



Scholars Research Library  
(<http://scholarsresearchlibrary.com/archive.html>)



ISSN : 2231- 3176  
CODEN (USA): JCMMDA

## Quantitative structure activity relationship study on the inhibitory activity of Schiff bases against *Escherichia coli* (*E.coli*)

John Philip Ameji<sup>1\*</sup>, Onoyima Christian Chinweuba<sup>2</sup> and Olusupo Sabitu B.<sup>3</sup>

<sup>1</sup>Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria

<sup>2</sup>Department of Chemistry, Nigeria Police Academy Wudil, Kano State Nigeria

<sup>3</sup>Department of Chemistry, Kano State University of Science and Technology, Wudil, Kano Nigeria

### ABSTRACT

The emergence of multi-drug resistant strain of *Escherichia coli* (*E.coli*) has necessitated the exploration and development of newer structural moiety of Schiff bases as anti- *E. coli* agents owing to their enormous inhibitory activity against this bacterium. In this present study, a Genetic function approximation (GFA) QSAR analysis of some selected Schiff bases with anti- *E. coli* activity was performed using OD, 1D, 2D and 3D descriptors resulting in the generation of three statistically significant models from which an octa-parametric model was selected as the most robust model with  $R^2 = 0.9622$ ,  $R^2_{adj} = 0.9470$ ,  $Q^2 = 0.9132$ ,  $R^2 - Q^2 = 0.049$ ,  $R^2 - R_0^2 / R^2 = 0.00$ ,  $R^2 - R_0^2 / R^2 = 0.001$ ,  $K = 1$ ,  $K' = 0.9622$ . The optimization model hinted the predominance of the size descriptor ETA-Eta-P-F-L (Local functionality contribution EtaF\_local relative to molecular size) in influencing the observed anti-*E. coli* activity of Schiff bases. It is envisaged that the QSAR results identified in this study will offer important structural insight into designing novel anti-*E. coli* drugs from Schiff bases.

**Keywords:** *Escherichia coli*, QSAR, descriptors, Schiff bases, MIC.

### INTRODUCTION

The role of antimicrobial drugs in decreasing illness and death associated with infectious diseases in animals and humans cannot be overemphasized. However, selective pressure exerted on existing antimicrobial drug has orchestrated the emergence and spread of drug-resistance traits among disease causing and commensal bacteria [1]. Of serious concern is the development of resistance by *E.coli* strains to the current antibiotics such as ampicillin, sulfonamide, gentamicin, streptomycin, ciprofloxacin, trimethoprim, amoxicillin [2, 3]. *E.coli* is often times a commensal bacterium of humans and animals but Pathogenic variants cause intestinal and extraintestinal infections, including gastroenteritis, urinary tract infection, meningitis, peritonitis, and septicemia [4, 5]. This trend of resistance exhibited by this organism poses serious threat to human and animals' health, necessitating the search for newer antibiotics.

In recent years, Schiff bases have received considerable attention because of their physiological and pharmacological activities [6]. This class of organic compounds have also demonstrated significant inhibitory activity against the growth of *E. coli* [7-10] making them potential drug candidate in man's quest to curb the dangerous trend of multi-drug resistance posed by this pathogenic micro-organism.

Conventional drug discovery and development is characterized by trial and error approach. This is time consuming, costly due to the enormous expense of failures of candidate drugs late in their development and a threat to green chemistry due to enormous waste released into the environment. QSAR offer important structural insight in the design of novel anti-microbial drugs by exploring and harnessing the structural requirements controlling the observed anti-microbial activities as well as providing predictive model for bio-activities of potential drug candidates, reducing the requirement for lengthy, costly and hazardous laboratory test. QSAR is based on the conception that there exists a close relationship between bulk properties of compounds and their molecular structure. Thus, it is the basic tenet of chemistry to identify these assumed relationships and then to quantify them allowing a clear connection between the macroscopic and the microscopic properties of matter [11].

The aim of this work is to build a statistically robust, predictive and rational Genetic function approximation (GFA) based QSAR model for inhibitory activity of Schiff bases against *E. coli*.

### MATERIALS AND METHODS

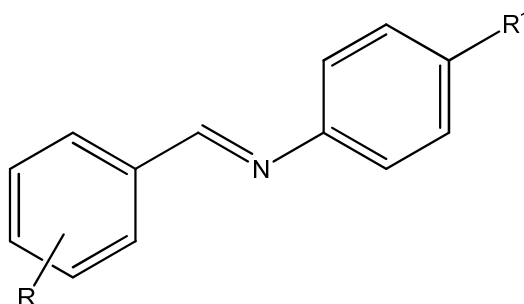
The molecular modeling studies were performed using the molecular modeling program SPARTAN'14 V1.1.0 ([www.wavefun.com](http://www.wavefun.com)). The "Cascade method" of molecular optimization was invoked for this study [12]. Here, the molecules were first pre-optimized with the molecular mechanics procedure included in Spartan'14 V1.1.0 software and the resulting geometries were further refined by means of a semi-empirical method (PM3). The choice of this method is anchored on the fact that it makes calculations less computationally taxing by relegating initial geometry calculations to less computationally intensive (and possibly more inaccurate) methods. The lowest energy structure was used for each molecule to calculate their physicochemical properties (molecular descriptors).

#### Data set

A data set comprising of series of 41 schiff bases with inhibitory activities expressed as minimum inhibitory concentration (MIC) against *Escherichia coli* was taken from reported articles [7-10] for this study. 70% of the data set (29 compounds) was used as training set for building the models while the remaining 30% was used as test set for external validation of the most statistically significant QSAR model. The notation, structure, MIC and pMIC of the compounds are shown in Table 1 below.

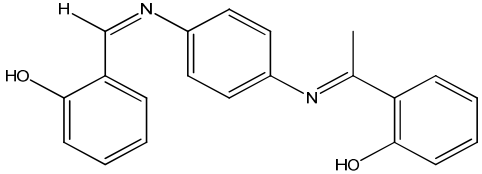
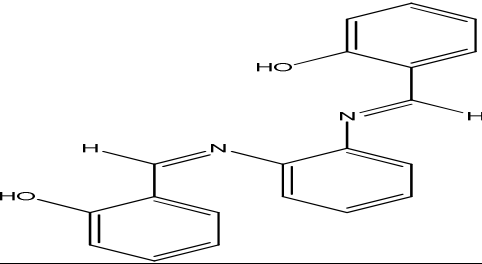
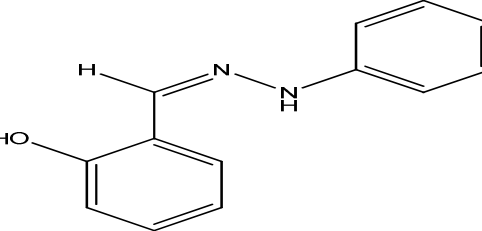
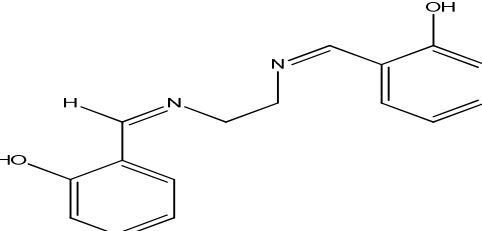
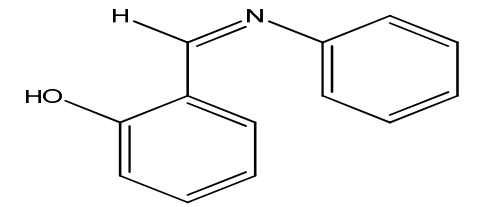
Figure 1, 2, and 3 gives the parent structures of the Schiff bases with their experimentally determined anti-*Escherichia coli* inhibitory activity in form of Minimum inhibitory activity (MIC). R, R<sup>1</sup>, R<sup>2</sup> and R<sup>3</sup> are the substituents to the parent structures.

**Table 1: Anti-S.typhi activity of the compounds (MIC µg/ml and pMIC)**



**Figure 1: parent structure for compound 1-8**

| Compound label | R                         | R <sup>1</sup>    | MIC | pMIC  |
|----------------|---------------------------|-------------------|-----|-------|
| C1             | 3-OCH <sub>3</sub>        | 4-CH <sub>3</sub> | 22  | 1.342 |
| C2             | 3,4-OCH <sub>3</sub>      | 4-CH <sub>3</sub> | 18  | 1.556 |
| C3             | 3,4,5-OCH <sub>3</sub>    | 4-CH <sub>3</sub> | 36  | 1.079 |
| C4             | 3-OCH <sub>3</sub> , 4-OH | 4-CH <sub>3</sub> | 12  | 1.806 |
| C5             | 4-F                       | 4-CH <sub>3</sub> | 64  | 1.806 |
| C6             | 4-Cl                      | 4-CH <sub>3</sub> | 26  | 2.093 |
| C7             | 4-Br                      | 4-CH <sub>3</sub> | 64  | 1.342 |
| C8             | 4-I                       | 4-CH <sub>3</sub> | 124 | 1.556 |

|     |  |  |     |       |
|-----|--|--|-----|-------|
| C12 |    |  | 200 | 2.301 |
| C13 |    |  | 200 | 2.301 |
| C14 |   |  | 250 | 2.398 |
| C15 |  |  | 100 | 2.000 |
| C16 |  |  | 200 | 1.362 |

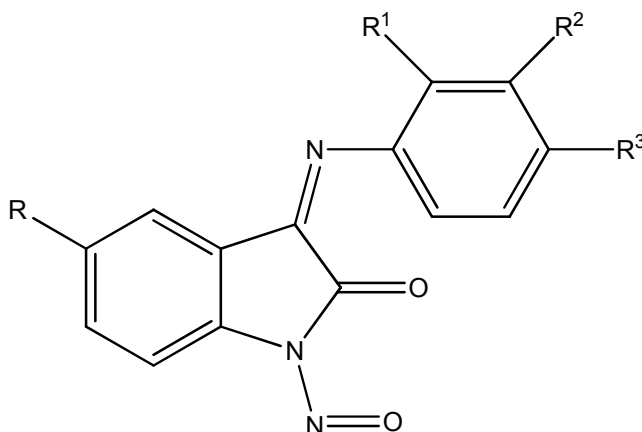


Figure 2: parent structure for compound 17-32

| Compound label | R  | R <sup>1</sup>  | R <sup>2</sup> | R <sup>3</sup>   | MIC | pMIC  |
|----------------|----|-----------------|----------------|------------------|-----|-------|
| C17            | H  | H               | H              | Cl               | 23  | 1.362 |
| C18            | H  | H               | H              | Br               | 21  | 1.322 |
| C19            | H  | H               | H              | F                | 24  | 1.38  |
| C20            | H  | H               | Cl             | F                | 42  | 1.623 |
| C21            | H  | H               | H              | CH <sub>3</sub>  | 20  | 1.301 |
| C22            | H  | H               | H              | OCH <sub>3</sub> | 39  | 1.591 |
| C23            | H  | H               | H              | NO <sub>2</sub>  | 22  | 1.342 |
| C24            | H  | NO <sub>2</sub> | H              | NO <sub>2</sub>  | 16  | 1.204 |
| C25            | Br | H               | H              | Cl               | 40  | 1.602 |
| C26            | Br | H               | H              | Br               | 18  | 1.255 |
| C27            | Br | H               | H              | F                | 22  | 1.342 |
| C28            | Br | H               | Cl             | F                | 20  | 1.301 |
| C29            | Br | H               | H              | CH <sub>3</sub>  | 21  | 1.322 |
| C30            | Br | H               | H              | OCH <sub>3</sub> | 24  | 1.38  |
| C31            | Br | H               | H              | NO <sub>2</sub>  | 17  | 1.23  |
| C32            | Br | NO <sub>2</sub> | H              | NO <sub>2</sub>  | 23  | 1.362 |

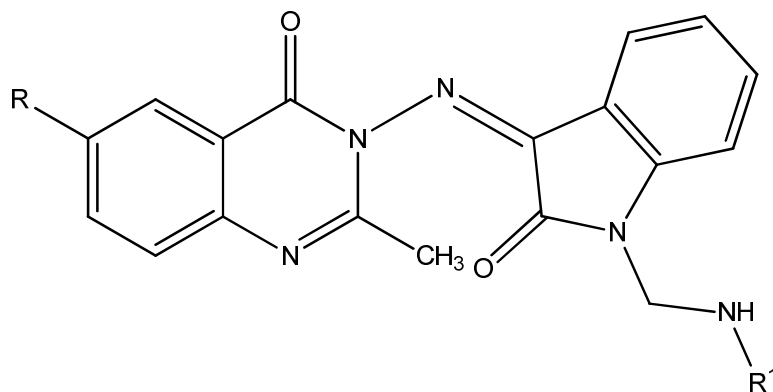
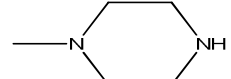
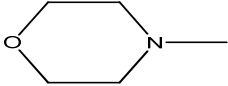
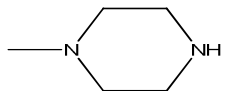
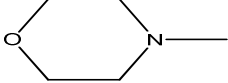


Figure 3: parent structure for compound 33-44

|     | R | R <sup>1</sup>  | MIC(μg/ml) | pMIC  |
|-----|---|---|------------|-------|
| C33 | H | N(CH <sub>3</sub> ) <sub>2</sub>  | 20         | 1.301 |
| C34 | H | N(C <sub>2</sub> H <sub>5</sub> ) <sub>2</sub>                                      | 16         | 1.204 |
| C35 | H | N(C <sub>6</sub> H <sub>5</sub> ) <sub>2</sub>                                      | 20         | 1.301 |
| C36 | H | N(C <sub>6</sub> H <sub>11</sub> ) <sub>2</sub>                                     | 19         | 1.279 |
| C37 | H |  | 22         | 1.342 |

|     |    |   |    |       |
|-----|----|---|----|-------|
| C38 | H  |  | 26 | 1.415 |
| C39 | Br | N(CH <sub>3</sub> ) <sub>2</sub>  | 16 | 1.204 |
| C40 | Br | N(C <sub>2</sub> H <sub>5</sub> ) <sub>2</sub>                                    | 18 | 1.255 |
| C41 | Br | N(C <sub>6</sub> H <sub>5</sub> ) <sub>2</sub>                                    | 17 | 1.23  |
| C42 | Br | N(C <sub>6</sub> H <sub>11</sub> ) <sub>2</sub>                                   | 19 | 1.279 |
| C43 | Br |  | 21 | 1.322 |
| C44 | Br |  | 27 | 1.431 |

### Model building

The computed descriptors were subjected to regression analysis with the experimentally determined minimum inhibitory concentration on logarithmic scale (pMIC) as the dependent variable and the selected descriptors as the independent variables using Genetic function approximation (GFA) method in Material studio software. To develop the optimization model, 29 samples were included in the training set. The number of descriptors in the regression equation was set to 5, and Population and Generation were set to 1,000 and 5,000, respectively. The number of top equations returned was 5. Mutation probability was 0.1, and the smoothing parameter was 0.5. The statistical significance of the generated models were assessed based on Friedman's LOF and the optimum model was selected based on this parameter.

### Model Validation

A reliable validation procedure is required in order to confirm the existence of chance correlations as well as ascertaining the fitting ability, stability, reliability and predictive ability of the developed models. The optimum model was validated and its validation parameters were compared with the standards shown in Table 2 below.

**Table 2: Validation metrics for a generally acceptable QSAR model**

| S/n | Metric symbol   | Name   | Threshold             |
|-----|---|--|-----------------------|
| 1   | R <sup>2</sup>  | Coefficient of determination                         | ≥ 0.6                 |
| 2   | Q <sup>2</sup>  | LOO cross validation coefficient                     | > 0.5                 |
| 3   | R <sup>2</sup> <sub>pred.</sub>                               | External test set's coefficient of determination     | ≥ 0.6                 |
| 4   | R <sup>2</sup> - Q <sup>2</sup>                               | Difference between R <sup>2</sup> and Q <sup>2</sup> | ≤ 0.3                 |
| 5   | F value   | Variation ratio                                      | High                  |
| 6   | R <sup>2</sup> - R <sub>0</sub> <sup>2</sup> / R <sup>2</sup> | Golbraikh and Tropsha condition                      | < 0.1                 |
| 7   | R <sup>2</sup> - R <sub>0</sub> <sup>2</sup> / R <sup>2</sup> | Golbraikh and Tropsha condition                      | < 0.1                 |
| 8   | K and K'  | Intercept  | 0.85 ≤ k or k' ≤ 1.15 |

Source: Roy *et al.*; Ravinchandran *et al.*; Golbraikh and Tropsha [13, 14, 15]

### Internal validation parameters

This validation is done using the data that created the model. The various internal validation parameters invoked in this study are presented thus;

**R<sup>2</sup> (the square of the correlation coefficient):** describes the fraction of the total variation attributed to the model. The closer the value of R<sup>2</sup> is to 1.0, the better the regression equation explains the Y variable. R<sup>2</sup> is the most commonly used internal validation indicator and is expressed as follows:

$$R^2 = 1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - Y_{training})^2} \quad (1)$$

Where, Y<sub>obs</sub>; Y<sub>pred</sub>; Y<sub>training</sub> are the experimental property, the predicted property and the mean experimental property of the samples in the training set, respectively [16].

**Adjusted R<sup>2</sup> (R<sup>2</sup><sub>adj</sub>):** R<sup>2</sup> value varies directly with the increase in number of regressors i.e. descriptors, thus, R<sup>2</sup> cannot be a useful measure for the goodness of model fit. Therefore, R<sup>2</sup> is adjusted for the number of explanatory variables in the model. The adjusted R<sup>2</sup> is defined as:

$$R^2_{adj} = 1 - (1 - R^2) \frac{n-1}{n-p-1} = \frac{(n-1)R^2 - p}{n-p+1} \quad (2)$$

Where p = number of independent variables in the model [17]

**Q<sup>2</sup> (Leave one out cross validation coefficient):** The LOO cross validated coefficient (Q<sup>2</sup>) is given by;

$$Q^2 = 1 - \frac{\sum(Y_p - Y)^2}{\sum(Y - Y_m)^2} \quad (3)$$

Where Y<sub>p</sub> and Y represent the predicted and observed activity respectively of the training set and Y<sub>m</sub> the mean activity value of the training set [17].

**Variance Ratio (F):** this parameter is used to judge the overall significance of the regression coefficient. It is the ratio of regression mean square to deviations mean square defined as:

$$F = \frac{\frac{\sum(Y_{cal} - Y_m)^2}{p}}{\frac{\sum(Y_{obs} - Y_{cal})^2}{N-p-1}} \quad (4)$$

Where Y<sub>obs</sub> stands for the observed response value, while Y<sub>calc</sub> is the model-derived calculated response and Y<sub>m</sub> is the average of the observed response values. The F value has two degrees of freedom: p, N – p – 1. The computed F value of a model should be significant at p < 0.05. A high F value is an indication that the regression coefficients are significant [13].

**Standard error of estimate (s):** Low standard error of estimate is an indication of a good model. It is defined as follows:

$$S = \sqrt{\frac{\sum(Y_{obs} - Y_{cal})^2}{N-p-1}} \quad (5)$$

Its degree of freedom is N-p-1 [13].

**Leave one out cross validation (LOOCV):** in this cross validation approach, the model is repeatedly refit leaving out a single observation and then used to derive a prediction for the left-out observation. For the model to have an excellent prediction ability, Q<sup>2</sup> must be > 0.5 and R<sup>2</sup> – Q<sup>2</sup> value should not exceed 0.3. The equation for CV is:

$$Q^2 = 1 - \frac{PRESS}{\sum(Y_i - Y_m)^2} \quad (6)$$

$$PRESS = \sum(Y_{pred, i} - Y_i) \quad (7)$$

Q<sup>2</sup> = LOOCV cross validation coefficient, R<sup>2</sup> = coefficient of determination.

Y<sub>i</sub> is the data value(s) not used to construct the CV model, PRESS is the predictive residual sum of the squares, Y<sub>m</sub> = mean of the experimental bioactivity (pMIC), Y<sub>pred, i</sub> is the predicted Y<sub>i</sub> [14].

#### Metrics for external validation

External validation of QSAR model is performed in order to ensure the predictability and applicability of the developed QSAR model for the prediction of untested molecules. The various external validation metrics used in this work are highlighted thus:

**Predictive R<sup>2</sup> (R<sup>2</sup><sub>pred</sub>):** R<sup>2</sup><sub>pred</sub> is termed the predictive R<sup>2</sup> of a development model and is an important parameter that is used to test the external predictive ability of a QSAR model. The predicted R<sup>2</sup> value is calculated as follows;

$$R_{\text{pred}}^2 = 1 - \frac{\sum [Y_{\text{obs}}(\text{test}) - Y_{\text{pred}}(\text{test})]^2}{\sum [Y_{\text{obs}}(\text{test}) - Y_{\text{m}}(\text{training})]^2} \quad (8)$$

$Y_{\text{pred}}(\text{test})$  and  $Y_{\text{obs}}(\text{test})$  indicate predicted and observed activity values respectively of the test set compounds and  $Y_{\text{m}}(\text{training})$  indicates mean activity value of the training set [14].

**Golbraikh and Tropsha's criteria:** according to Golbraikh and Tropsha, models are considered satisfactory, if all the following conditions are met.

- (a)  $R_{\text{test}}^2 > 0.5$
- (b)  $(R^2 - R_0^2 / R^2) < 0.1$
- (c)  $(R^2 - R_0'^2 / R^2) < 0.1$
- (d)  $0.85 \leq k \leq 1.15$
- (e)  $0.85 \leq k' \leq 1.15$

Parameters  $R^2$  and  $R_0^2$  are the squared correlation coefficients between the observed and predicted values of the compounds with and without intercept, respectively. The parameter  $R_0'^2$  bears the same meaning with  $R_0^2$  but uses the reversed axes.  $K$  is the intercept of the plot of the observed and predicted values of the compounds and  $K'$  the reversed axes intercept [15].

## RESULTS AND DISCUSSION

The equations are the QSAR mathematical models while the terms are the molecular descriptors. Detailed definition of the descriptors that appeared in the models are given Table 5.

**Table 3: GFA derived QSAR models for the  $K_{\text{OA}}$  of the selected POPs**

| Model | Equation  | Definition of terms  |
|-------|---|--|
| 1.    | $pMIC = 0.028330211 * X7$ $+ 0.754079982 * X32$ $- 5.518138173 * X57$ $+ 47.796805665 * X62$ $+ 0.494140811 * X73$ $- 0.146023548 * X144$ $- 0.000885791 * X149$ $- 0.069505442 * X195$ $- 8.198314822$                       | X7 : ATSm5<br>X32 : VP-3<br>X57 : ETA_EtaP_F<br>X62 : ETA_EtaP_F_L<br>X73 : nHBDon<br>X144 : RNCS<br>X149 : GRAV-1<br>X195 : Wlambda2.volume                               |
| 2.    | $pMIC = 0.034301705 * X7$ $- 6.239202363 * X57$ $+ 36.506138720 * X62$ $+ 0.564599476 * X73$ $+ 0.025662395 * X80$ $- 0.158090685 * X144$ $- 0.075859857 * X195$ $+ 0.026847150 * X208$ $- 4.155706232$                       | X7 : ATSm5<br>X57 : ETA_EtaP_F<br>X62 : ETA_EtaP_F_L<br>X73 : nHBDon<br>X80 : nAtomP<br>X144 : RNCS<br>X195 : Wlambda2.volume<br>X208 : Wlambda2.eneg                      |
| 3.    | $pMIC = 0.030272104 * X7$ $+ 0.019990735 * X47$ $- 6.413499672 * X57$ $+ 33.104729852 * X62$ $+ 0.538824561 * X73$ $+ 0.030292997 * X80$ $- 0.146988859 * X144$ $- 0.083336920 * X195$ $+ 0.025485956 * X208$ $- 3.293980187$ | X7 : ATSm5<br>X47 : ETA_Beta_ns<br>X57 : ETA_EtaP_F<br>X62 : ETA_EtaP_F_L<br>X73 : nHBDon<br>X80 : nAtomP<br>X144 : RNCS<br>X195 : Wlambda2.volume<br>X208 : Wlambda2.eneg |

**Table 4: Validation Parameters of the models**

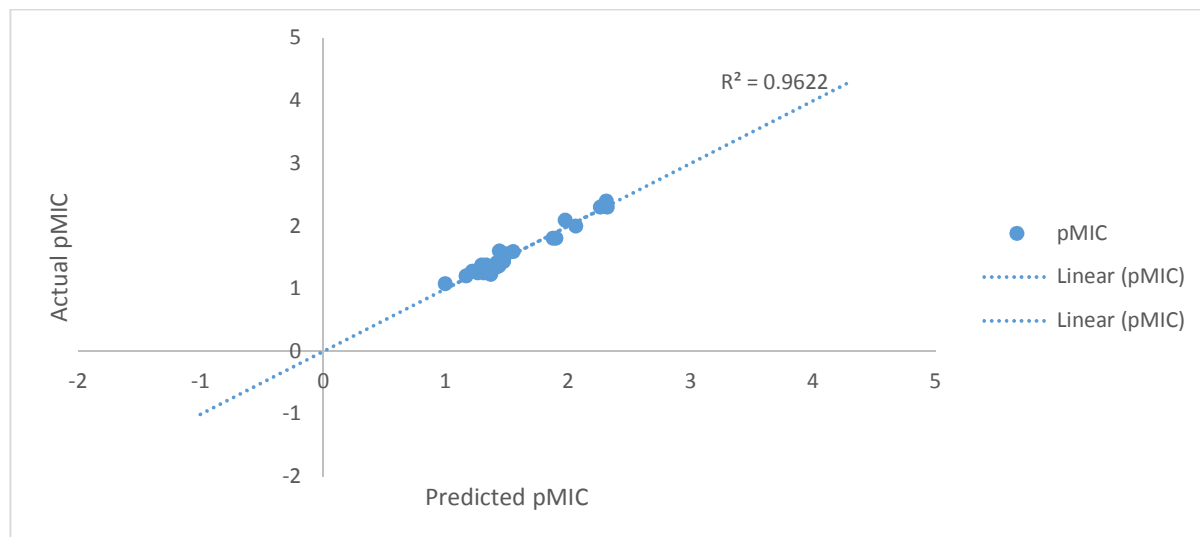
| S/n | Parameters                                    | Model 1     | Model 2     | model 3     |
|-----|---|-------------|-------------|-------------|
| 1   | Friedman LOF                                  | 0.02221800  | 0.02241300  | 0.02297500  |
| 2   | R-squared                                     | 0.96216800  | 0.96183600  | 0.97122700  |
| 3   | Adjusted R-squared                            | 0.94703500  | 0.94657000  | 0.95759800  |
| 4   | Cross validated R-squared                     | 0.91317900  | 0.91730900  | 0.94418000  |
| 5   | Significant Regression                        | Yes         | Yes         | Yes         |
| 6   | Significance-of-regression F-value            | 63.58191400 | 63.00667100 | 71.26126300 |
| 7   | Critical SOR F-value (95%)                    | 2.45092100  | 2.45092100  | 2.42509900  |
| 8   | Replicate points                              | 0           | 0           | 0           |
| 9   | Computed experimental error                   | 0.00000000  | 0.00000000  | 0.00000000  |
| 10  | Min expt. error for non-significant LOF (95%) | 0.06688400  | 0.06717700  | 0.05953900  |

The GFA algorithm makes use of a population of many models rather than generating a single model. The models are scored using Friedman's "lack of fit" (LOF) measure as the evaluation function [18, 19]. Thus, based on model with the least LOF score, model 1 is selected as the optimization model for predicting the MIC of Schiff bases.

The symbols and detailed definition of the descriptors that appeared in the models depicted in Table 3 are given in Table 5 below.

**Table 5: Detailed definition of descriptors**

| S/n | Descriptor symbol | Definition   |
|-----|-------------------|--|
| 1   | ATSm5             | ATS autocorrelation descriptor, weighted by scaled atomic mass             |
| 2   | VP-3              | Valence path, order 3  |
| 3   | ETA_EtaP_F        | Functionality index EtaF relative to molecular size                        |
| 4   | ETA_EtaP_F_L      | Local functionality contribution EtaF_local relative to molecular size     |
| 5   | ETA_Beta_ns       | A measure of electron-richness of the molecule                             |
| 6   | nHBDon            | Number of hydrogen bond donors.  |
| 7   | RNCS              | Relative negative charge surface area -- most negative surface area * RNCG |
| 8   | GRAV-1            | Gravitational index of heavy atoms   |
| 9   | Wlambda2.volume   | Directional WHIM, weighted by van der Waals volumes                        |
| 10  | Wlambda2.eneg     | Directional WHIM, weighted by Mulliken atomic electronegativities          |
| 11  | nAtomP            | Number of atoms in the largest pi system                                   |

**Figure 4: Plot of actual pMIC against predicted pMIC**



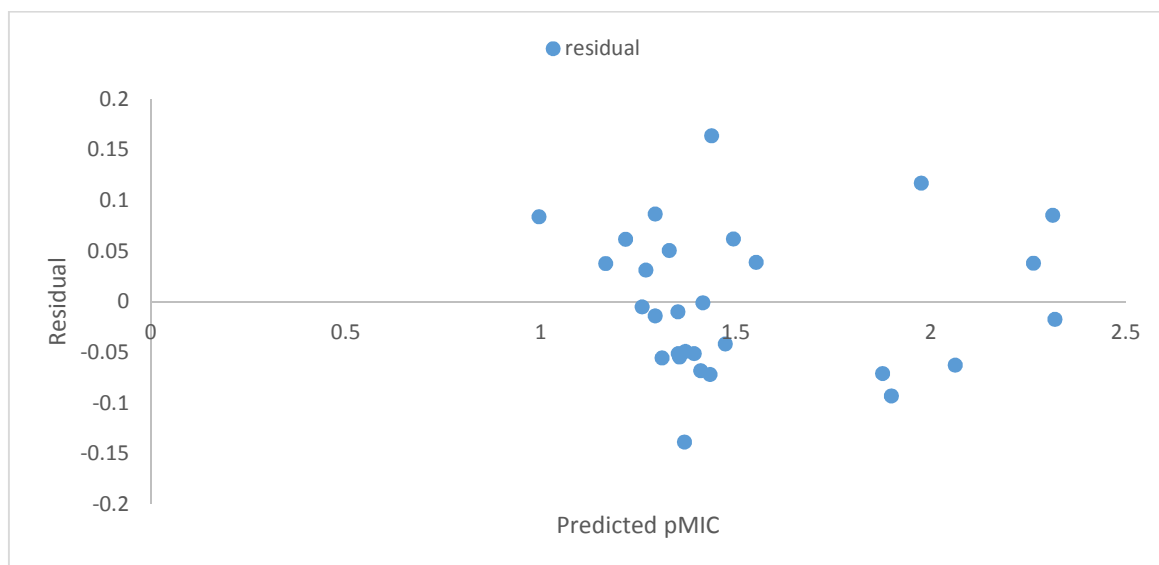


Figure 5: Residual plot of model 1

Table 6 gives the predicted pMIC ( $Y_{pred}$ ) of the anti-*E. coli* Schiff bases used as external test set for model validation obtained by substituting the values of the descriptors in column 3,4,5,6,7,8,9, and 10 into the optimum QSAR model (model 1). Column 1 gives the experimentally determined pMIC ( $Y_{obs}$ ) values of the test set compounds;  $Y_m$  is the mean pMIC of the training set compounds.

Table 6: External validation of Model 1

| Test cpd | $Y_{obs}$ | ATSm5  | VP-3     | ETA_EtaP_F | ETA_EtaP_F_L | nHBDOn | RNCS     |
|----------|-----------|--------|----------|------------|--------------|--------|----------|
| C2       | 1.255     | 22.159 | 2.80426  | 0.89724    | 0.29502      | 0      | 5.734714 |
| C6       | 1.415     | 14.450 | 2.518645 | 0.8326     | 0.28839      | 0      | 3.793112 |
| C17      | 1.362     | 33.242 | 3.183458 | 1.04066    | 0.31         | 0      | 5.834262 |
| C20      | 1.623     | 38.266 | 3.297687 | 1.08743    | 0.31395      | 0      | 5.097443 |
| C24      | 1.204     | 48.912 | 3.562283 | 1.3036     | 0.33392      | 0      | 3.733877 |
| C29      | 1.322     | 44.595 | 3.686629 | 0.99604    | 0.2931       | 0      | 5.701515 |
| C32      | 1.362     | 62.217 | 4.110085 | 1.28505    | 0.3157       | 0      | 4.588109 |
| C35      | 1.301     | 70.502 | 6.431879 | 1.28689    | 0.30113      | 0      | 2.149159 |
| C37      | 1.342     | 55.530 | 5.582728 | 1.04743    | 0.23984      | 2      | 1.436629 |
| C39      | 1.204     | 65.475 | 5.106156 | 1.07431    | 0.26108      | 1      | 1.547756 |
| C41      | 1.23      | 84.472 | 6.976827 | 1.25252    | 0.25711      | 1      | 0.674175 |
| C43      | 1.322     | 68.835 | 6.13053  | 1.03114    | 0.22901      | 2      | 1.416237 |

| Test set | GRAV-1   | Wlambda2.volume | $Y_{pred}$ | $Y_m$ | $(Y_{obs}-Y_{pred})^2$ | $(Y_{obs}-Y_m)^2$  |
|----------|----------|-----------------|------------|-------|------------------------|--------------------|
| C2       | 1.70E+03 | 1.908207        | 1.22       | 1.53  | 0.001445               | 0.073441           |
| C6       | 1.45E+03 | 1.205697        | 1.38       | 1.53  | 0.001262               | 0.012321           |
| C17      | 2.10E+03 | 0.06367         | 1.50       | 1.53  | 0.019272               | 0.026896           |
| C20      | 2.23E+03 | 2.347013        | 1.49       | 1.53  | 0.016399               | 0.009409           |
| C24      | 2.69E+03 | 1.980924        | 1.58       | 1.53  | 0.139091               | 0.103684           |
| C29      | 2.29E+03 | 2.650138        | 1.31       | 1.53  | 5.11E-05               | 0.041616           |
| C32      | 2.95E+03 | 0.360036        | 1.35       | 1.53  | 0.000116               | 0.026896           |
| C35      | 3.64E+03 | 15.41845        | 1.33       | 1.53  | 0.000997               | 0.050625           |
| C37      | 3.03E+03 | 1.218562        | 1.28       | 1.53  | 0.003887               | 0.033856           |
| C39      | 3.01E+03 | 5.518675        | 1.27       | 1.53  | 0.004778               | 0.103684           |
| C41      | 4.01E+03 | 6.505369        | 1.22       | 1.53  | 0.000132               | 0.087616           |
| C43      | 3.29E+03 | 1.377386        | 1.40       | 1.53  | 0.006037               | 0.041616           |
|          |          |                 |            |       | $\Sigma = 0.193466$    | $\Sigma = 0.61166$ |

The predicted  $R^2$  value for the test set compounds is calculated using the formulae in equation 8.

$$\text{Thus, } R^2_{\text{pred.}} = 1 - \left( \frac{0.193466}{0.61166} \right) = 0.6837$$

**Table 7: Golbraikh and Tropsha external validation parameters for model 1**

| s/n | Parameter | Value  |
|-----|-----------|--------|
| 1   | $r^2$     | 0.9622 |
| 2   | $r_0'^2$  | 0.9608 |
| 3   | $r_0^2$   | 0.9622 |
| 4   | k         | 1      |
| 5   | K         | 0.9622 |

Based on the parameters above;

$$R^2 - R_0'^2 / R^2 = \frac{0.9622 - 0.9622}{0.9622} = 0.000$$

$$R^2 - R_0^2 / R^2 = \frac{0.9622 - 0.9608}{0.9622} = 0.001$$

**Table 8: Comparison of Yobs (training) and Ypred.(training) of model 1**

| Name | Y <sub>obs</sub> | Y <sub>pred</sub> | Residual |
|------|------------------|-------------------|----------|
| C1   | 1.342            | 1.39333           | -0.05133 |
| C3   | 1.556            | 1.494089          | 0.061911 |
| C4   | 1.079            | 0.995213          | 0.083787 |
| C5   | 1.806            | 1.899129          | -0.09313 |
| C7   | 1.806            | 1.876937          | -0.07094 |
| C8   | 2.093            | 1.975795          | 0.117205 |
| C12  | 2.301            | 2.263063          | 0.037937 |
| C13  | 2.301            | 2.318343          | -0.01734 |
| C14  | 2.398            | 2.312507          | 0.085493 |
| C15  | 2                | 2.062778          | -0.06278 |
| C16  | 1.362            | 1.433831          | -0.07183 |
| C18  | 1.322            | 1.37101           | -0.04901 |
| C19  | 1.38             | 1.293494          | 0.086506 |
| C21  | 1.301            | 1.352256          | -0.05126 |
| C22  | 1.591            | 1.551997          | 0.039003 |
| C23  | 1.342            | 1.410229          | -0.06823 |
| C25  | 1.602            | 1.438125          | 0.163875 |
| C26  | 1.255            | 1.260037          | -0.00504 |
| C27  | 1.342            | 1.352068          | -0.01007 |
| C28  | 1.301            | 1.355503          | -0.0545  |
| C30  | 1.38             | 1.329319          | 0.050681 |
| C31  | 1.23             | 1.368738          | -0.13874 |
| C33  | 1.301            | 1.269841          | 0.031159 |
| C34  | 1.204            | 1.166293          | 0.037707 |
| C36  | 1.279            | 1.217333          | 0.061667 |
| C38  | 1.415            | 1.416018          | -0.00102 |
| C40  | 1.255            | 1.310682          | -0.05568 |
| C42  | 1.279            | 1.293169          | -0.01417 |
| C44  | 1.431            | 1.472874          | -0.04187 |

#### **Euclidean based applicability domain for the optimum QSAR models:**

Applicability domain (AD) is the physicochemical, structural or biological space, knowledge or information on which the training set of the model has been developed. The resulting model can be reliably applicable for only those compounds which are inside this domain. It is based on distance scores calculated by the Euclidean distance norms. At first, 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 [20].

The Euclidean based applicability domain for test and training set compounds are shown in Tables 9 and 10 respectively.

**Table 9: Euclidean based applicability domain for test set compounds**

| Test cpd. | Distance Score | Mean Distance | Normalized Mean Distance |
|-----------|----------------|---------------|--------------------------|
| C2        | 19999.12       | 689.625       | 0.172                    |
| C6        | 24832.34       | 856.288       | 0.325                    |
| C17       | 15298.04       | 527.519       | 0.024                    |
| C20       | 14590.55       | 503.123       | 0.001                    |
| C24       | 18687.86       | 644.409       | 0.131                    |
| C29       | 14574.45       | 502.567       | 0.001                    |
| C32       | 23492.69       | 810.093       | 0.283                    |
| C35       | 39427.3        | 1359.562      | 0.787                    |
| C37       | 25008.91       | 862.376       | 0.331                    |
| C39       | 24635.35       | 849.495       | 0.319                    |
| C41       | 49658.22       | 1712.353      | 1.11                     |
| C43       | 30733.76       | 1059.785      | 0.512                    |

**Table 10: Euclidean based applicability domain for training set compounds**

| Training set cpd | Distance Score | Mean Distance | Normalized Mean Distance |
|------------------|----------------|---------------|--------------------------|
| C1               | 23621.15       | 814.522       | 0.287                    |
| C3               | 17317.72       | 597.163       | 0.087                    |
| C4               | 21392.38       | 737.668       | 0.216                    |
| C5               | 25083.4        | 864.945       | 0.333                    |
| C7               | 22150.96       | 763.826       | 0.24                     |
| C8               | 21050.16       | 725.868       | 0.205                    |
| C12              | 14551.67       | 501.782       | 0                        |
| C13              | 14636.81       | 504.717       | 0.003                    |
| C14              | 24082.77       | 830.44        | 0.301                    |
| C15              | 19400.24       | 668.974       | 0.153                    |
| C16              | 27512.74       | 948.715       | 0.41                     |
| C18              | 14601.45       | 503.498       | 0.002                    |
| C19              | 15367.83       | 529.925       | 0.026                    |
| C21              | 15997.68       | 551.644       | 0.046                    |
| C22              | 14878.55       | 513.054       | 0.01                     |
| C23              | 14600.75       | 503.474       | 0.002                    |
| C25              | 14741.56       | 508.33        | 0.006                    |
| C26              | 15793.31       | 544.597       | 0.039                    |
| C27              | 14691.04       | 506.588       | 0.004                    |
| C28              | 15794.43       | 544.636       | 0.039                    |
| C30              | 15159.66       | 522.747       | 0.019                    |
| C31              | 17089.16       | 589.281       | 0.08                     |
| C33              | 17839.9        | 615.169       | 0.104                    |
| C34              | 19880.23       | 685.525       | 0.169                    |
| C36              | 39149.11       | 1349.969      | 0.778                    |
| C38              | 25578.98       | 882.034       | 0.349                    |
| C40              | 27056.27       | 932.975       | 0.395                    |
| C42              | 46174.93       | 1592.239      | 1                        |
| C44              | 31653.11       | 1091.486      | 0.541                    |

Table 11 gives the P-value of the optimum GFA derived QSAR model at 95% confidence level. SS, DF, MS, and F in the table represents sum of squares, degree of freedom, mean square and F-ratio respectively.

**Table 11: P-value of the model at 95% confidence level**

| Source     | SS     | DF | MS       | F       | p-value |
|------------|--------|----|----------|---------|---------|
| Difference | 3.5803 | 8  | 0.4475   | 63.4162 | <0.0001 |
| Error      | 0.1411 | 20 | 0.007057 |         |         |
| Null model | 3.7214 | 28 |          |         |         |

Tables 3, 4, and 5 give the GFA derived QSAR models for predicting minimum inhibitory concentration (MIC) of Schiff bases, validation parameters of the models, and detailed definition of the descriptors used in the models respectively. Based on the validation parameters, the octa-parametric model (model 1) was selected as the optimization model for predicting the MIC of Schiff bases. The Genetic Function Algorithm derived QSAR model is good agreement with the minimum standard shown in Table 2 as  $R^2 = 0.9622$ ,  $R^2_{adj} = 0.9470$ ,  $Q^2 = 0.9132$ ,  $R^2_{pred} = 0.6837$  and the Golbraikh and Tropsha criteria for robust QSAR model were also met, the result in Table 7 confirms

this. The predictability of model 1 is evidenced by the low residual values observed in Table 8 which gives the comparison of observed and predicted MIC of the molecules. Also, the plot of predicted pMIC against observed pMIC shown in Figure 1 indicates that the model is well trained and it predicts well the pMIC of the compounds. Furthermore, the plot of observed pMIC versus residual pMIC (Figure 2) indicates that there was no systemic error in model development as the propagation of residuals was observed on both sides of zero [21].

Table 11 gives the P-value of the optimization model (model 1). The value of P ( $< 0.0001$ ) at 95% confidence level shows that the alternative hypothesis that the magnitude of the observed inhibitory activity of Schiff bases against *E. coli* is a direct function of the empirical property (ies) or the theoretical parameter(s) which makes the descriptor of the total chemical structure of the molecules under investigation takes preference over the null hypothesis which state otherwise.

The applicability domain of the optimization model (model 1) was also defined for test set (Table 9) and training set (Table 10) compounds using Euclidean based approach. The results showed that all the compounds fall within the applicability domain of the model as their normalized mean distance score fall within the range of 0 and 1.

The result of the QSAR modelling hinted the predominance of the size descriptor ETA-Eta-P-F-L (Local functionality contribution EtaF\_local relative to molecular size) over other descriptors in the model in influencing the anti-salmonella typhi bioactivity of the studied Schiff bases owing to its relatively high numerical coefficient. The positive value of the coefficient of the descriptor implies that the minimum inhibitory concentration (MIC) of schiff bases is directly proportional to the value of this descriptor. Thus, the inhibitory activity of Schiff bases decreases with the increase in value of this descriptor since activity of drug varies inversely with its minimum inhibitory concentration.

Molecular size is a factor affecting the distribution of extremely large molecules. There is a high tendency of very large molecular sized drugs to be largely confined to the extracellular fluid or plasma compartment affecting their distribution in the body [22]. The decrease in bioactivity of the Schiff bases against *E. coli* with increase in molecular size as depicted in the optimization model (Model 1) may be due to the possibility of the large molecules being largely confined to the extracellular fluid or plasma compartment.

#### RECOMMENDATION

In the future design of novel Schiff bases as anti-*E. coli* drug, it is recommended based on this research that the compounds should be made less bulkier as possible since molecular size is negatively correlated to the bioactivity of the compounds as shown in the GFA derived model.

#### CONCLUSION

The generated QSAR models, performed to explore the structural requirements controlling the observed antibacterial properties, hinted that the biological activities were predominantly affected by size descriptor ETA-Eta-P-F-L (Local functionality contribution EtaF\_local relative to molecular size). The robustness and applicability of QSAR equation has been established by internal and external validation techniques. It is envisaged that the wealth of information in this QSAR model will provide an insight to designing a novel bioactive nickel-schiff base complex that will curb the emerging trend of multi-drug resistant strain of the bacterium, *E. coli*.

#### CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this paper. Also, they declare that this paper or part of it has not been published elsewhere.

#### CONTRIBUTION OF THE AUTHORS

This work was carried out in collaboration among all authors. Authors JPA and OCC designed the study and wrote the protocol. Author JPA, OCC, and OSB did the literature search and performed the statistical analysis. Authors JPA and OSB wrote the first draft of the manuscript. All authors read and approved the final manuscript.

#### REFERENCES

- [1] F.M Aarestrup; H.C Wegener; P. Collignon. *Expert Rev Anti Infect Ther.* **2008**, 6:733–50.

- 
- [2] L. Blaettler; D. Mertz; R. Frei; L. Elzi; A.F. Widmer; M. Battegay, M. *Infection*. **2009**, 37: 534–9.
- [3] G. Kronvall. Antimicrobial resistance. *APMIS*. **2010**; 118:621–39.
- [4] H. Von Baum; R. Marre. *Inter. J. of Med.Micro*. **2005**, 295: 503-11.
- [5] S.V. Sodha; M. Lynch; K. Wannemuehler; M. Leeper; M. Malavet; J. Schaffzin. *Epidemiol Infect*. **2011**, 139:309–16.
- [6] U.R. Misbah; I. Muhammad; A. Muhammad. *Amer. J. Applied Chem*. **2013**, 1(4): 59-66.
- [7] M.A. Hafiz; A. Dildar; M. Hira. *Pak. J. Pharm. Sci.*, **2015**, 28(2): 449-455.
- [8] A. Malik; K.S. Sher; R. Wajid; H. A. Zonera; I. Muhammad. *African J. Biotechnol.* **2011**, 10(2): 209-213.
- [9] K.B.Santhosh; K.G. Parthiban. *Asian J. Pharm. Clin. Res*. **2011**, 4 (4).
- [10] S.K. Sahu; M.D. Afzal; M. Banerjee; S. Acharya; C.C. Behera; S. Si. *J. Braz. Chem. Soc.*, **2008**: 19 (5).
- [11] O. P. Sanja; J.B. Dijana; D.C Dragoljub. *APTEFF*, **2008**; 39, 1-212.
- [12] W.J. Hehre, W. J. *Practical Strategies for Electronic Structure Calculations*, Wavefunction, Inc., Irvine, CA. 2005.
- [13] K. Roy, K. *Springer Briefs in Molecular Science*. **2015**: DOI 10.1007/978-3-319-17281.
- [14] V. Ravichandran; R. Harish; J. Abhishek; S. Shalini; P.V. Christopher; K.A. Ram. *Int. J. Drug Design Discov.*, **2011**, 2: 511-519.
- [15] A. Golbraikh; A.J.Tropsha. *Mol. Graphics Mod*. **2002**, 20, 269-276.
- [16] W. Wu; C. Zhang; W. Lin; Q. Chen; X. Guo; Y. Qian. *PLoS ONE*, **2015**, 10(3): e0119575.
- [17] K. Brandon; O. Aline. Comprehensive R archive network (CRAN): [http:// CRAN.R-project.org](http://CRAN.R-project.org). Retrieved July 3<sup>rd</sup>, **2015**.
- [18] D. Rogers; A.J. Hopfinger. *J Chem Inf Comput Sci*. **1994**, 34: 854– 866. doi/abs/10.1021/ci00020a020
- [19] J.F. Friedman. *Multivariate Adaptive Regression Splines, Technical Report No. 102*, **1990**.
- [20] A. Pravin. DTC\_Euclidean Programme, Drug Theoretics & Cheminformatics (DTC) Laboratory, Jadavpur University, **2013**.
- [21] M.J. Heravi; A. Kyani. *J. Chem. Inf. Comput. Sci*. **2004**, 44: 1328–1335.
- [22] D.J.Birkett (2002). *Pharmacokinetics Made Easy*, 2nd ed. Roseville, Australia, McGraw-Hill Australia.