The Journal of Engineering and Exact Sciences – JCEC, Vol. 05 N. 01 (2019)
journal homepage: https://periodicos.ufv.br/ojs/jcec
doi: https: 10.18540/jcecvl5iss1pp0063-0078
OPEN ACCESS – ISSN: 2527-1075

INSILICO MODELLING ON SOME C_{14}-UREA TETRANDRINE COMPOUNDS AS POTENT ANTI-CANCER INHIBITORS AGAINST HUMAN ERYTHROLEUKEMIA (HEL) CELL LINE

A. MUSTAPHA[1,*], G. A. SHALLANGWA[1], A. TIJJANI[2] and A. UZAIRU[1]

[1] Ahamdu Bello University, Faculty of Physical Sciences, Chemistry Department, Zaria, Kaduna State, Nigeria
[2] Department of Applied Chemistry, Federal University Dustin-ma, Katsina State, Nigeria

**Corresponding author. Ahmadu Bello University, Department of Chemistry, Zaria, Kaduna State, Nigeria, Phone: +2348064057773*
e-mail address: mustychem19@gmail.com (A.MUSTAPHA).

ARTICLE INFO

Article history:
Received 2018-12-20
Accepted 2019-02-11
Available online 2019-03-08

keywords
QSAR
Mean Effect
Validation
Descriptors
Model
Y-randomization

ABSTRACT

Insilico modelling was executed on 28 C_{14}-urea tetrandrine compounds as inhibitors of leukemic (HEL) cell lines using Quantitative Structure-Activity Relationship (QSAR) method. The structure of the inhibitors was correctly drawn, then geometrically optimized at Density Functional Theory (DFT) level (DFT/B3LYP/6-31G) with Spartan 14 V1.1.4. Also, molecular descriptors of the inhibitors were calculated with PaDEL calculator, and the results were partitioned into training and test set after data pretreatment. The training set was used to generate a model by employing genetic function approximation in choosing best descriptors to form the model. The validation parameters of the model include; R^2_{train} as 0.8067, LOF as 0.037, $r^2(Qcv)$ as 0.6378, R^2_{test} as 0.7629 and cR^2_p as0.6990 which have passed the criteria for acceptability of a QSAR model worldwide. In addition, the model depicted four (4) descriptors, AATS4v, AATS5i, AATSC5i, and GATS5m with positive mean effects signifying that increase in these descriptors will positively influence and increase the activity of the inhibitors. This study depicts a route in designing and synthesizing new* C14-*urea tetrandrine compounds with better inhibitory potentials.*

RESUMO

*A modelagem Insilico foi realizada em 28 compostos de tetrandrina C14-ureia como inibidores de linhagens leucêmicas (HEL) usando o método de Relação Estrutura-Atividade Quantitativa (QSAR). A estrutura dos inibidores foi corretamente desenhada, depois geometricamente otimizada ao nível da Teoria do Funcional da Densidade (DFT) (DFT / B3LYP / 6-31G *) com o Spartan 14 V1.1.4. Além disso, os descritores moleculares dos inibidores foram calculados com a calculadora PaDEL, e os resultados foram divididos em treinamento e teste após o pré-tratamento dos dados. O conjunto de treinamento foi utilizado para gerar um modelo empregando a aproximação da função genética na escolha dos melhores descritores para formar o modelo. Os parâmetros de validação do modelo incluem; R2train como 0.8067, LOF como 0.037, r2 (Qcv) como 0.6378, R2test como 0.7629 e cR2p como 0.6990 que passaram os critérios de aceitabilidade de um modelo QSAR em todo o mundo. Além disso, o modelo descreve quatro (4) descritores, AATS4v, AATS5i, AATSC5i e GATS5m, com efeitos médios positivos, significando que o aumento desses descritores influenciará positivamente e aumentará a atividade dos inibidores. Este estudo descreve uma rota na concepção e síntese de novos compostos de tetrandrina C14-ureia com melhores potenciais inibitórios*

1. INTRODUCTION

Human erythroleukemia (HEL) is a nasty syndrome formed as a result of some infrequent heterogeneous cells corresponding to about 3 to 4% of acute myeloid leukemia (AML) which give rise to red blood cells (Davey et al., 1995). Thus, the body produces a large amount of abnormal, immature white and red blood cells (erythrocytes). According to the World Health Organization (WHO), human erythroleukemia can be grouped into three subgroups. These include; (a) leukemia having multi-lineage dysplasia; (b) therapy-based acute myeloid leukemia and myelodysplastic disorders and (c) acute erythroid leukemia subdivided in erythroleukemia (erythroid/myeloid) and pure erythroid leukemia (Kowal-Vern et al., 2000). HEL is well-known to effects males, and the age for spreading of the disease seems to be bimodal, with a minimum of below 20 years and maximum in the seventh decade of life (Kowal-Vern et al., 2000). Tetrandrine, on the other hand, is a dibenzyltetrahydroisoquinoline compound derived from Chinese medicinal plant called *Stephania tetrandra* and it is reported to have anti-tumor activities, proliferation chemotherapeutic drugs and converses multidrug resistance (MDR) of tumor cell (Liu., 2016).

In recent decades, there was a significant number of studies that proved the success of the Quantitative Structure-Activity Relationship (QSAR) approach for prediction of various properties, such as solubility, lipophilicity, toxicity, mutagenicity, activities (Lan et al., 2017). By definition, a QSAR model is a mathematical linear equation involving molecular descriptors used in predicting the biological activity of a compound which is ought to be very useful in designing a new compound with better activity. Therefore, the main aim of this research was to develop a QSAR model of some C_{14}-urea tetrandrine compounds which can be used to predict the biological activities of Human erythroleukemic (HEL) cells using Genetic Function Approximation (GFA) method.

2. METHODOLOGY

2.1 Data set collection

A data set of twenty-eight (28) C_{14}-urea tetrandrine compounds as potent anti-cancer agents for this study were sourced from the literature (Lan et al., 2017). The biological activities of the inhibitors against leukemia (HEL) cell line were measured in IC_{50} (μM) which is the concentration of compound required to reduce 50% of the cell viability. This is further transformed to a logarithm scale (Eq. 1) so as to reduce skewness in the concentration values. The 2D structures of the compounds were drawn using Chem Draw software version 12.0.2, then aligned with their respective IC_{50} values as showed in Appendix table A1.

$$pIC_{50} = -\log(IC_{50} \times 10^{-6}) \quad (1)$$

2.2 Geometry Optimization

The molecular geometries of all the compounds were obtained by engaging Spartan V.14 at the density functional theory level (DFT/B3LYP/6-31G*) at ground state, (Becke, 1993; Lee et al., 1988). The geometry optimization is the process of computing the lowest energy of conformation for a given compound which also corresponds to its most stable structure.

2.3 Molecular descriptor calculation and Data pretreatment

The optimized twenty-four (24) molecules were subjected to PaDEL calculator to compute their molecular descriptors including electronic, spatial, structural, constitutional, geometrical, physiochemical, autocorrelation, thermodynamic, and topological descriptor (Alisi et al., 2018). The data generated from the PADEL- software in MS Excel (.csv) format were observed to contain redundant data, zero or non-informative descriptors, as such the data were further subjected to the pre-treatment process using a data pre-treatment malware downloaded from Drug Theoretical and Cheminformatics (DTC lab) website so as curate the results (Ambure et al., 2015).

2.4 Data Set Division

The pre-treated data were split into two sets (training and test sets) by employing Kennard-Stone's algorithm division technique using a division software also gotten from DTC Lab and (Kennard and Stone, 1969).

2.5 Model Generation and Validation

The training set was exported to material studio software for model building using genetic function approximation (GFA) approach, where the dependent variable is the inhibitory concentration (IC_{50}) and the independent variables are the molecular descriptors. The fitness score of the resultant GFA model during the evolution process was measured using Friedman formula (Eq. 2) which determines the finest fitness score defined as; (Friedman, 1991). In Materials Studio, LOF expression (Eq. 2) is slightly different from the original Friedman expression (1991).

$$LOF = \frac{SSE}{M\left[1-\beta\left(\frac{c+d\times p}{M}\right)\right]^2} \quad (2)$$

where c represents the number of the terms in the model, **d** represents a scaled smoothing factor, **p** corresponds to the entire number of descriptors in the model, $\boldsymbol{M}$ represents the number of inhibitors or compounds that made up training set and $\boldsymbol{\beta}$ is a safety factor with a value of 0.99 which guarantee that the denominator of the equation can never be equal to zero (Khaled and Abdel-shafi, 2011). **SSE** is the Sum of Squares of Errors and it is defined by the expression (Eq. 3) below;

$$SSE = \sqrt{\frac{(Y_{exp} - Y_{pred})^2}{N - P - 1}} \quad (3)$$

SSE value gives an idea about the quality of a model, low SEE value signifies better model and vice versa.

2.6 Internal Validation

The validity of the QSAR model examined using leave-one-out (LOO) cross-validation which provides a rigorous internal check on the model. It is also used to check the true predictive power or reliability of the model. The cross-validation regression coefficient, r^2 (Q^2_{cv}) were also calculated using Eq. 4:

$$r^2(Q^2_{cv}) = 1 - \left[\frac{\Sigma(y - y_{pred})^2}{\Sigma(y - \bar{y}_{tr})^2}\right] = 1 - \frac{PRESS}{SST} \quad (4)$$

Where $\overline{y}_{tr}$ is the average observed activities for the training set, y is the observed activity, and y_{pred} is the predicted activity of the training set respectively (Brandon, 2015). PRESS is the predictive sum of squares of a model, SST is the total sum of squares which correspond to the mean-corrected sum of squares of the responses over the entire data set.

The cross-validation coefficient $r^2(Q^2_{cv})$ is one of the key parameters that assess the predictive sway of a model. However, r^2 (Q^2_{cv}) score closer to 1.0 depicts high predictive influence. Furthermore, r^2 (Q^2_{cv}) score ought to be nearly close to the regression coefficient (R^2). But r^2 (Q^2_{cv}) that is far less than R^2 probably suggest data overfitting by the model. The r^2 (Q^2_{cv}) score of 0.0 by a model means no predictive consistency at all, according to the cross-validation criterion.

The values of regression coefficient (R^2) are directly proportional to the number of descriptors. However, the R^2 values are not consistent for evaluating the strength of the model. Thus, R^2 is adjusted with the mandate to refurbish and stabilize the model equation. The R^2 (adjusted) is given as in Eq. 5:

$$R^2_{adj} = \frac{R^2 - p\,(n-1)}{n - p + 1} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} \quad (5)$$

Where **p** is the number of descriptors or parameters in the regression model and **n** is the number of compounds that made up the training set (Brandon, 2015). The adjusted r^2 is significant only when there are some degrees of freedom for lack-of-fit. For a model with an additional parameter to be an improvement, the addition of the parameter is required to decrease SSE at least proportionately to the reduction in the degrees of freedom.

2.7 External Validation

The model developed was further subjected to external validation in order to measure its prediction competency using the test set and the regression coefficient(R^2_{pred}) value is given in Equation 6;

$$R^2_{pred} = 1 - \frac{\Sigma(Ypred_{test} - Y_{exp_{test}})^2}{\Sigma(Ypred_{test} - \overline{Y}_{training})^2} \quad (6)$$

Where; $Y_{pred_{test}}$ and $Yexp_{test}$ are the observed and predicted activity of the test set respectively. $\overline{Y}_{training}$ is mean scores of observed activity of the training set (Tropsha et al., 2003).

2.8 MLR Y-Randomization

In order to have confidence in the model built, Y-Randomization test was executed on the training set descriptors matrix (Tropsha, 2010). This is done by randomly shuffling the inhibitory concentrations (dependent variable) while keeping the descriptors (independent variables) constant resulting in the generation of random MLR models (Roy et al., 2012). The new QSAR models are anticipated to have significantly low R^2 and Q^2 values for 10 trials, which certify that the models are robust and cR^2_p is also calculated which should be more than 0.5, defined as in Eq. 7:

$$cR^2_p = R \times [R^2 - (R_r)^2]^{1/2} \quad (7)$$

where c R^2_p is the coefficient of determination, R is the coefficient of regression and R_r is average 'R' of random models.

2.9 Bias-variance estimation

The model generated was also assessed by examining the residuals (prediction errors) according to bias-variance evaluation. This method allows QSAR users to understand the contribution of the two components of the prediction errors, namely systematic error (bias) and random error (variance) in the model (Roy, 2017). The estimation was successfully achieved using a Bias-Variance Estimator downloaded from DTC lab website, and it uses bootstrapping procedure as a resampling process. The output parameters are $bias^2$ and variance defined as in the equations below;

$$Bias^2 = \frac{1}{N_c}\sum_{i=1}^{Nc}(\bar{Y}_{Pred(i)} - Y_{obs})^2 \quad (8)$$

$$\bar{y}_{pred(i)} = \frac{\sum_{j-1}^{N_B} y^{B(j)}_{pred(i)}}{N_B} \quad (9)$$

$$Variance = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{1}{N_B}\sum_{j=1}^{N_B}\left(Y^{B_j}_{Pred(i)} - \bar{Y}^{B}_{pred(i)}\right)^2 \quad (10)$$

Where N_c represents the number of compounds in the test set, $Y_{exp(i)}$ is the experimental response value of the compound 'i', $\bar{Y}_{Pred(i)}$ is the mean predicted response value of compound 'i' from 'i' bootstrap models, $Y^{B_j}_{Pred(i)}$ is the predicted response value of compound i from the bootstrap model 'j'

2.10 Statistical analysis of the descriptors

2.10.1 Mean Effect

The mean effect score of a descriptor is used to estimate its relative significance and contribution in the model and it is defined as:

$$Mean\ Effect = \frac{\alpha_j \sum_i^n d_j}{\sum_j^m(\alpha_j \sum_i^n d_j)} \quad (11)$$

where α_j represents the coefficient of the descriptor *j*, *dj* correspond to the value of each descriptor in the data matrix for each molecule in the training set, **m** represents the number of the descriptors in the model and n is the number of molecules in the training set (Minovski et al., 2013).

2.10.2 Varian Inflation Factor (VIF)

The variance inflation factor is a measure of the multicollinearity among the descriptors, usually expressed as:

$$VIF = (1 - R^2)^{-1} \quad (12)$$

where R^2 is the correlation coefficient. The VIF values ranging from 1 to 5 depicts that the model is stable and acceptable. Hence, VIF value corresponding to unity means that there is no inter-correlation between the variables. But, VIF value greater than 10 suggests that the model is unstable and unacceptable

(Myers, 1990)

2.11 Applicability Domain

A QSAR model applicability domain is usually tasked to explore the area where the compound predictions can be dependably useful. As such, chemical compounds that fall outside the applicability domain cannot make a very good prediction (Eriksson et al., 2003; Nandi et al., 2013). Consequently, the prediction that is interpolated in the chemical space is acceptable while extrapolated predictions in the chemical space are rejected as well. The leverage technique was engaged in evaluating the domain of applicability for the model generated and it is defined as the leverage values for the *ith* compound (Eq. 13) (Gramatica et al., 2007)

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

where X(i) is a vector of molecular descriptors of the compound, X is a matrix of descriptors for compounds from the training set, and X^T is the transpose matrix of X used in developing the model. The threshold leverage (h^*) is defined as borderline of normal leverage scores for *X* outliers (Eq. 14):

$$\boldsymbol{h}^* = 3\frac{(r+1)}{n} \quad (14)$$

Where $\boldsymbol{n}$ is the number of training compounds and $\boldsymbol{r}$ is the number of descriptors in the model.

3. RESULTS AND DISCUSSION

3.1 Descriptor Calculations

The QSAR study was performed to generate a model that relates data from the structure of C_{14}-urea tetrandrine compounds with its inhibitory activity against Human erythroleukemia (HEL) cell lines. Initially, the 32 quantum chemical descriptors for all the drawn compounds were obtained from Spartan 14 software via the optimization process. These were pooled with the 1875 molecular descriptor calculated by PaDEL-Descriptor calculator V2.20 to give 1907.

3.2 Data Pretreatment and Division

The PaDEL-Descriptor output in MS Excel (.csv) were subjected to data pretreatment which removed non-informative constant data and pair of variables with a correlation coefficient greater than 0.7 using the Data pre-treatment software. The data set results from the pretreatment process was divided by using Kennard-Stone algorithm method, where 19 compounds are considered as training set and 9 compounds are the test set. The division was successfully done using the Dataset Division GUI 1.2 software.

3.3 Model Building and Validation

In building the QSAR model, four (4) optimum descriptors were selected via Genetic Function Approximation (GFA) of material studio software and the model generated is illustrated below:

pIC_{50}= 0.056049326 ***AATS4v** + 0.125292658 * **AATS5i** + 4.178045312 * **AATSC5i** + 2.792929285 * **GATS5m** -28.6592 (15)

The validation parameters of the model were presented in Table 1, which clearly shows that the model passed the criteria of acceptability. Also, the coefficients of regression (R-squares) are 0.8053 and 0.7629 for both the training and test set inhibitors respectively. This is an indication of a good relationship between the predicted and observed activities.

Table 1 – Internal validation parameters

Model Validation			
Parameter	**Value**	**Threshold**	**Ref.**
LOF(Friedman)	**0.0367**	-	
R^2 (training)	0.8067	≥ 0.6	(Tropsha, 2010)
R^2 (Adjusted)	0.7515	-	
$Q_{CV}(r^2)$	0.6398	≥ 0.5	(Tropsha, 2010)
S. Regression	Yes	-	
Critical F-value	3.1601	-	
Lack of fit	14	-	
R^2 (test)	0.7629	≥ 0.6	(Tropsha, 2010)
No. of Bootstrap Models	10,000	-	
Bias^2	0.0186	-	
Variance	0.00483	-	
R^2_{Pred}	0.6209	≥ 0.6	(Tropsha 2003)

The values for model external validation of the test set inhibitors was reported in Table 2, and the R^2_{Pred} was computed as 0.6209. The entire model descriptors scores of the dataset, model prediction results which encompasses observed, predicted inhibitory concentration (pIC_{50}) and their residual scores were presented in Table 3 and 4 respectively.

Table 2 – External validation of the test set compounds

ID No.	Y_{pred}	Y_{exp}	$(Y_{pred}-Y_{exp})$	$(Y_{pred}-Y_{exp})^2$	$(Y_{pred}-Y_{tr})$	$(Y_{pred}-Y_{tr})^2$
22	4.9068	4.9362	-0.0294	0.000869	-0.27606	0.076211
20	5.4456	5.4788	-0.0332	0.001105	0.262744	0.069034
14	5.1639	5.1890	-0.0251	0.000634	-0.01897	0.00036
17	4.9460	5.0958	-0.1498	0.022445	-0.23687	0.05611
2	5.1296	5.3439	-0.2142	0.045885	-0.05319	0.002829
1	5.0808	5.3196	-0.2387	0.057016	-0.102	0.010404
4	4.9161	5.0259	-0.1098	0.012056	-0.26673	0.071145
24	5.3523	5.3306	0.0217	0.000471	0.169502	0.028731
27	5.4456	5.3746	0.0709	0.005036	0.262773	0.069049
				$\sum = 0.1455$		$\sum = 0.3838$

Hence, $R^2_{Pred} = 1 - \frac{0.1455}{0.3838} = 0.6209$

Table 3 - Descriptors and their scores

Name	*AATS4v*	*AATS5i*	*AATSC5i*	GATS5m
1 [a]	191.2172	158.9763	0.113391	0.941708
2 [a]	188.7672	159.5503	0.141485	0.940577
3	179.6935	159.7312	0.181305	1.019785
4 [a]	189.4507	158.6309	0.126677	0.913801
5	190.2577	159.6393	0.166333	0.923922
6	195.5674	158.9422	0.130958	0.800991
7	189.6495	158.8044	0.106531	0.95271
8	191.3768	159.1293	0.102356	0.932943
9	194.9906	159.0844	0.097067	0.945155
10	194.7097	159.0556	0.100522	0.930606
11	197.6377	160.5527	0.080365	0.921723
12	195.1452	159.0005	0.134983	0.917208
13	193.6035	160.2301	0.112096	0.935366
14 [a]	196.5985	158.7824	0.110915	0.875849
15	195.3012	159.1388	0.119859	0.940639
16	197.698	160.1874	0.152365	0.784384
17 [a]	191.4124	158.9583	0.090172	0.925043
18	190.1657	159.3008	0.102022	0.949016
19	197.5386	158.5713	0.071864	0.958166
20 [a]	206.3831	157.2727	0.074365	0.902759
21	182.7771	159.9879	0.1438	0.945972
22 [a]	186.465	159.221	0.114748	0.96175
23	195.8065	159.7764	0.140898	0.848063
24 [a]	190.3263	159.7573	0.179237	0.923259
25	196.6429	158.3903	0.144702	0.855029
26	191.9665	158.7985	0.16282	0.944654
27 [a]	191.3396	158.2101	0.13436	1.072861
28	193.8159	158.2857	0.13424	0.952818

[a] superscript signify test set

Table 4 - QSAR predictions results

Training Set				Test Set			
ID No.	Activity	Predict	Residuals	ID No.	Activity	Predict	Residual
3	4.9248	5.0313	-0.1065	1	5.3196	5.0808	0.2387
5	5.3957	5.2816	0.11408	2	5.3439	5.1296	0.2142
6	5.0296	5.0008	0.02884	4	5.0259	4.9161	0.1097
7	4.9093	4.9735	-0.0641	14	5.1890	5.1639	0.0251
8	5.0925	5.0384	0.0541	17	5.0958	4.9460	0.1498
9	5.3133	5.2473	0.0660	20	5.4788	5.4456	0.0332
10	5.1771	5.2017	-0.0246	22	4.9362	4.9068	0.0294
11	5.3957	5.4444	-0.0486	24	5.3306	5.3523	-0.0217
12	5.3506	5.3258	0.0248	27	5.3746	5.4456	-0.070
13	5.3242	5.3485	-0.0243	-			
15	5.1844	5.3541	-0.1697	-			
16	5.2668	5.3192	-0.0525	-			
18	5.0491	5.0355	0.0136	-			
19	5.2831	5.2569	0.0262	-			
21	4.8291	4.8735	-0.0443	-			
23	5.2588	5.2917	-0.0328	-			
25	5.0366	5.200288	-0.1636				
26	5.4341	5.315348	0.118804				
28	5.2189	5.258146	-0.03918				

By definition, Residual score is the differences between observed and predicted activity, and lower residual values signify high extrapolative ability of the model. In addition, the model generated was assessed by developing 10,000 bootstrap models of the same sample size starting from the training set, in order to estimate the magnitude of systematic (bias) and random (variance) errors (Roy, 2017). The bias, variance and mean square errors were very insignificant, which depicts that the model predictions are good.

Table 5 - Pearson's correlation analysis

Descriptors	*AATS4v*	*AATS5i*	*AATSC5i*	*GATS5m*
AATS4v	1			
AATS5i	-0.17113	1		
AATSC5i	-0.46341	0.08196	1	
GATS5m	-0.58045	-0.11521	-0.13927	1

3.4 Statistical Analysis of the Descriptors

In order to assess the relationships among descriptors in the model, values of the four (4) descriptors were extracted from the training set, then subjected to Pearson's correlation analysis and the results were described in Table 5. The result shows that there is an insignificant inter-correlation among the descriptors because the correlation coefficients between all pairs are less than 0.6.

The results in Table 6 illustrates some statistical parameters of descriptors in the developed model. It shows that the variance inflation factor (VIF) scores of all descriptors in the model are not greater than 4, which is acceptable. Similarly, the p-values of all descriptors in the model are less than 0.05, which means that there is a relationship between the descriptors and the inhibitory concentration of the compounds. The output of Y-Randomization test was presented in Table 7. The cR^2_p value was calculated as 0.6990 which is greater than 0.5.

Table 6 - Statistical parameters

Descriptors	*Coefficients*	*VIF*	*P-value*	*Mean Effect*
AATS4v	0.0573	3.1278	3.1E-06	0.3291
AATS5i	0.1213	1.1495	0.0038	0.5754
AATSC5i	4.1771	1.9406	0.0008	0.0156
GATS5m	2.9181	2.4751	0.0001	0.0798

Table 7- Y-randomization test

Model	R	R^2	Q^2
Original	0.8981	0.8067	0.6398
Random 1	0.4163	0.1733	-0.6024
Random 2	0.3546	0.1257	-0.5840
Random 3	0.3187	0.1015	-0.5506
Random 4	0.4918	0.2419	-0.2730
Random 5	0.4504	0.2028	-0.4480
Random 6	0.5457	0.2978	-0.2576
Random 7	0.5147	0.2649	-0.3244
Random 8	0.1304	0.0170	-0.8243
Random 9	0.6325	0.4001	-0.1143
Random 10	0.6271	0.3932	-0.1854
Average r :	0.4482		
Average r^2 :	0.2218		
Average Q^2 :	-0.416		
cRp^2 :	0.6990		

A Plot of standardized residual versus observed inhibitory concentration "Fig 1" showed a random scattering around the baseline of data at the standardized residual equal to zero which depicts the absence of systematic error.

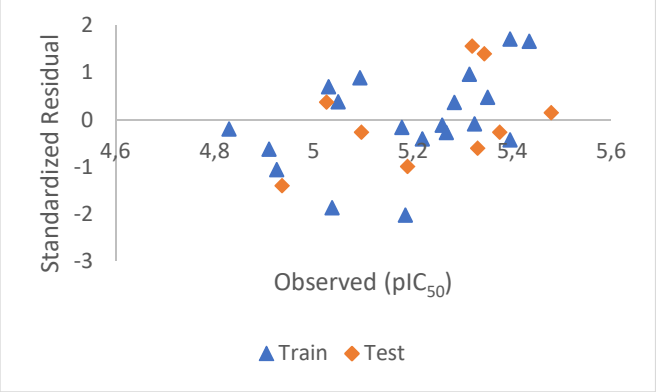


Figure 1 – Plot of Standardized residual against Observed (pIC_{50}).

A scatter plot for standardized residuals against the leverages also termed as Williams Plot was presented in "Fig 2" so as to detect the structural outliers or influential compounds. The plot revealed dispersion of inhibitors within ± 2 square area of standard deviation unit which means there is no Y-outlier. However, the calculated threshold leverage (h*) is 0.78, which revealed that three (3) test set compounds (i.e., compound 4, 20 and 27) are considered as structural X-outliers because their leverages are more than the threshold score. The reason is that of the differences in the substitution pattern of the chemical structure in the dataset.

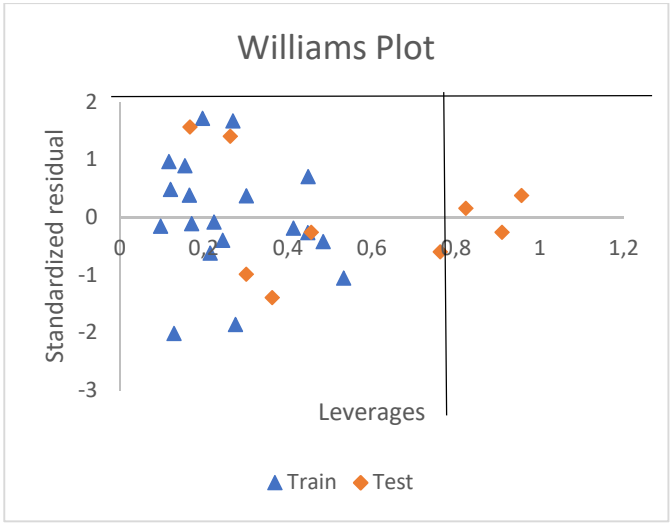


Figure 2 - Williams plot (Standardized residuals vs Leverages)

Plot of predicted versus observed activity (pIC_{50}) was presented in "Fig 3" which clearly shows the that the training set inhibitors are in agreement with the test set inhibitors.

3.5. **Meaning of the descriptors**

The four (4) descriptors in the model belong to the autocorrelation descriptor java class, and their descriptions were reported in Table 8.

Average Broto-Moreau autocorrelation descriptors (ATS_k) are generally computed as the graph invariant describing how the property considered is distributed along the topological structure. It is obtained by dividing each term by the corresponding number of contributions, thus avoiding any dependence on molecular size as in equation 16:

$$\overline{ATS_k} = \frac{1}{2\Delta_k} \sum_{i=1}^{A} \sum_{j=1}^{A} w_i \cdot w_j \delta(d_{ij}; k) \qquad (16)$$

Where Δ_k is the sum of the Kronecker delta function which corresponds to the total number of vertex pairs at distance equal to k (Todeschini and Consonni, 2009). AATS5i and AATS4v descriptors has the highest contribution with the mean effect of 0.5754 and 0.3279 respectively.

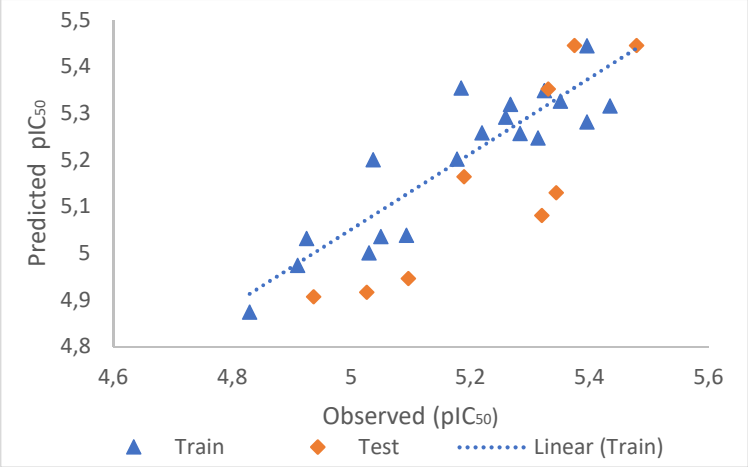


Figure 3 – Plot of predicted against observed activities (pIC_{50}).

Table 8- Descriptor class and descriptions

Descriptor	Description	Class
AATS4v	Average Broto-Moreau autocorrelation/ measured by van der Waals volumes (lag 4)	2D
AATS5i	Average Broto-Moreau autocorrelation - lag 5 / weighted by first ionization potential	2D
AATSC5i	Average centered Broto-Moreau autocorrelation - lag 5 / weighted by first ionization potential	2D
GATS5m	Geary autocorrelation - lag 5 / weighted by mass	2D

The Geary autocorrelation - interval 8 per weighted by the Vander Waals volumes (**GATS5m)** is also a 2D autocorrelation descriptor, obtained from molecular graphs by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (lag 5) (Todeschini and Consonni, 2009). The positive mean effect of these four (4) descriptors in this study inferred that an increase in their values will positively influenced and increases the inhibitory concentrations.

CONCLUSION

In conclusion, this research has successfully achieved its aim of generating a statistically significant model for predicting the inhibitory potentials of C_{14}-urea tetrandrine compounds against Human erythroleukemia (HEL) cell line using Genetic Function Approximation (GFA) method. Our research findings revealed molecular descriptors AATS4v, AATS5i, AATSC5i, and GATS5m with positive mean effects depicts that an increase in the descriptors score, increases the activity of the inhibitors. Hence, this knowledge could be of vital importance in designing and synthesizing new C_{14}-urea tetrandrine compound with excellent inhibitory potentials.

REFERENCES

AMBURE, P.; RAHUL, B.A.; AGNIESZKA, G.; TOMASZ, P.; KUNAL, R. "NanoBRIDGES" Software: Open Access Tools to Perform QSAR and Nano-QSAR Modeling. *Chemical Intelligence Laboratory Systems.* v.147, p.1–13, 2015.

ALISI, I.O.; UZAIRU, A.; ABECHI, S.E.; IDRIS, S.O. Quantitative Structure activity relationship analysis of coumarins as free radical scavengers by genetic function algorithm. *Iranian Chemical Society*. v.6, p.208–222

BECK, A.D. Becke's three parameter hybrid method using the LYP correlation functional. *Journal of Chemical Physics*. v.98, p.5648–5652, 1993

BRANDON, V.; ORR, K.A. Comprehensive R archive network (CRAN): http:// CRAN.Rproject.org; 2015:112-113

DAVEY, F.R.; ABRAHAM, J.R.N.; BRUNETTO, V.L.; MACCALLUM, J.M.; NELSON, D.A.; BALL, E.D. Morphologic characteristics of erythroleukemia (acute myeloid leukemia; FAB-M6): a CALGB study. *American Journal of Hematology*, v.49, p. 29–38, 1995.

ERIKSSON, L.; JAWORSKA, J.; WORTH, A.P.; CRONIN, M.T.D.; MCDOWELL, R.M.; GRAMATICA, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-baes QSARs. *Environmental Health Perspectives*, v.111, p.1361-1375, 2003.

FAN, Y.; LU, H.; AN, L.; WANG, C.; ZHOU, Z.; FENG, F.; ZHAO, Q. Effect of active fraction of Eriocaulon sieboldianum on human leukemia K562 cells via proliferation inhibition, cell cycle arrest and apoptosis induction. *Environmental Toxicology and Pharmacology*, v.43, p.13-20, 2016.

FRIEDMAN JH, Multivariate Adaptive Regression Splines. *The Annals of Statistics*, p.1–67, 1991.

GRAMATICA, P.; GIANI, E.; PAPA, E. Statistical external validation and consensus modeling: A QSPR case study for KOC prediction. Journal of Molecular Graphics Modelling, v.25, p.755-66. DOI: 10.1016/j.jmgm.2006.06.005, 2007.

KENNARD, R.W.; STONE, L.A. Computer Aided Design of Experiments. *Technometrics*, v.11 (1), p.137–48, 1969,

DOI: 10.1080/00401706.1969.

KHALED, K.F.; ABDEL-SHAFI, N.S. Quantitative structure and activity relationship modeling study of corrosion inhibitors: Genetic function approximation and molecular dynamics simulation methods. *International Journal of Electrochemical Science*, v.6, p.4077-4094, 2011

KOWAL-VERN, A.; MAZZELLA, F.M; COTELINGAM, J.D; SHRIT, M.A.; RECTOR, J.T.; SCHUMACHER, H.R. Diagnosis and characterization of acute erythroleukemia subsets by determining the percentages of myeloblasts and proerythroblasts in 69 cases. *American Journal of Hematology*, v.65, p. 5–13, 2000.

LIU, T.; LIU, X.; LI, W.H. Tetrandrine, a Chinese Plant-Derived Alkaloid, Is a Potential Candidate for Cancer Chemotherapy, *On Co-Target*, v.7, p.480100–480115, 2016.

LAN, J.; HUANG, L.; LOU, H.; CHEN, C.; LIU, T.; HU, S.; YAO, Y.; SONG, J.; LUO, J.; LIU, Y.; XIA, B.; XIA, L; ZENG, X.; BEN-DAVID, Y.; PAN, W. Design and Synthesis of Novel Tetrandrine Derivatives as Potential Anti-Tumor Agents against Human Hepatocellular Carcinoma. *European Journal Medicinal Chemistry*, v.12, p. 3-4, 2017 DOI: 10.1016/j.ejmech.11.007s

LEE, C.; YANG, W.; PARR, R.G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, p.37-785, 1988

MINOVSKI, N.; ŽUPERL, Š.; DRGAN, V.; NOVIČ, M. Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: a case study. *Analytica Chemical Acta*, v.759, p.28–42, 2013.

MYERS, R.H. Classical and modern regression application. *2nd edition. Duxbury press*. CA. 1990

NANDI, S.; MONESI, A.; DRGAN, V.; MERZEL, F.; NOVIČ, M. Quantitative structure-activation barrier relationship modeling for Diels-Alder ligations utilizing quantum chemical structural descriptors. *Chemistry Central Journal*, v.7, p.1-13, 2013. DOI.org/10.1186/1752153X-7-171

PERKINS, R.; FANG, H.; TONG, W.; WELSH, W.J. Quantitative Structure-Activity Relationship Methods: *Perspectives on Drug Discovery and Toxicology*, v. 22(1), p.1666–1679, 2003.

ROY, K.; AMBURE, P.; & AHER, R. B. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? *Chemometrics and Intelligent Laboratory Systems*, v*162*, p44-54. 2017.

ROY, K.; MITRA, I.; KAR, S.; OJHA, P.K.; DAS, R.N.; KABIR H. Comparative studies on some metrics for external validation of QSPR models. *Journal of Chemical Informatics and Modellin*g. v.52, p.:396–408. 2012. DOI:10.1021/ci200520g

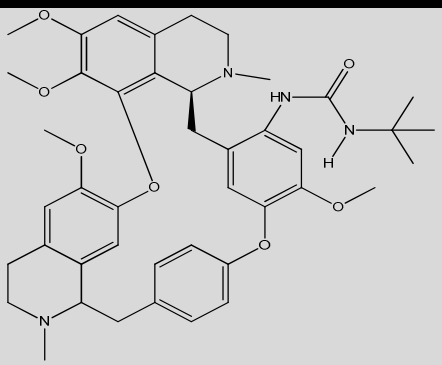
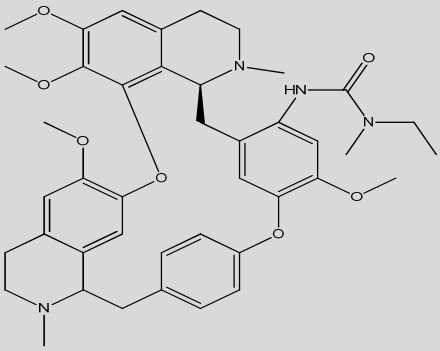
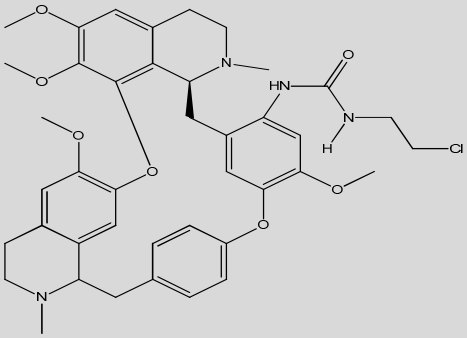
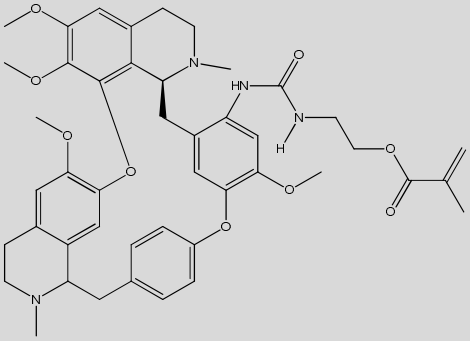
TODESCHINI, R.; CONSONNI, V. Molecular descriptors for chemo-informatics. Weinheim: Wiley- VCH, (Methods and principles in medicinal chemistry). 2009. ISBN: 9783527318520

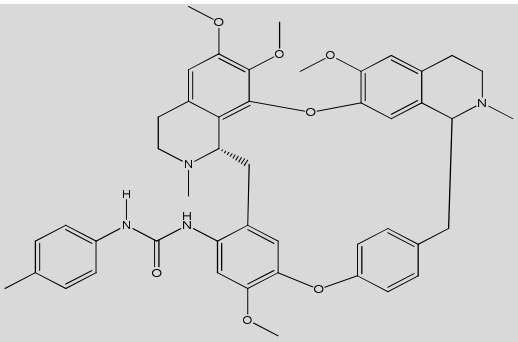
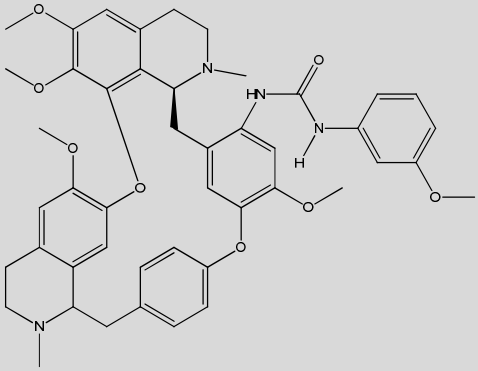
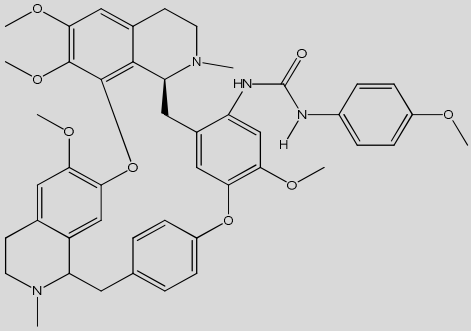
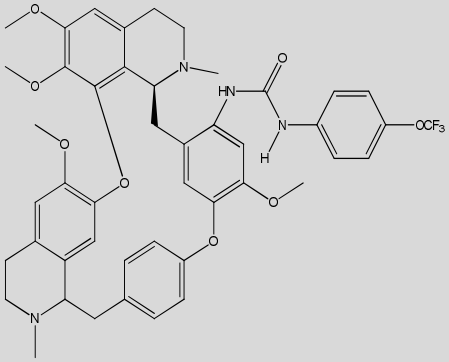
TROPSHA, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, v.29 (6–7), p.476–88, DOI: 10.1002/minf.201000061, 2010

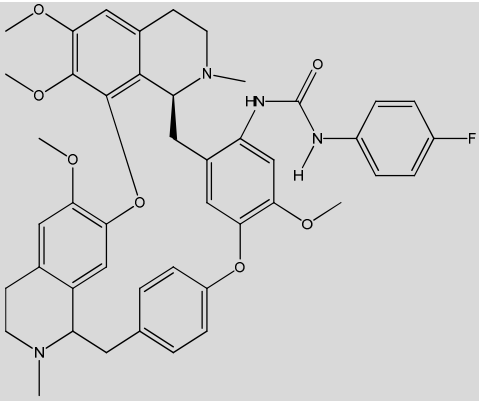
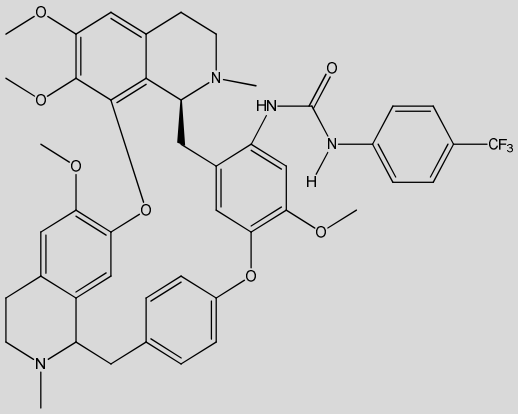
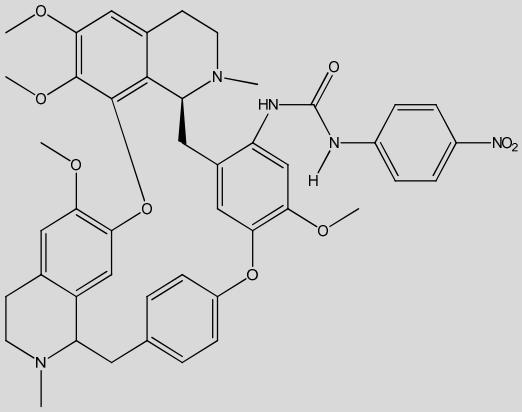
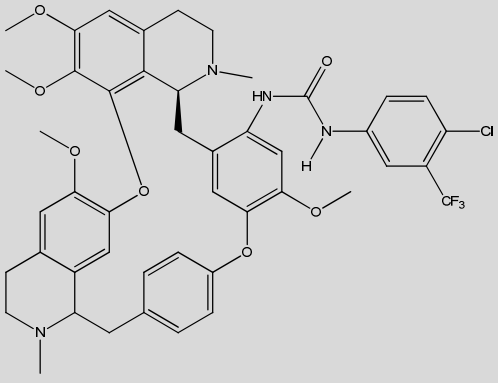
TROPSHA, A.; GRAMATICA, P.; & GOMBAR, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, v*22* (1), p69-77. 2003

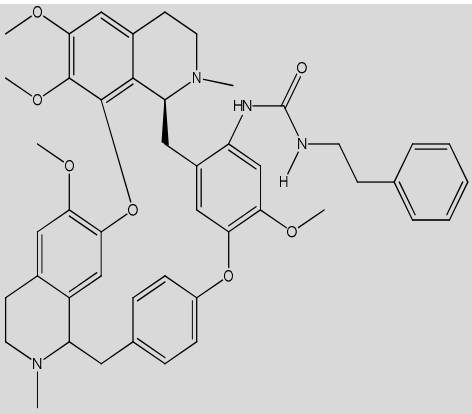
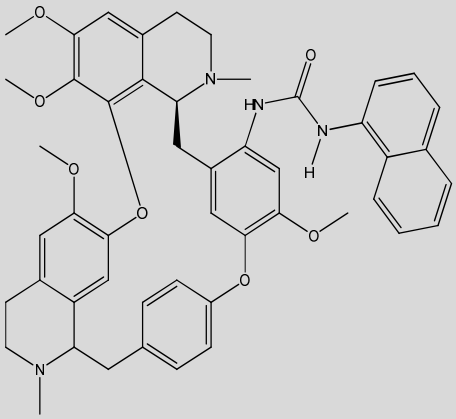
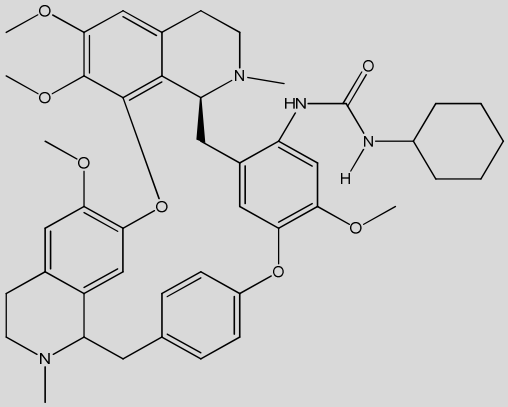
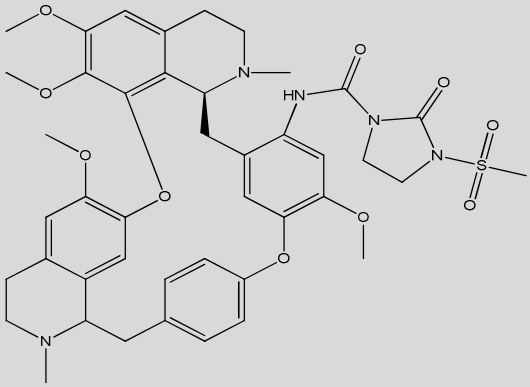
APENDIX A1

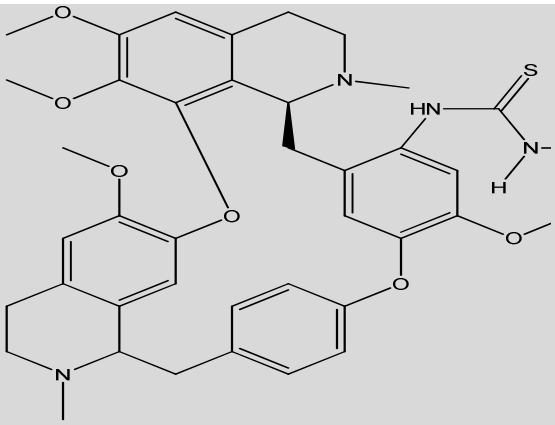
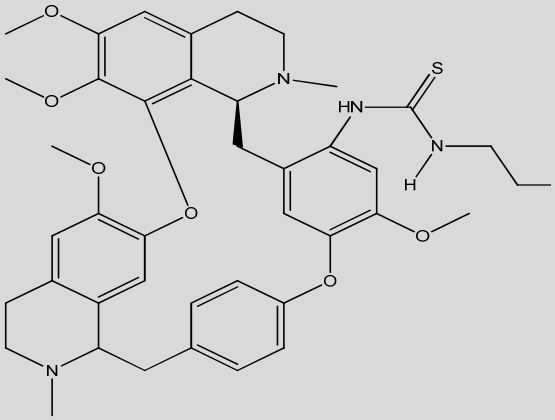
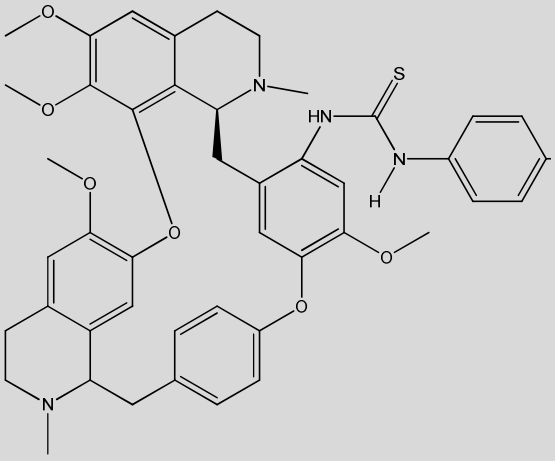
Table A1 - C_{14}-urea tetrandrine compounds and their Inhibitory Concentrations (IC_{50}) in μM against HEL Leukemic cell line.

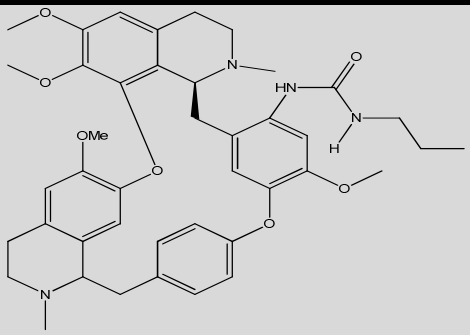
ID No.	Training set compound	IC50 (μM)	pIC50
3		1.19	5.9244
5		4.02	5.3957
6		9.34	5.0296
7		11.23	4.9492

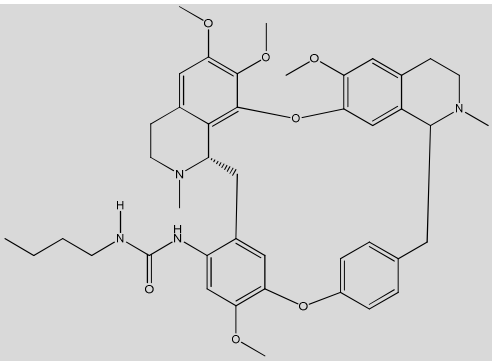
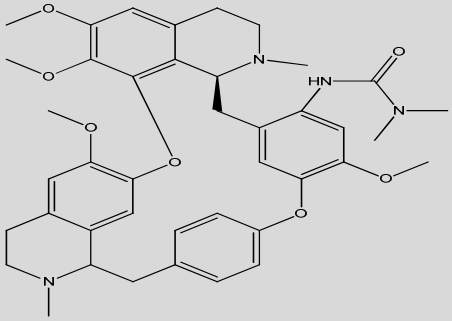
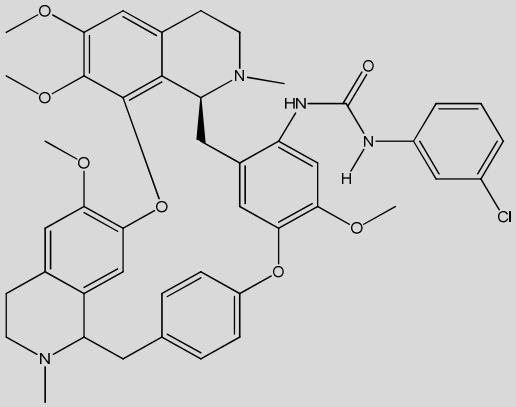
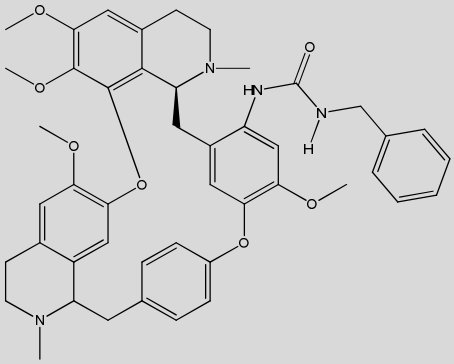
8		8.08	5.092
9		4.86	5.3133
10		6.65	5.1771
11		4.02	5.3957

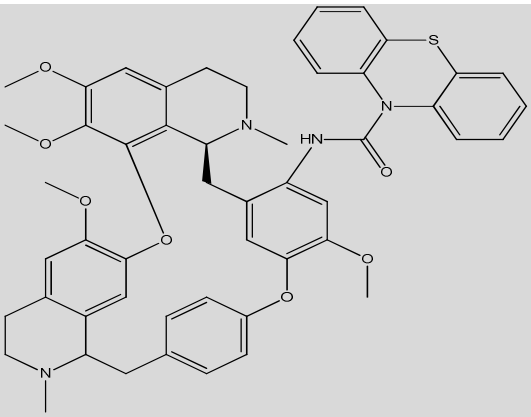
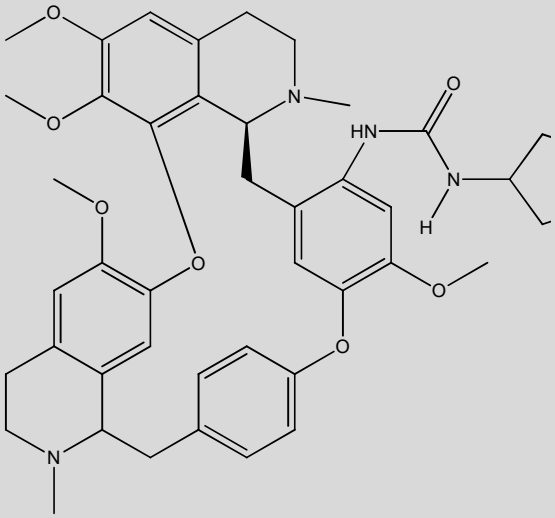
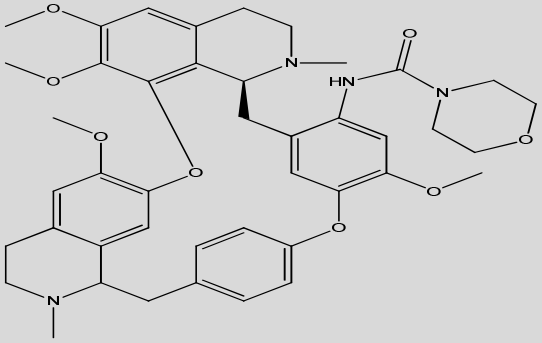
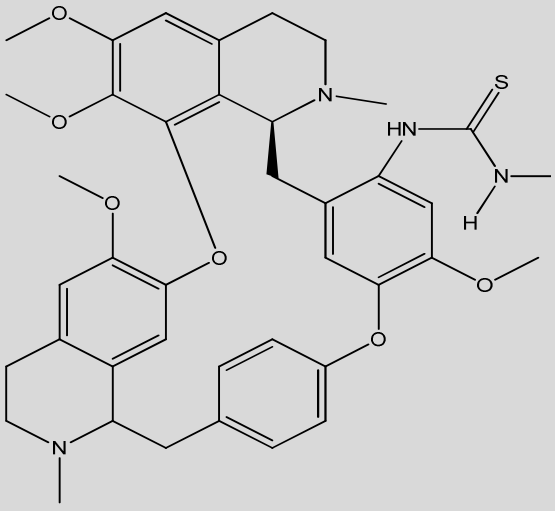
12		4.46	5.3506
13		4.74	5.3242
15		6.54	5.1844
16		5.41	5.2668

18		8.93	5.0491
19		5.18	5.2856
21		11.48	4.8297
23		5.51	5.2586

25		9.19	5.0366
26		3.68	5.4341
28		6.04	5.2189

ID No.	Test set Compound	IC50 (μM)	pIC50
1		4.79	5.3196

2		4.53	5.3439
4		9.42	5.0259
14		6.47	5.1890
17		8.02	5.0958

20		3.32	5.4788
22		11.16	4.9523
24		4.67	5.3306
27		4.22	5.3746