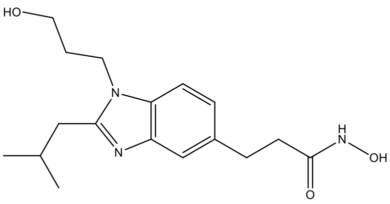


Figure 1: QSAR methodology flowchart. Source: Ameji*et al*. [9]

Table 1: Chemical Structure and Experimental pIC$_{50}$ of the Data set

Cpd	Structure	pIC$_{50}$	Cpd	Structure	pIC$_{50}$
1*		2.000	8		1.000
2		1.620	9		0.854
3		1.420	10*		0.824
4*		1.398	11		0.721
5		1.310	12		0.721
6		1.398	13*		0.678

7*		1.301	15		0.523
14		0.638	17		0.463
16*		0.469	19*		0.452
18		0.456	21		0.385
20		0.398	23		0.102
22*		0.337	25		-0.029
24		0.097	26*		-0.447

					
27		-0.845	29		-1.544
28		-0.531			

Key: *Test set compound

The selected descriptors were subjected to regression analysis with the pIC_{50} as the dependent variable and the selected descriptors as the independent variables using Genetic function approximation (GFA) method in Material studio software. GFA could create a population of models rather than a single model, a distinctive feature of this method. GFA algorithm, selecting the basic functions genetically, developed better models than those made using stepwise regression methods. And then, the models were estimated using the "lack of fit" (LOF), which was measured using a slight variation of the original Friedman formula, so that best model received the best fitness score [10].

In Materials Studio, LOF is measured using a slight variation of the original Friedman formula[11].The revised formula is:

$$LOF = \text{SSE} / (1 - \frac{C+dp}{M})^2 \qquad (1)$$

Where SSE is the sum of squares of errors, c is the number of terms in the model, other than the constant term, d is a user-defined smoothing parameter, p is the total number of descriptors contained in all model terms (ignoring the constant term) and M is the number of samples in the training set. Unlike the commonly used least squares measure, the LOF measure cannot always be reduced by adding more terms to the regression model. While the new term may reduce the SSE, it also increases the values of c and p, which tends to increase the LOF score. Thus, adding a new term may reduce the SSE, but actually increases the LOF score. By limiting the tendency to simply add more terms, the LOF measure resists over fitting better than the SSE measure[12].

The internal validation of the best model was performed using the well-known scheme of "leave-one-out" (LOO) cross-validation. Usually, the square of LOO cross-validation coefficient (q^2)should be > 0.5 for a reliable model. Other validation parameters deployed in this study include the square of the correlation coefficient, R^2 (threshold of ≥ 0.6)[13]. External validation is also crucial to obtain QSAR models with more reliable predictive abilities. The optimum QSAR model was externally validated using the test set of 9 molecules (Table 5)with the aid of equation 2. Generally, a QSAR model is accepted to own high predictive power only if the square of predictive correlation coefficient (R^2pred) is greater than 0.6 for the test set [13].

$$R^2_{pred.} = 1 - \frac{\sum[Ypred(te) - Yobs(te)]^2}{\sum[Yobs(te) - Ym(tr)]^2} (2)$$

Ypred.(te) and Y(te) indicate predicted and observed activity values respectively of the test set compounds and Ym(tr) indicates mean activity value of the training set [13].

3.0 RESULT AND DISCUSSION

Model 1gives the best Genetic Function Approximation derived (GFA) QSAR model for predicting the $pIC_{50}$of the studied HDAC Inhibitors.Model 1 was chosen as the best model owing to its least LOF value.Likewise, its validation parameters are in good agreement with the standard validation metrics for a robust QSAR model proposed byRavinchandran*et al*. [13]. The definition of the descriptors in the models are presented in Table 2.

Model 1:

$$pIC_{50} = 0.486918520\,ALogP - 0.220621099\,nH - 0.147538828\,nN - 0.222922364\,nBondsM + 0.065648498\,C2SP2 - 0.005082877\,ECCEN + 0.129605484\,MDEC - 23 + 3.057207376\,MDEO - 22 + 0.033084474\,MW - 2.068570970$$

LOF = 0.075, R^2 = 0.987, R^2_{adj} = 0.975, Q^2_{LOO} = 0.921, F-value = 83.581, Min expt. error for non-significant LOF (95%) = 0.09

Table 2: Detailed definition of descriptors

Descriptor	Definition
ALogP	Octanol-water partition coefficient
Nh	Number of hydrogen atom
Nn	Number of nitrogen atom
nBondsM	Total number of bonds that have bond order greater than one
ECCEN	A topological descriptor combining distance and adjacency information
MDEC-23	Molecular distance edge between all secondary and tertiary carbons
MDEO-22	Molecular distance edge between all secondary oxygens
MW	Molecular weight
C2SP2	Doubly bound carbon bound to two other carbons

The closeness of coefficient of determination (R^2) to its absolute value of 1.0 is an indication that the model explained a very high percentage of the response variable (descriptor) variation, high enough for a robust QSAR model. The high adjusted R^2 (R^2_{adj}) value and its closeness in value to the value of R^2 implies that the model has excellent explanatory power to the descriptors in it. Also, the high and closeness of Q^2 value to R^2

revealed that the model was not over fitted. F value judges the overall significance of the regression coefficients. The high F value of the model is an indication that the regression coefficients are significant

The comparison of observed and predicted antibacterial activities of the complexes is presented in Table 3. The sound predictability of model 1 is evidenced by the low residual values observed in the Table. Also, the high linearity of the plot of predicted pIC_{50}against observed pIC_{50} shown in Figure 2.1 indicates that the model is well trained and it predicts well the pIC_{50} of the compounds.

To ascertain whether there exists a systematic error in the model development, the residual pIC_{50} was plotted against observed pIC_{50} (Figure 2.2). The propagation of residuals on both sides of zero indicated that there was no systemic error in model development [14].

The P-value of the optimization model at 95% confidence level shown in Table 4 has α value < 0.05. This reveals that the alternative hypothesis that the magnitude of the observed HDACinhibitory activity of the molecules is a direct function of the descriptors of their total chemical structure takes preference over the null hypothesis which states otherwise.

The positive coefficients of the descriptors; ALogP, MDEC-23, MDEO-22, MW implies that the Histone Deacetylase Inhibitory activity of the compound varies directly with the values of these descriptors. Thus, for an enhanced inhibitory activity of a molecule against the target enzyme, the values of these descriptors should be considerably high. In a similar vein, the values of nH, nN, nBondsM and ECCEN descriptors should be made minimal in the Histone Deacetylase inhibitors for their enhanced bioactivity against the enzyme owing to their negative correlation to pIC_{50} as shown in the optimization model.

Table 3: Comparison of experimentalpIC_{50} and predictedpIC_{50} of model 1

ExperimentalpIC_{50}	ExperimentalpIC_{50}	PredictedpIC_{50}	Residual
1.62000000	1.62000000	1.60382000	0.01618000
1.42000000	1.42000000	1.58151000	-0.16151000
1.31000000	1.31000000	1.17246400	0.13753600
1.39800000	1.39800000	1.43368400	-0.03568400
1.00000000	1.00000000	0.99154800	0.00845200
0.85400000	0.85400000	0.74430200	0.10969800
0.72100000	0.72100000	0.64920600	0.07179400
0.72100000	0.72100000	0.72100000	0.00000000
0.63800000	0.63800000	0.59138600	0.04661400
0.52300000	0.52300000	0.61872700	-0.09572700
0.46300000	0.46300000	0.58007300	-0.11707300
0.45600000	0.45600000	0.27709000	0.17891000
0.39800000	0.39800000	0.51330000	-0.11530000
0.38500000	0.38500000	0.28826700	0.09673300
0.09700000	0.09700000	0.13364200	-0.03664200
-0.02900000	-0.02900000	0.03188700	-0.06088700
-0.59000000	-0.59000000	-0.63942300	0.04942300
-0.84500000	-0.84500000	-0.73832900	-0.10667100
-0.53100000	-0.53100000	-0.57947800	0.04847800
-1.54400000	-1.54400000	-1.50967700	-0.03432300

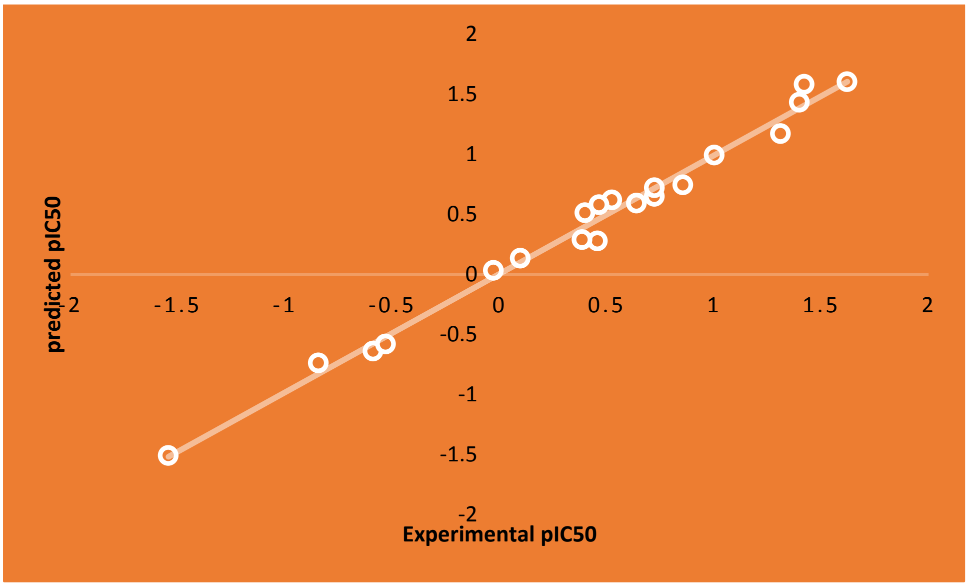


Figure 2.1: Plot of PredictedpIC$_{50}$ against predicted pIC$_{50}$

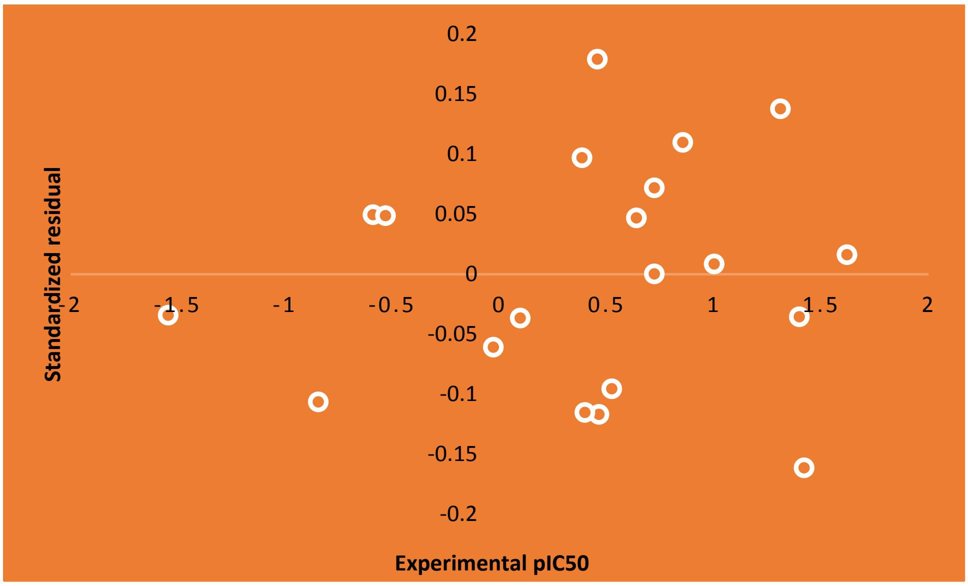


Figure 2.2: Residual plot of model 1

Table4: P-value of model 1 at 95% confidence level

Source	SS	DF	MS	F	p-value
Difference	7.019	4	0.7612	18.2945	<0.0001
Error	0.449	18	0.0416		
Null model	7.468	22	0.311		

The model predicted well the external test set compounds as shown in Table 5 except for Compounds 1, 13 and 26having abnormally high residual value, thus they are treated as structural outliers. The plot of the experimental pIC$_{50}$ verses predicted pIC$_{50}$ shown in Figure 2.3 shows a high degree of agreement (i.e R^2 =

0.7055), an indication that the model is capable of providing valid predictions for new molecules that falls within its applicability domain.

Table 5: External validation Table for the optimum model (Model 1)

Test set compound	Experimental pIC_{50}	Predicted pIC_{50}	Residual
13	0.678	1.957729	-1.27973
16	0.469	0.576994	-0.10799
10	0.824	0.771003	0.052997
19	0.452	0.497421	-0.04542
26	-0.447	0.791366	-1.23837
22	0.337	0.386201	-0.0492
1	2	-1.391	3.390998
4	1.398	0.946251	0.451749
7	0.824	0.445174	0.378826

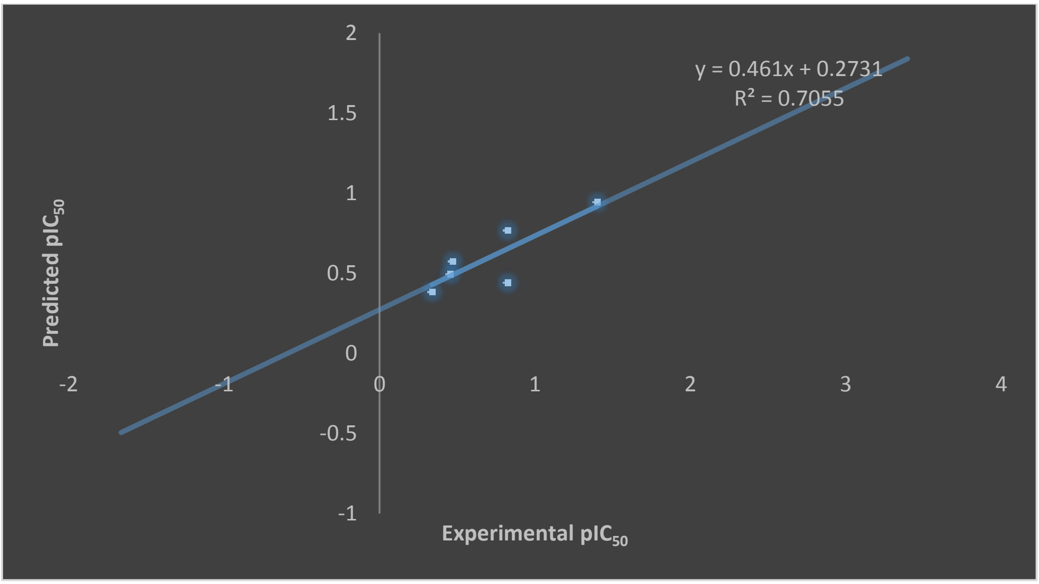


Figure 2.3: Plot of Predicted pIC_{50} against Experimental pIC_{50} of test set compounds

4.0 CONCLUSION

The aim of this study has been fully achieved; the dominant structural features responsible for HDAC inhibitory activities of the studied molecules has been successfully harnessed. The validity of the optimum QSAR model has been ascertained internally and externally. The wealth of information in this work will undoubtedly be of immense help in the structural modifications of the studied molecules as a guide to discover additional HDAC inhibitors with greater therapeutic utility.

REFERENCES

1. M. Mottamal, S. Zheng, T.L. Huang, G. Wang. Histone Deacetylase Inhibitors in Clinical Studies as Templates for New Anticancer Agents. *Molecules* **2015**, 20, 3898-3941; doi:10.3390/molecules20033898.

2. Cancer Facts and Figures (CFF). Corporate Center: American Cancer Society Inc. 250 Williams Street, NW, Atlanta, GA 30303-1002 404-320-3333.

3. J.Ferlay, H.R. Shin, F. Bray, D. Forman, C.D. Mathers, D. Parkin D. GLOBOCAN 2008, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No.10 [Internet]. Lyon, France: International Agency for Research on Cancer. 2010; Available from: http://globocan.iarc.fr

4. J.E. Bolden, M.J. Peart, R.W. Johnstone. Anticancer activities of histone deacetylase inhibitors. Nat. Rev. Drug Discov. 2006, 5, 769–784.

5. A.K. Rathod. Antifungal and Antibacterial activities of Imidazolylpyrimidines derivatives and their QSAR Studies under Conventional and Microwave-assisted. International Journal of PharmTech Research. 2011; 3 (4),1942-1951.

6. E.N. Bharath, S.N. Manjula, A. vijaychand. In silicodrug design tool for overcoming the innovation deficit in the drug discovery process**.** *International Journal of Pharmaceutical Sciences*, 2011; 3 (2): 8-12.

7. S. Thangapandian, S. John, K.W. Lee.Genetic Function Approximation and Bayesian Models for the Discovery of Future HDAC8 Inhibitors. Interdisciplinary Bio Central. IBC 2011; 3:15, 1-11 • DOI: 10.4051 / ibc.2011.3.4.0015

8. C. Hansch, T.P.Fujita. Analysis: A Method for the correlation of biological activity and chemical structure. *Journal of American Chemical Society*. 1964; *86:* 1616-1626.

9. J.P. Ameji, A. Uzairu, S.O. Idris. Quantitative structure activity relationship study of nickel schiff base complexes as potent anti-candida albicans agents. Journal of Computational Methods in Molecular Design, 2015, 5 (3):120-134.

10. W. Wu, C. Zhang, W. Lin, Q. Chen, X. Guo, Y. Qian. Quantitative Structure-Property Relationship (QSPR) Modeling of Drug-Loaded Polymeric Micelles via Genetic Function Approximation. *PLoSONE*, 2015; 10(3): e0119575. doi: 10.1371/journal.pone.0119575.

11. J.F. Friedman. *Multivariate Adaptive Regression Splines, Technical Report No. 102*, Stanford University, 1990

12. K.F. Khaled, N.S. Abdel-Shafi. Quantitative Structure and Activity Relationship Modeling Study of Corrosion Inhibitors: Genetic Function Approximation and Molecular Dynamics Simulation Methods. *International Journal of Electrochemical Science*, 2011; 6: 4077 – 4094.

13. V. Ravichandran, R. Harish,J. Abhishek, S. Shalini, P.V. Christapher, K.A. Ram. Validation of QSAR models - strategiesandimportance.*International Journal of Drug Design and Discovery*, 2011, 2: 511-519.

14. M. Heravi-Jalali,A. Kyani. Use of computer-assisted methods for the modelling of retention time of a variety of volatile organic compounds: a PCA –MLR-ANN approach. *Journal of chemical informatics and computer science*, 2004; 44: 1328-1335