Journal of Computational Methods in Molecular Design, 2016, 6 (3):55-63


ISSN : 2231- 3176
CODEN (USA): JCMMDA

# Towards the Discovery of Novel Anti-cancer agents through Semi-empirical (PM3) Based QSAR Modelling of Histone Deacetylase Inhibitors

**Bamidele M. Omotola[1*], Jimoh Abdullateef[2], Ojilere C. Juliet[2] and Ibraheem A. Wasiu[2]**

*[1]Nigerian Institute of Leather Science and Technology, Zaria Nigeria*
*[2]Ahmadu Bello University, Zaria 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, Semi-empirical, Kier1, Cancer, QSAR

---

## INTRODUCTION

Cancer is a group of diseases characterized by uncontrolled growth and spread of abnormal cells. 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 [1]. More recently, the 2016 Cancer Facts and Figures (CFF) revealed that about 1,685,210 new cancer cases are expected to be diagnosed in 2016 globally [2], with these figures, the disease poses serious health risk to man. Incessant search for newer therapeutic agents is certainly not debatable.

Histone dacetylases (HDACs) stand as promising therapeutic targets for treatment of cancer because these compounds are highly implicated in this disease. HDAC inhibitors 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. To date, the FDA has approved three HDAC inhibitors for cutaneous/peripheral T-cell lymphoma and many more HDAC inhibitors are in different stages of clinical development for the treatment of hematological malignancies as well as solid tumors [3].

Concerted efforts aimed at discovering new and hopefully more therapeutically efficacious HDAC inhibitors entails adequate harnessing of the molecular descriptors having direct link with this bioactivity (i.e., HDAC inhibitory

properties). These descriptors can be optimized in the molecules to achieve the aforementioned aim. This can be achieved via Quantitative Structure Activity (QSAR) Modelling.

Quantitative structure activity relationship (QSAR) study provides medicinal chemists valuable information that is useful for drug design and prediction of drug activity. QSAR models are mathematical equations which construct a relationship between chemical structures and their biological activities as a linear regression model in the form Y = Xb + e. This equation may be used to describe a set of predictor variables (X) with a predicted variable (y) by means of a regression vector (b) [4].The physicochemical properties predicted from structure are helpful in the search for new molecules of similar or increased biological activity. QSAR studies enable the investigators to establish reliable quantitative relationships, to derive a QSAR model, and predict the activity of novel molecules prior to their synthesis. These studies reduce the trial-and-error element in the design of compounds by establishing mathematical relationships between physical, chemical, biological, or environmental activities of interest and measurable or computable physicochemical, electronic, topological, or stereo chemical parameters [5].

The aim of this work is two-fold; to harness the principal molecular descriptors responsible for the observed HDAC Inhibitory activities of the studied molecules and to build robust QSAR model for predicting this bioactivity in HDAC inhibitors.

## MATERIALS AND METHODS

The general scheme for this work is depicted in Fig. 1. A set of 29HDAC Inhibitors were gotten from literature [6].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.

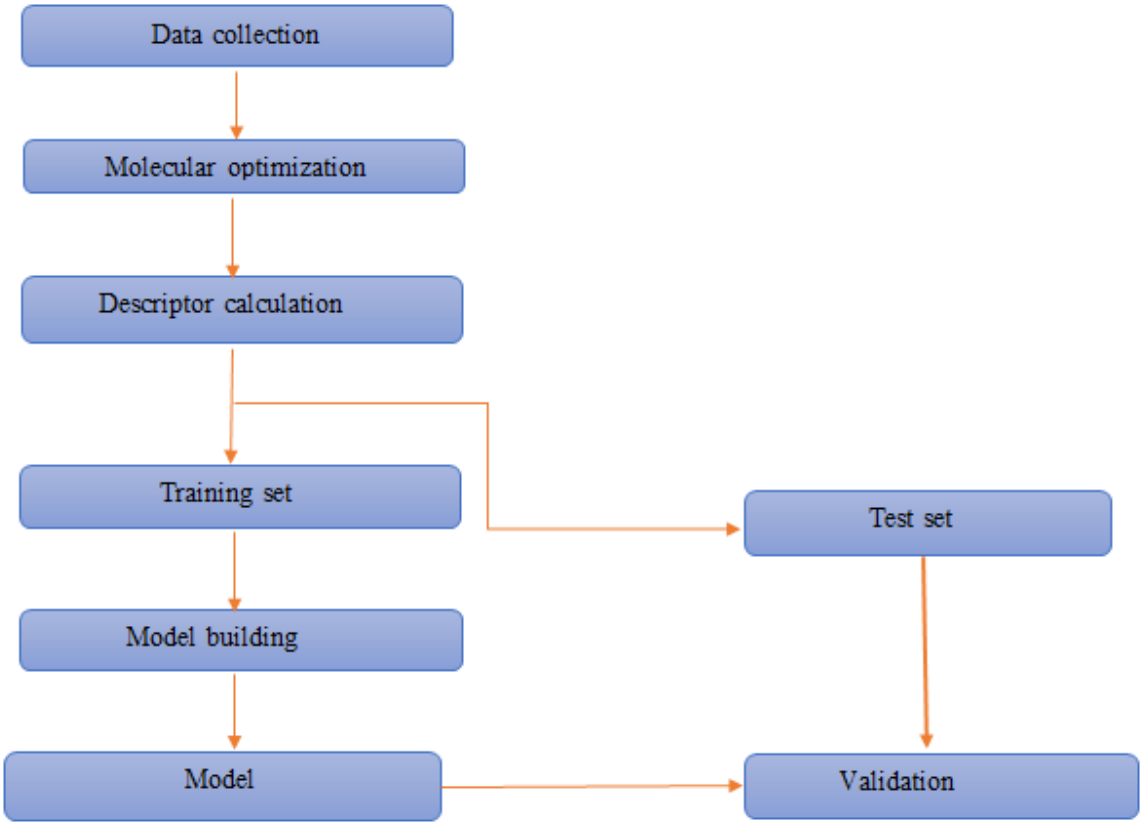
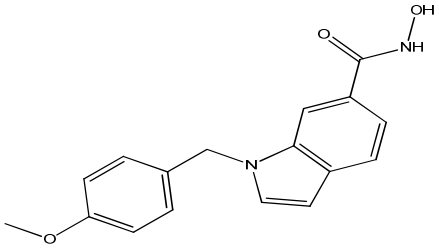
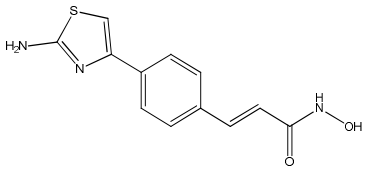
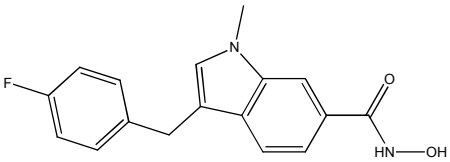
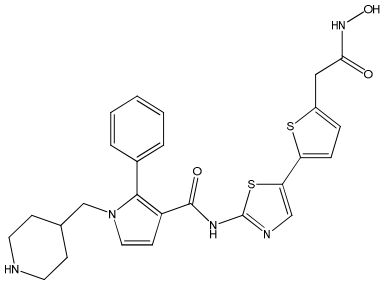
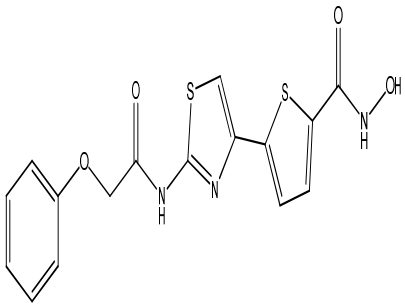
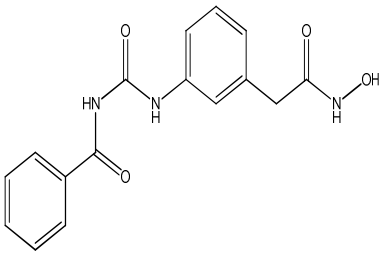
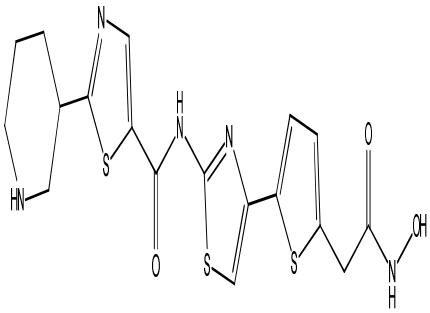
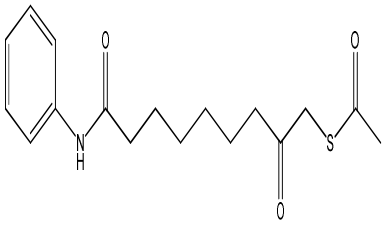
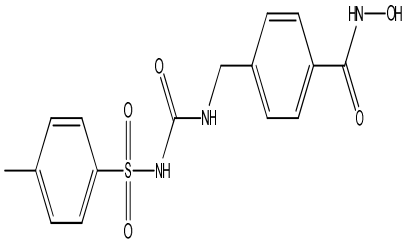
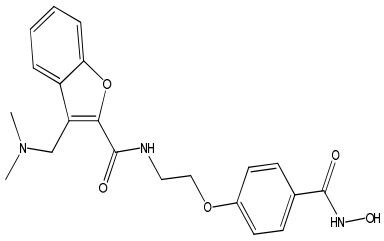
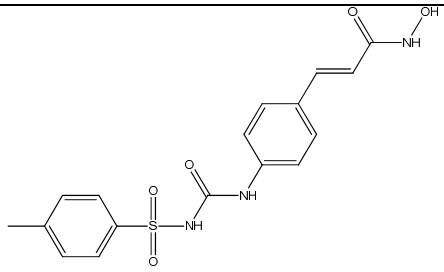
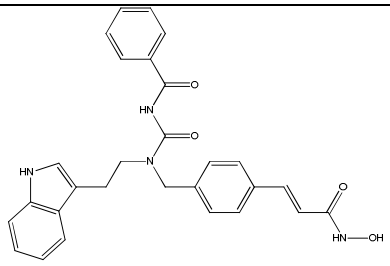
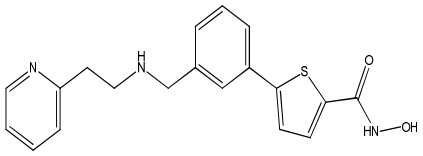
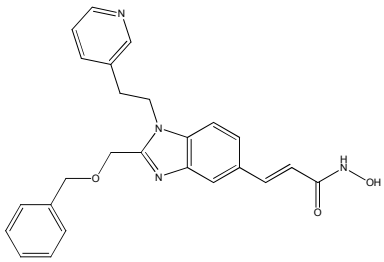
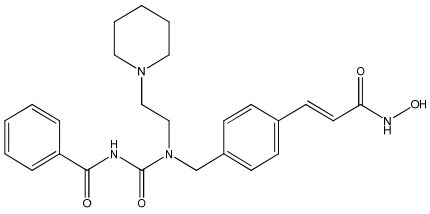
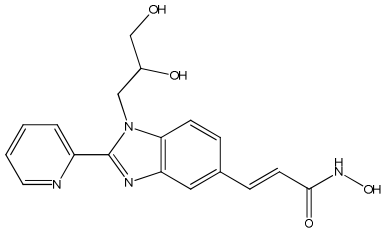
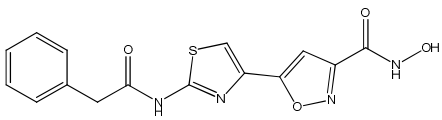
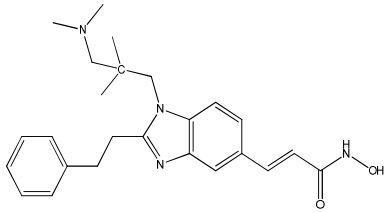
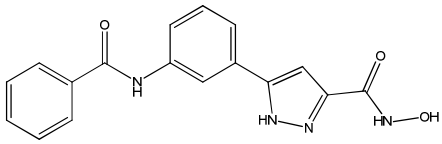
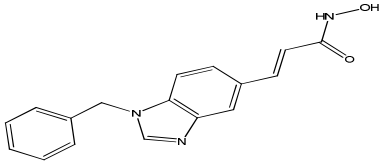
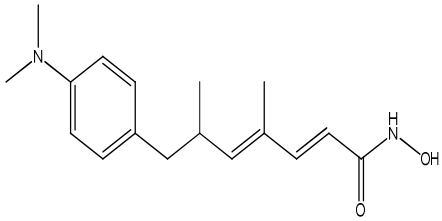
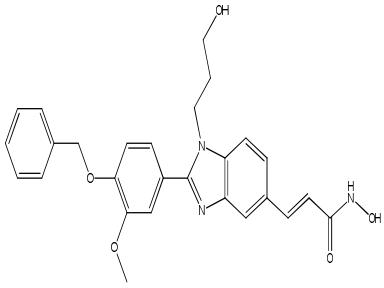
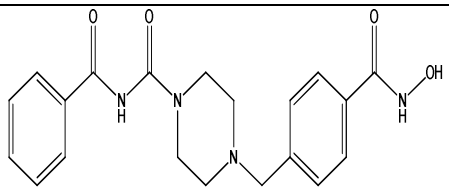
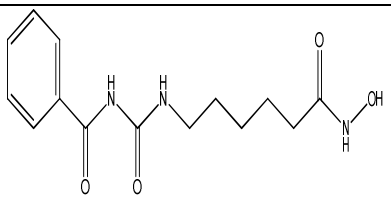
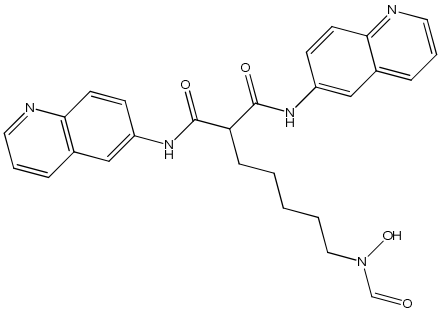
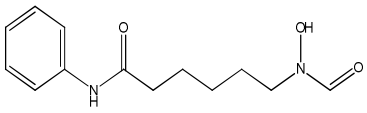
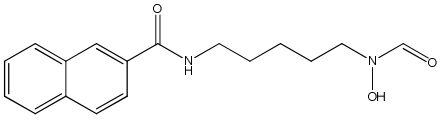
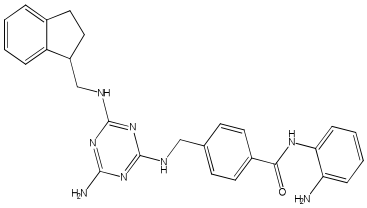
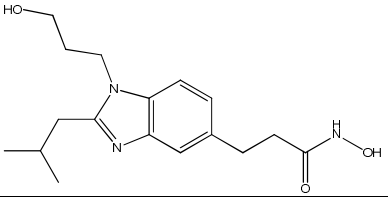



**Fig 1: QSAR methodology flowchart**

**Table 1: Chemical Structure and Experimental $pIC_{50}$ of the Data set**

| Cd | Structure | $pIC_{50}$ | Cd | 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*

Molecular optimization was performed at Semi-empirical (PM3) level of theory with the aid of ***Spartan 14 V.1.1.0*** program (Spartan 14). The 0D, 1D, 2D and 3D descriptors of the optimized molecules were calculated using ***Padel descriptor tool kit*** and ***Spartan'14 softwares***. Descriptors are the numerical representation of molecular structures. The information about any molecular structure is encoded by descriptors.

The data set of the 29 compounds was manually 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 to select the suitable descriptors for the GFA regression analysis. 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. The best model is the one with the least *lack of fit* (LOF) score. LOF is measured using a slight variation of the original Friedman formula, so that best model received the best fitness score [7].

In ***Materials Studio***, LOF is measured using a slight variation of the original Friedman formula (Friedman, 1990). 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 [8].

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) [9]. 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 with the aid of equation 2.1. Generally, a QSAR model is accepted to own high predictive power only if the square of predictive correlation coefficient ($R^2$pred) is greater than 0.5 for the test set [9].

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

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 [9].

## RESULTS AND DISCUSSION

Model 1gives the best Genetic Function Approximation derived QSAR model for predicting the $pIC_{50}$HDAC Inhibitors. Likewise, its validation parameters are in good agreement with the standard validation metrics for a robust QSAR model proposed by Ravinchandran *et al*. [9]. The definition of the descriptors in the models are presented in Table 2.

**Model 1:**

$$\boldsymbol{pIC_{50}} = 5.138993864\,\boldsymbol{FMF} - 0.209794614\,\boldsymbol{Kier1} + 0.575285910\,\boldsymbol{Kier3} - 30.269180407\,\boldsymbol{MDEN\text{-}11} + 0.319484327\,\boldsymbol{n5HeteroRing} + 7.620614156\,\boldsymbol{globalTopo} - 13.171621619$$

LOF = 0.137, $R^2$ = 0.933, $R^2_{adj}$ = 0.902, $Q^2_{LOO}$ = 0.841, F-value = 30.239

**Table 2: Detailed definition of descriptors**

| Descriptor symbol | Definition |
|---|---|
| FMF | Complexity of a molecule |
| Kier1 | First kappa shape index |
| Kier3 | Third kappa shape index |
| MDEN-11 | Molecular distance edge between all primary nitrogens |
| n5HeteroRing | Number of rings containing heteroatoms |
| globalTopo | Global topological charge index |

The predictability of model 1 is evidenced by the low residual values observed in Table 3 which gives the comparison of observed and predicted$pIC_{50}$ of the HDAC inhibitors. Also, the high linearity of the plot of predicted $pIC_{50}$against observed$pIC_{50}$shown in Fig. 2 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 $pIC_{50}$(Fig. 3). The propagation of residuals on both sides of zero indicated that there was no systemic error in model development (Heravi and Kyani, 2004).

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 HDAC inhibitory activitiesof the studied molecules is a direct function of the descriptors of their total chemical structure takes preference over the null hypothesis which states otherwise.

**Table 3: Comparison of actual $pIC_{50}$ and pred. $pIC_{50}$ of model 1**

| Compound | Experimental $pIC_{50}$ | Predicted $pIC_{50}$ | Residual |
|---|---|---|---|
| 2 | 1.62000000 | 1.30047000 | 0.31953000 |
| 3 | 1.42000000 | 1.40865000 | 0.01135000 |
| 5 | 1.31000000 | 1.51651600 | -0.20651600 |
| 6 | 1.39800000 | 1.53903100 | -0.14103100 |
| 8 | 1.00000000 | 0.78159500 | 0.21840500 |
| 9 | 0.85400000 | 0.99135200 | -0.13735200 |
| 11 | 0.72100000 | 0.64291400 | 0.07808600 |
| 12 | 0.72100000 | 0.41195200 | 0.30904800 |
| 14 | 0.63800000 | 0.59705000 | 0.04095000 |
| 15 | 0.52300000 | 0.37852000 | 0.14448000 |
| 17 | 0.46300000 | 0.19018900 | 0.27281100 |
| 18 | 0.45600000 | 0.87932100 | -0.42332100 |
| 20 | 0.39800000 | 0.33491000 | 0.06309000 |
| 21 | 0.38500000 | 0.48849900 | -0.10349900 |
| 23 | 0.09700000 | 0.10872100 | -0.01172100 |
| 24 | -0.02900000 | -0.24556900 | 0.21656900 |
| 25 | -0.59000000 | -0.37695400 | -0.21304600 |
| 27 | -0.84500000 | -0.75526500 | -0.08973500 |
| 28 | -0.53100000 | -0.18290000 | -0.34810000 |
| 29 | -1.54400000 | -1.54400000 | 0.00000000 |

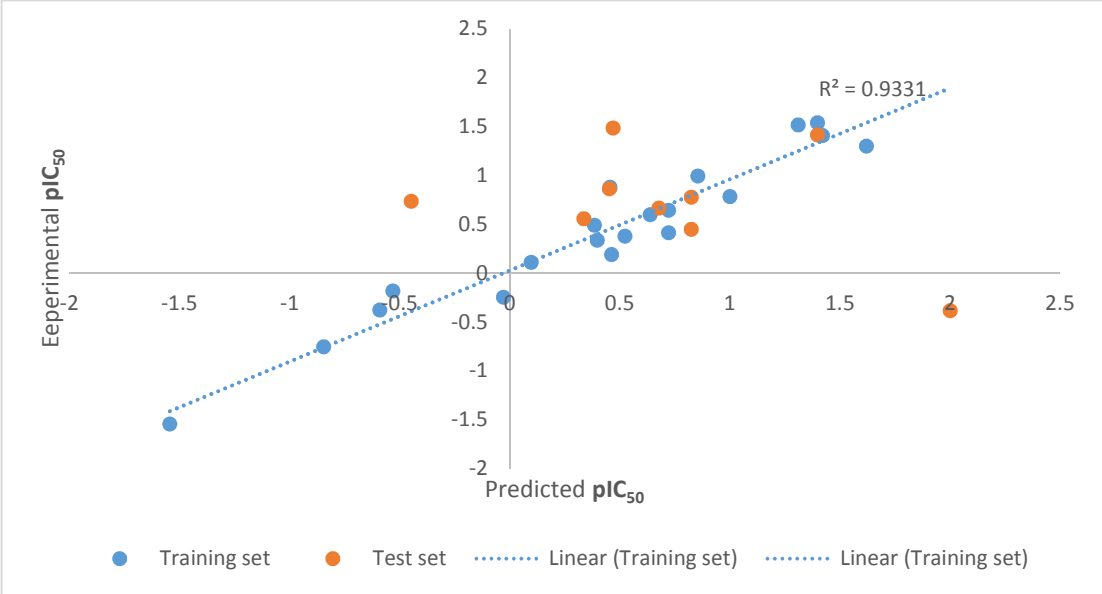



**Fig. 2: Plot of experimental $pIC_{50}$ against predicted $pIC_{50}$**


**Fig. 3: Plot of Standardized residual verses Experimental $pIC_{50}$ (Residual Plot)**

**Table 4: 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 | | |

**Table 5:External Validation of Model 1**

| Test set compounds | $[Ypred(te) - Yobs(te)]^2$ | $[Yobs(te) - Ym(tr)]^2$ |
|---|---|---|
| 1 | 0.000147 | 0.025921 |
| 2 | 0.002136 | 0.094249 |
| 3 | 0.168959 | 0.004225 |
| 4 | 0.04791 | 0.0324 |
| 5 | 0.000257 | 0.776161 |
| 6 | 0.141636 | 0.094249 |
| | **$\sum$ = 0.361045** | **$\sum$ = 1.027205** |

Using equation 2 above, $R^2_{pred.} = 1 - {}^{0.361045}/_{1.027205} = 0.6495$

Table 5 gives the external predictability of the model. Three compounds were remove from the external test set prior to the calculation of $R^2_{pred.}$ because they were discovered to be outliers. The high $R^2_{pred.}$ value of 0.6495 above the minimum threshold of 0.5 recommended for a standard QSAR model indicates that the model possesses robust external predictive ability.

The positive coefficient of the descriptors; FMF, Kier3, n5HeteroRing, globaltopo is an indication that they vary directly with the inhibitory activities of the molecules. Thus, the higher the values of these descriptors in the molecules, the higher their HDAC inhibitory bioactivities and vice versa. In a sharp contrast to other descriptors in the optimization model, Kier1 descriptor varies inversely with HDAC inhibitory activities of the molecules. The implication is; for an enhanced HDAC inhibitory activities of the molecules, the value of this descriptor should be as low as possible.

## 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] J Ferlay;HR Shin;F Bray;D Forman;CD Mathers;D Parkin, retrieved from: http://globocan.iarc.fr.**2008**.
[2] Cancer Facts and Figures, **2016**. Corporate Center: American Cancer Society Inc. 250 Williams Street, NW, Atlanta, GA 30303-1002 404-320-3333.
[3] M Mottamal; S Zheng;LT Huang; G Wang,*Molecules*,**2015**, 20, 3898-3941
[4] R Sabet;M Mohammadpour;A Sadeghi;A Fassihi,*Eur. J. Med. Chem.***2010**, 45, 1113–1118
[5] OP Sanja;J Lidija;JStrahinja; Z Kovačevića; ND Kalajdžijaa,*APTEFF*, **2012**, 43, 1-342.
[6] S Thangapandian; S John; KW Lee,*Inter. Bio. Central*, **2011**, 3, 1-11
[7] W Wu;C Zhang; W Lin;Q Chen;X Guo; Y Qian,*PLoSONE*, **2015**, 10(3), doi: 10.1371/journal.pone.0119575.
[8] KF Khaled;NS Abdel-Shafi,*Inter. J. of Electrochemical Sci.*, **2011**, 6, 4077 – 4094.
[9] V Ravichandran;R Harish;J Abhishek;S Shalini; PV Christapher; KA Ram,*Inter.J. of Drug Des. and Discov.*, **2011**, 2: 511-519.