Insilico Modelling of Quantitative Structure-Activity Relationship of PGI50 Anticancer Compounds on k-562 Cell Line

David Ebuka Arthur, Adamu Uzairu, Paul Mamza, Stephen Eyije Abechi, Gideon Shallangwa

Publisher: Cogent OA

Journal: *Cogent Chemistry*

DOI: http://doi.org/10.1080/23312009.2018.1432520

Insilico Modelling of Quantitative Structure-Activity Relationship of PGI50 Anticancer Compounds on k-562 Cell Line

David Ebuka Arthur, Adamu Uzairu, Paul Mamza, Stephen Eyije Abechi, Gideon Shallangwa

Authors and Affiliations

David Ebuka Arthur*, Adamu Uzairu, Paul Mamza, Stephen Eyije Abechi, Gideon Shallangwa

Department of Chemistry,

Ahmadu Bello University (ABU) Zaria, Kaduna State, Nigeria

Corresponding Author

David Ebuka Arthur*

Salutation: Mr (PhD in view physical and Theoretical Chemistry)

Email: hanslibs@myway.com

TEL +2348138325431

Adamu Uzairu

Salutation: Prof. (PhD physical and Theoretical Chemistry)

Paul Mamza

Salutation: Prof. (PhD Polymer Chemistry)

Steven Abechi

Salutation: Dr (PhD Analytical Medicinal Chemistry)

Gideon Shallangwa

Salutation: Dr (PhD Inorganic Chemistry)

Abstract:

The pGI_{50} cytotoxicity values of 112 compounds on K-562 cancer cell line were modelled in order to illustrate the Quantitative structure activity relationship (QSAR) of the compounds. The data set were divided into training and test set through Kennard-stone algorithm, while the pool of molecular descriptors calculated with paDEL descriptor metric program was subjected to genetic functional algorithm (GFA) for selection of descriptor to be modeled. The statistical significance of the model was verified by calculating the values of Q^2_{LOO} (0.845), Q^2_{F1} (0.9397), Q^2_{F2} (0.6862) and R^2_{pred} (0.6862) needed to evaluate the strength and robustness of the model. The result of the internal and external validation of the model indicates that the model is good and could be used to predict the GI_{50} of anticancer compounds on K-562 leukemia cell line.

.

Keywords: K-562 cell line, QSAR, GFA-MLR, anticancer, Williams plot

Introduction:

Cancer is one of deadliest diseases in the world, it is caused by uncontrolled cellular growth. The disease is best seen as the inhibition of the defence mechanism responsible for the eradication of cells, which has been the backbone of carcinogenesis.

Cancer, reportedly kills 135,000 people a year, which is a bit higher than the from heart disease (News, 2003). Most cancer noticed have been reportedly linked to mutations caused by chemical exposure from environmental pollutants, food constituents, tobacco smoking etc. (Ferlay et al., 2010; Iuliano et al., 2012; Organization, 2002). Cancerous tumours are of two types, one malignant or Benign in nature (Siegel, Miller, & Jemal, 2015), and the other metastasis, which is the spread of cancer from the main site to other neighbouring organs, is the major cause of mortality in cancer suffering patients (Parkin, Boyd, & Walker, 2011). Some tumour cells have been reported to resist the effect of present day chemotherapeutic agents, given rise to a problem involving the clinical treatment of cancer, and so bringing our search for novel anticancer agents that selectively induce apoptosis.

K562 cells were the first human immortalised myelogenous leukemia line to be recognized. They are of the erythroleukemia type, and the cell line was gotten from a 53-year-old female chronic myelogenous leukemia patient in blast crisis (Drexler, 2000; Lozzio & Lozzio, 1975).The cells are non-adherent and rounded, they are positive for the BCR/ABL fusion gene, and bear some proteomic similarity to undifferentiated erythrocytes (Andersson, Nilsson, & Gahmberg, 1979). In culture they display much less clattering than many other suspension lines, probably due to the down regulation of surface adhesion molecules by bcr/abl. Though, additional study proposes that BCR/ABL over-expression may actually increase cell adherence to cell culture plastic (Karimiani et al., 2014). The issue with K562 cells, and numerous other cancer cell sorts, is an excess of Aurora kinases (Fan et al., 2016). These kinases assume a part in the

development of spindles, partition of chromosomes, and cytokinesis (Fan et al., 2016). These functions are important in cells so as to divide and regenerate tissues, and assume a support part in homeostatic capacities. Be that as it may, the excess of Aurora kinases takes into consideration uncontrolled cell division, bringing about tumor (Fan et al., 2016). Inhibiting these kinases is an essential direction mechanism of cancer, since it keeps cells from advancing into mitosis.

Computational design of novel molecule is a tool that has been used to accelerate discovery process, resulting in its acknowledgement and popularity. This is due to its tendency to reduce the classical trial and error approach (Roy, Kar, & Das, 2015b). Also, development of molecular modeling techniques such as quantitative activity relationship (QSAR), application of conformational search methodologies like molecular dynamics and Monte-Carlo simulations and so on have also contributed greatly to discovery and development of new molecules (Sabet, Mohammadpour, Sadeghi, & Fassihi, 2010; A. Speck-Planche, V. V. Kleandrova, F. Luan, & M. N. D. Cordeiro, 2012; A. Speck-Planche, V. V. Kleandrova, F. Luan, & M. N. D. S. Cordeiro, 2012). The purpose of this study to develop a new in silico QSAR model, that can be used to screen the bioactivity of known and hypothetical molecules against K-562 cancer cell line, and further design new active molecules by altering molecular descriptors and chemical fragments which were found to be significant within the applicability domain of the model.

Experimental Section:

The computational hardware and software used in this work includes: Computer (HP pavilion Intel(R) core i5-4200U with 1.63Hz and 2.3Hz processor and windows 8.1 operating system), Spartan 14 (Hehre & Huang, 1995), ChemBio Ultra 12.0 (Evans, 2014; Li, Wan, Shi, & Ouyang, 2004), Padel-descriptor (Yap, 2011), MS Excel (Denton, 2001).

The data set contained 112 molecules used to evaluate the relationship between the chemical fingerprints of the compounds and their anticancer activities on human leukaemia (K-562) cell line (Marx, O'Neil, Hoffman, & Ujwal, 2003). The chemical structures of the data set, NSC and CAS number were taken from the drug discovery and development arm of the National Cancer Institute (NCI) (https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data). The data contains aminopterin and camptothecin derivatives, colchicine analogues and so on. The anticancer activity results are shown in GI_{50}, which is the concentration for 50% of cancer cell proliferation (Marx et al., 2003). Some the compounds containing salts or small fragments were treated separately, the metal ions and chloride ions were removed since they play no significant contribution to the activity of the drugs, this was collaborated by authors such as Fatemi (Fatemi, Heidari, & Gharaghani, 2015) and Roy (Kar & Roy, 2012; Roy, Kar, & Das, 2015a). the counterpart of the ions was optimized at a protonated state, as they should in solution.

The biological activity ($-LogGI_{50}$) of the studied compounds were presented in Table 1 and the data set of the activities ranges from 2.2 to 9.3. Further literature (Chopade, Phadnis, Hodage, Wadawale, & Jain, 2015), showing the wide range of activities data set used to improve the quality of information gotten from the compounds.

Generation of molecular descriptors

The 2D structure of each of the compounds was generated using the sketch option on Spartan 14 and was converted into 3D structure by using the view option on Spartan 14. From the build option on the program, the structures were minimize using molecular mechanic force field (MMFF) option to remove any strain present in the molecular structure. In addition this ensures a well define conformer relationship between the compounds under study

(Viswanadhan, Ghose, Revankar, & Robins, 1989). From the set up calculation option on Spartan 14, the calculation was set to equilibrium geometry at the ground state using density functional theory at B3LYP. After optimization, Spartan molecular descriptor were obtained from the display-output and display-properties option on Spartan 14 GUI. The fully optimized 3D structure without symmetry restrictions, were saved as SD file through the file option on the Spartan 14 GUI. The fully optimized 3D structure in SD file were then open with ChemBio 3D ultra 12.0 to calculate molecular topological descriptors using the calculation option on the ChemBio 3D ultra 12.0 GUI

Splitting of data-set into modelling sets and evaluation test sets

The data set was divided into two sets, the modelling set and test set. The modelling set is used in developing the model, it contains eighty percent of the entire data set. While the test set which constitutes the remaining twenty percent of the whole data set were not used in the construction of the model but to ascertain the predictive ability of the model (Tropsha, 2010).

Data Division

In order to obtain validated QSAR models the dataset was divided into training and test sets. Ideally, this division should be performed such that points representing both training (80% of compounds) and test sets (20% percent of compounds) are distributed within the whole descriptor space occupied by the entire dataset, and each point of the test set is close to at least one point of the training set. This partitioning ensures that a similar principle can be employed for the activity prediction of the test set. Kennard-Stone Algorithm will be applied for dividing Dataset into a Training and Test set (Rajer-Kanduč, Zupan, & Majcen, 2003), (Wu et al., 1996), (Kennard & Stone, 1969).

Objective function = $\sum_{i=1}^{K+1}\{[\mu(i)_{train} - \mu(i)_{test}] + [\sigma(i)_{train} - \sigma(i)_{test}]\}$

K is the number of inputs and μ and σ are mean and standard deviation of the input or output variable, respectively. With this technique, all objects are considered as candidates for the training set. The selected candidates are chosen sequentially. KS algorithm can be summarized as follows: First, the KS algorithm takes the pair of samples with the largest Eucledian distance of *x*-vectors (predictors) and then it sequentially selects a sample to maximize the Eucledian distance between *x*-vectors of already selected samples and the remaining samples. This process is repeated until the required number of samples is achieved. For each pair of samples *i* and *j,* the Eucledian distance in *x* space is defined as (Wu *et al*., 1996; (Saptoro, Tadé, & Vuthaluru, 2012); Kennard *et al*., 1969). The algorithm employs Eucledian distance ED_x (p q), between the *x* vectors of each pair (p,q) of samples in order To ensure a uniform distribution of such a subset along the *x* data space

$$\boldsymbol{ED_x(p,q)} = \sqrt{\sum_{j=1}^{N}[x_p(j) - x_q(j)]2} \qquad \boldsymbol{p,q} \in [\mathbf{1}, \boldsymbol{M}]$$

N is the number variables in *x* and M is the number of samples, while x_p *(j)* and x_q *(j)* are the *j* the variable for samples *p* and *q* , respectively.

Model development

Multiple Linear Regression was used to show the relationship between the dependent variable Y (pGI_{50}) and independent variable X (atomic descriptors). The model is fit such that sum-of-squares difference between the experimental and predicted values of set biological activity is minimized. In regression analysis, contingent mean of dependant variable (pGI_{50}) Y relies on (descriptors) X.

Evaluation of the QSAR Model

The QSAR models developed were validated by reviewing some of its parameters like: R^2 (the squared correlation coefficient); F test (Fischer's Value) for statistical significance; Q^2 (cross-validated correlation coefficient); pred R^2 (R^2 for external test set)

Validation of the QSAR model

The ability of a QSAR equation to predict the bioactivity of unknown compounds was determined using the leave-one-out cross-validation method. The cross-validation regression coefficient (Q_{cv^2}) was calculated with the following equation:

$$q_{CV}^2 = 1 - {}^{PRESS}/_{TOTAL} = 1 - \frac{\sum_{i=1}^{n} (y_{exp} - y_{pred})^2}{\sum_{i=1}^{n} (y_{exp} - \bar{y})^2}$$

where y_{pred}, y_{exp}, and $\tilde{y}$ are the predicted, experimental, and mean values of experimental activity, respectively. It has been reported that high estimation of statistical attributes is not enough to justify the ability of a model, and so to assess the predictive capacity of the new QSAR model, the method depicted by Golbraikh and Tropsha (Golbraikh & Tropsha, 2002) and Roy et al. (Roy, Kar, & Ambure, 2015) were utilized. The coefficient of determination for the test set R_{test}^2, was calculated through the accompanying mathematical statement

$$R_{Test}^2 = 1 - \frac{\sum \left(Y_{pred_{test}} - Y_{Test}\right)^2}{\sum \left(Y_{pred_{test}} - \overline{Y}_{Training}\right)^2}$$

where $Y_{pred_{test}}$ and Y_{Test} are the predicted value founded on the QSAR equation (model response) and experimental activity values, respectively, of the external test set compounds. $\bar{Y}_{Training}$ is the average activity value of the training set compounds (Tropsha, Gramatica, & Gombar, 2003). Additional assessment of the predictive ability of the QSAR model for the test set compounds was done by determining the value of ($\mathbf{r_m^2}$), using the rm^2 metric calculator developed by Roy et al. (Roy et al., 2013).

Evaluation of the applicability domain of the model

The applicability domain of the QSAR model is imperative in establishing the model ability to make predictions within the chemical space for which it was developed (Tropsha et al., 2003). The leverage tactic was used in unfolding the applicability domain of the QSAR models (Gramatica, Giani, & Papa, 2007). Leverage of a given chemical compound $\boldsymbol{h_i}$, is defined as: $\boldsymbol{h_i = x_i(X^TX)^{-1}x_i^T}$ $\boldsymbol{(i = 1,\ldots,m)}$, where $\boldsymbol{x_i}$ is the descriptor row-vector of the query compound $\boldsymbol{i}$, and $\mathbf{X}$ is the $\boldsymbol{n \times k}$ descriptor matrix of the training set compounds used to develop the model. As a prediction tool, the warning leverage (h*) which is the limit of normal values for X outliers and is defined as: $\boldsymbol{h^* = 3(k+1)/n}$, where $\boldsymbol{n}$ is the number of training compounds, $\boldsymbol{k}$ is the number of descriptors in the model. The test compounds with leverages $\boldsymbol{h_i < h^*}$ are considered to be reliably predicted by the model. The Williams plot, a plot of standardized residuals versus leverage values, is utilized to translate the relevance area of the model in terms of chemical space. The domain of unfailing prediction for external test set molecules' is defined as compounds which have leverage values within the threshold $\left(\boldsymbol{h_i < h^*}\right)$ and standardized residuals no greater than 3α (3 standard deviation units), hence they are accepted as Y outlier. Test set compounds where $\left(\boldsymbol{h_i > h^*}\right)$ are thought to be

unreliably anticipated by the model because of considerable extrapolation. For the training set, the Williams plot is utilized to recognize compounds with the best structural influence $\left(h_i > h^*\right)$ in developing the model

Results and Discussion

A QSAR analysis was performed to explore the structure–activity relationship of different 112 compounds with different organic moiety acting as anticancer. In a QSAR study, generally, the quality of a model is expressed by its fitting and prediction ability (Table 2).

QSAR on K-562 cell line dataset

K-562 cell line

$$pGI_{50} = -5.524\,(\boldsymbol{Methanal}) + 5.514\,(\boldsymbol{PSA}) - 6.097(\boldsymbol{ATS7e}) - 2.255\,(\boldsymbol{ATSC5c}) - 1.219\,(\boldsymbol{naasN}) - 2.813(\boldsymbol{minHBint7}) - 2.162\,(\boldsymbol{minHBint10}) + 1.482\,(\boldsymbol{maxHBint5}) - 4.484\,(\boldsymbol{hmax}) + 7.419(\boldsymbol{MDEC-11}) + 8.762(\boldsymbol{MDEC-23}) - 3.254(\boldsymbol{RDF155v}) + 6.467$$

$$N_{train} = 90,\ R^2_{train} = 0.915,\ R^2_{adjusted} = 0.902,\ F_{train} = 69.298,\ Q^2_{LOO} = 0.845,\ Outliers > 3.0 = 5, N_{test} = 22$$

N is the number of compounds, R^2 is the squared correlation coefficient, Q^2_{LOO}, is the squared cross-validation coefficients for leave one out, F is the Fisher F statistic, and RMSE is the root mean square error.

The built model was used to predict the test set data, and the results are presented in Table 1. The predicted pGI_{50} values for the compounds in the training and test sets for K-562 leukaemia cell line were plotted against the experimental pGI_{50} values in Figure 1, Likewise, the plot of the residuals values for both the training and test sets against the experimental pGI_{50} estimations is presented in Figure 2. As can be seen from Table 1, Figure 1 and Figure

2, the computed values for the pGI_{50} are in great concurrence with those of the test set, hence the model did not demonstrate any relative and systematic error, since the arrangement of the residuals on both sides of zero is arbitrary.

The QSAR of K-562 model in this literature was reported to have an R^2 value of 0.902 and Q^2_{CV} value of 0.845, while for the external validation R^2_{pred}, Q^2F_1 and Q^2F_2 values were reported in Table 3 as 0.672, 0.916 and 0.581. The result justifies that the classic metric test for 100% developed by Roy et al. (Roy, Kar, et al., 2015a) for a QSAR model biasness test is good and in well agreement with other standards stated by Tropsha and Golbraikh (Golbraikh & Tropsha, 2002).

QSAR model validation

The genuine value of QSAR models is not only their capacity to reproduce known activities of a compound, confirmed by their fitting power (R^2), but for the most part is their potential for predicting biological activity. Therefore, the internal consistency of the training set was confirmed by using leave-one-out (LOO) cross-validation method to guarantee the strength of the model (Supratik Kar, 2010).

The leverages for every compound in the data set were plotted against their standardized residuals, leading to discovery of outliers and influential chemicals in the models. Figure 3 shows the Williams plot of K-562 dataset. The applicability domain is established inside a squared area within ±3 bound for residuals and a leverage threshold **h*** ($\boldsymbol{h}^* = \boldsymbol{3p^o/n}$, where $\boldsymbol{p^o}$ is the number of model parameters and n is the number of compounds. The Williams plot for the training set shown in Figure 3, establishes applicability domain of the model within **± 3d** and a leverage threshold **h*** = 0.433.

The Williams plot for K-562 data set shows two group of outliers, one of which is related to the difference in the structures of the compounds used as training set and the other directly related to the wide variations in their experimental data. Compound with these identification number (ID: 15, 37, 65, 70 and 72) from Table 1, were identified as outliers within the plot because of their incorrect experimental data used, the remaining three compounds (ID: 10, 64 and 84) which influences the scope of the model positively are structurally different from other compounds in the model (Roy, Kar, & Ambure, 2015). All these compounds have their leverage values greater than the warning leverage (**h***) value, their high leverages are responsible for swaying the performance of the model.

In order to assess the robustness of the model, the Y-randomization test was applied in this study. Y-randomization test confirms whether the model is obtained by chance correlation, and is a true structure–activity relationship to validate the adequacy of the training set molecules.

The new QSAR models (after several repetitions) was reported to have low R^2 and Q^2_{LOO} values for K-562 activity (Table 3). In the event that the opposite happens, then an adequate QSAR model can't be gotten for that particular modelling system and information. The after effects of Table 3 show that an adequate model is gotten by GA–MLR system, and the model created is measurably noteworthy and vigorous. In Table 2, statistical parameters such the mean absolute error (MAE) and root mean square error (RMSE) for training and test set were recorded to investigate the overall error included in the model (Roy, Kar, et al., 2015a). The slope of the models and their coefficients are also presented (Table 2), which validate the model strength and supports other results presented in Tables 3.

To examine the relative importance, and the contribution of each descriptor in the model, for each descriptor the value of the mean effect (MF) was calculated. This calculation was performed with the equation below

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_{j}^{m} \beta_j \sum_{i}^{n} d_{ij}}$$

MF_j represents the mean effect for the considered descriptor j, β_j is the coefficient of the descriptor j, d_{ij} stands for the value of the target descriptors for each molecule, and m is the descriptor's number in the model (Dimić, Mercader, & Castro, 2015).

The MF value provides important information on the effect of the molecular descriptors in the developed model, the signs and the magnitude of these descriptors combined with their mean effects reveals their individual strength and direction in influencing the activity of a compound. The mean effect values are presented in Table 4. The molecular edge descriptor (MEDC-23) (Liu, Cao, & Li, 1998), polar surface area (PSA) and maximum hydrogen electropological state (hmax) (Hall & Kier, 1995) were found to have the most pronounce effect on the model. The mean effects of MEDC-23 (-3.918) and PSA (-3.887) were negatively correlated with activities of the model, while that of hmax (2.978) contributes positively to the model. Hereby indicating that high polar surface area and molecules edge of the type (MEDC-23) were responsible for hindering the potency of these compounds on K-562 cancer cell line.

Interpretation of Descriptors in model

Methanal fragment count is a 2D molecular descriptor utilized by the model to predict the 50% reduction in proliferation of K-562 leukaemia cell line. This descriptor defines the number formaldehyde fragment that is within a molecule, its mean effect (0.184) to the model

though a little insignificant in magnitude is positively correlated to the activity of the compounds.

The **polar surface area** (**PSA**) of a molecule is defined as the surface sum over all polar atoms, primarily oxygen and nitrogen, also including their attached hydrogens, it is a commonly used medicinal chemistry metric for the optimisation of a drug's ability to permeate cells. The mean effect of PSA (-3.887) reported in Table 4 is significantly high and its responsible for decreasing the bioactivity of most of the compounds used in developing the model. Hence in the design of a hypothetical new drug a significant decrease in this descriptor is needed to improve its activity.

ATS7e is a 2D autocorrelation molecular descriptor developed by Todeschini and Consonni (Todeschini & Consonni, 2009), which is defined as Broto-Moreau autocorrelation - lag 7 / weighted by Sanderson electronegativities.

$$ATSdw = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij}(w_i w_j),$$

where, w_i and w_j are the weights of the atoms i and j, $\omega \in \{m, p, e, v\}$, and δ_{ij} is Kronecker delta, that is, δ_{ij} =1 if the ij^{th} entry in the Topological Level Matrix is = d, and δ_{ij} = 0 otherwise (Broto & Devillers, 1990; Broto, Moreau, & Vandycke, 1984; Gilles Moreau & Pierre Broto, 1980; G Moreau & P Broto, 1980). ATS7e descriptor with mean effect (1.837) is found to be a significant descriptor which is positively correlated to the bioactivity of the compounds, hence by increasing the magnitude of the descriptor its activity is also increased. Other autocorrelation descriptor used in the model includes ATSC5c, which is defined as Centered Broto-Moreau autocorrelation - lag 5 / weighted by charges. This molecular descriptor is weighted by the charges on the molecule unlike ATS7e which is related to the polarization of the molecules caused by highly electronegative elements present in a

compound, the former has a mean effect of 1.427, which indicates the direction of the descriptor influences the activity positively when increased.

The E - state and the HE – state indices may be used as atomic parameters to generate other topological indices. naasN is a 2D Atom type electrotopological state descriptor, which is defined as the number of atom-type N:- descriptor present in a compound. It is an example of a combination of electronic, topological, and valence state information developed by Hall and Kier (Hall & Kier, 1995) to relate the importance of nitrogen atom type of the order in affecting the topological feature of the overall compound and how this in turn affects the activity of the compound as a direct result of this effect. The calculated effect (0.162) of the descriptor to the model was directly correlated to the activity of anticancer agents. Three other E-state descriptor used in the model are minHBint7, minHBint10, maxHBint5 and hmax, they are defined as Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7, Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 10, maxHBint5 and Maximum H E-State respectively. The mean effects of the descriptors are presented in Table 4, their values vary in magnitude and direction with maxHBint5 which is negatively correlated to the activity of the molecules. Their values are given as 1.658, 1.286, -0.658 and 2.978 respectively, hmax had the highest value (2.978) while maxHBint5 (-0.658) which is negatively correlated to the activity of the molecules contributes the least to the model. Roy and Mitra (Ojha, Mitra, Das, & Roy, 2011) showed that the importance of the ability to encode the topology and electronic environment of molecular fragments in unison portrayed the E-state indices as an indispensable tool in the field of QSAR studies.

MDEC-11 and MDEC-23 are 2D Molecular distance edge descriptor developed by Liu et al. (Liu et al., 1998), MDEC-11 with a mean effect of -0.459, is defined as Molecular distance edge between all primary carbons. The magnitude of MDEC-11 descriptor in the model

shows that a decrease in the bond length of all primary carbons present in a potent anticancer agent increase the bioactivity of the molecule, while MDEC-23 descriptor defined as molecular distance edge between all secondary and tertiary carbons was reported with the mean effect of -3.918. The mean effect of MDEC-23 contributes the most in decreasing the activity of the molecules, its effect when compared to all other descriptor in the model is the most significant, hence the decrease in secondary and tertiary Carbon atoms in a molecule would greatly increase the activity of an anticancer agent or hypothetical compounds with potent effect on K-562 leukaemia cell line.

Radial distribution function is a 3D coordinates of the atoms of molecules transformed into a structure code that has a fixed number of descriptors irrespective of the size of a molecule, Formally, the Radial Distribution Function of an ensemble of *N* atoms can be interpreted as the probability distribution to find an atom in a spherical volume of radius *r*. RDF155v is one of the descriptor used in the model, it has a mean effect of 0.373 contributing very little to the overall effect of the descriptor to the model. The radial distribution function - 155 / weighted by relative van der Waals volumes as defined describes how the van der waal volume of the descriptor affects the activity of the molecule. Here the value of the mean effects implores the increase of the RDF-155 weighted by the molecular volume in influencing the positive action of anticancer agents to their target site.

Ligand Base Drug Design

Twenty-three (23) compounds were designed using the information derived from the model. The molecular descriptor PSA and hmax were the principal descriptor used in our design and this is owed to their significant mean effect on the model compared to other descriptors. We selected two lead compounds from our test set with low residual value from their predicted

pGI_{50}. This was done in order to minimize the possibility of statistical error in our design. The compound CAMPTOTHECIN ANALOGUE 3, was used to design 12 new analogues, while COLCHICINE DERIVATIVE was used as a lead compound in designing the remaining 11 compounds. The MF value of PSA descriptor suggest the removal of hetero atoms such as oxygen and nitrogen in order to reduce the polar surface area of the compounds, while hmax supports the conversion of unsaturated carbons to saturated carbons or replacing the (-O-) alkoxy groups with methylene carbons ($-CH_2-$), thereby making more room for hydrogen atoms and increasing the possibility of hydrogen bond formation with the receptor.

The pGI_{50} result of the designed analogues of CAMPTOTHECIN ANALOGUE 3 (**CA**) and COLCHICINE DERIVATIVE (**CD**) presented in Table 5 and 6 shows a correlation between the activity of the newly designed compounds with the mean effect values of hmax and PSA. pGI50 of more than 90% of the designed compounds were more than the lead compounds, thereby justifying the contribution of PSA and hmax descriptor to the activity of anticancer drugs in mitigating K562 cancer cell line.

Conclusion

For the robustness and statistical significance of the developed model, an initial division of dataset was done for training and test set compounds using Kennard-stone algorithm, before using GFA-MLR tool for building the model. The model is statistically robust both internally (Q^2: 0.845) and externally (Q^2_{F1}: 0.9397; Q^2_{F2}: 0.6862, R^2_{pred}: 0.6722) and satisfy the criteria of acceptable QSAR model proposed by different groups. The model indicates the importance of hydrogen bonding parameters (minHBint7, minHBint10, maxHBint5 and hmax), it indicates that a decrease in hydrogen bonding potentials of path length 7 and 10, as

well a decrease in the total polar surface area (PSA) for any compound is required to improve the pGI_{50} of anticancer agents.

Acknowledgment

We would like to acknowledge the National Cancer institute for providing the material data used for the QSAR study in the website (https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data).

Table 1: Chemical Names of Dataset with NSC numbers and their pGI_{50} values on K-562 Cell Lines

Serial Number (ID)	NAME	NSC	K-562 (Experimental pGI_{50})	K-562 (Predicted pGI_{50})	Residual	Standardized residual
1	11-FORMYL-20(RS)-CAMPTOTHECIN	606172	5.7	4.808	0.892	1.592
2	11-HYDROXYMETHYL-20(RS)-CAMPTOTHECIN	606173	5.6	6.165	-0.565	-1.009
3	14-CHLORO-20(S)-CAMPTOTHECIN HYDRATE	643833	5.7	6.521	-0.821	-1.466
4	2'-DEOXY-5-FLUOROURIDINE	27640	6.1	4.809	1.291	2.305
5	3-HP	95678	5.7	5.888	-0.188	-0.336
6	5,6-DIHYDRO-5-AZACYTIDINE	264880	5.5	5.571	-0.071	-0.127
7	5-AZA-2'-DEOXYCYTIDINE	127716	4[a]	4.243	-0.243	-0.596
8	5-AZACYTIDINE	102816	6.1	5.289	0.811	1.448
9	5-HP	107392	5.3	5.530	-0.230	-0.411

10	7-CHLOROCAMPTOTHECIN	249910	7.3[$]	6.897	0.403	0.720
11	9-AMINO-20-(R,S)-CAMPTOTHECIN	629971	7.5	7.307	0.193	0.345
12	ACIVICIN	163501	5.5 [a]	4.490	1.010	2.478
13	ALLOCOLCHICINE	406042	8 [a]	6.869	1.131	2.774
14	ALPHA-TGDR	71851	4.1	4.996	-0.896	-1.599
15	AMINOPTERIN DERIVATIVE1	132483	6.4* [a]	8.250	-1.850	-4.539
16	AMINOPTERIN DERIVATIVE2	184692	8	8.520	-0.520	-0.929
17	AMINOPTERIN DERIVATIVE3	134033	7.6	8.334	-0.734	-1.311
18	AMONAFIDE	308847	5.4	5.671	-0.271	-0.484
19	AN ANTIFOL	623017	7.6	7.344	0.256	0.457
20	ANTHRAPYRAZOLE DERIVATIVE	355644	6.7	5.929	0.771	1.377
21	APHIDICOLIN GLYCINATE	303812	5.3	5.744	-0.444	-0.793
22	ARA-C	63878	4.6	5.422	-0.822	-1.467
23	ASALEY	167780	5.2	5.811	-0.611	-1.498
24	AZQ	182986	5.3	5.203	0.097	0.174
25	BAKER'S SOLUBLE ANTIFOL	139105	6.8	6.653	0.147	0.262
26	BCNU	409962	4.3	3.858	0.442	0.789

27	BETA-TGDR	71261	6.2	5.348	0.852	1.521
28	BISANTRENE HCL	337766	7.3	6.931	0.369	0.659
29	BREQUINAR	368390	6.9 [a]	7.050	-0.150	-0.368
30	BUSULFAN	750	3.6 [a]	3.201	0.399	0.978
31	CAMPTOTHECIN	94600	7.3 [a]	6.766	0.534	1.311
32	CAMPTOTHECIN ANALOG	295500	6	6.655	-0.655	-1.169
33	CAMPTOTHECIN ANALOG2	606985	7.5	6.622	0.878	1.567
34	CAMPTOTHECIN ANALOG3	295501	7.5 [a]	7.019	0.481	1.179
35	CAMPTOTHECIN BUTYLGLYCINATE ESTER HYDROCHLORIDE	606499	6.3	6.528	-0.228	-0.408
36	CAMPTOTHECIN ETHYLGLYCINATE ESTER HYDROCHLORIDE	606497	6.1	6.466	-0.366	-0.654
37	CAMPTOTHECIN GLUTAMATE HCL	610459	6.5*[a]	8.558	-2.058	-5.049
38	CAMPTOTHECIN HEMISUCCINATE SODIUM SALT	610456	6.3	6.431	-0.131	-0.234
39	CAMPTOTHECIN LYSINATE HCL	610457	7.2 [a]	6.366	0.834	2.046
40	CAMPTOTHECIN PHOSPHATE	610458	6.2	4.868	1.332	2.379
41	CAMPTOTHECIN, 9-METHOXY-	176323	7.3	7.002	0.298	0.532
42	CAMPTOTHECIN, ACETATE	95382	5.5	6.050	-0.550	-1.349
43	CAMPTOTHECIN, HYDROXY-	107124	7.4	7.153	0.247	0.442

44	CAMPTOTHECIN, NA SALT	100880	7.3	7.424	-0.124	-0.222
45	CAMPTOTHECIN,20-O-((4-(2-HYDROXYETHYL)-1-PIPERAZINO)OAC	374028	6.1 [a]	7.211	-1.111	-2.726
46	CAMPTOTHECIN-20-O-(N,N-DIMETHYL)GLYCINATE HCL	618939	7.3 [a]	7.767	-0.467	-1.147
47	CCNU	79037	4.6	4.393	0.207	0.370
48	CHLORAMBUCIL	3088	4	4.608	-0.608	-1.086
49	CHLOROZOTOCIN	178248	3.2	2.824	0.376	0.671
50	CLOMESONE	338947	3.3 [a]	2.779	0.521	1.277
51	COLCHICINE	757	7.2	7.402	-0.202	-0.362
52	COLCHICINE DERIVATIVE	33410	7.9 [a]	7.947	-0.047	-0.116
53	CYANOMORPHOLINODOXORUBICIN	357704	8.3	8.023	0.277	0.494
54	CYCLOCYTIDINE	145668	3.4 [a]	4.465	-1.065	-2.612
55	CYCLODISONE	348948	4.1	3.032	1.068	1.906
56	DAUNORUBICIN	82151	7	6.565	0.435	0.777
57	DEOXYDOXORUBICIN	267469	7.4	7.731	-0.331	-0.591
58	DIANHYDROGALACTITOL	132313	3.9	4.369	-0.469	-0.838
59	DICHLORALLYL LAWSONE	126771	5.7	5.962	-0.262	-0.468

60	DOLASTATIN 10	376128	10.2	9.797	0.403	0.720
61	DOXORUBICIN	123127	7	7.485	-0.485	-0.865
62	FLUORODOPAN	73754	3.4 [a]	4.587	-1.187	-2.912
63	FTORAFUR (PRO-DRUG)	148958	3	4.029	-1.029	-1.838
64	GLYCINATE	364830	7	7.718	-0.718	-1.282
65	GUANAZOLE	1895	2.2*[a]	4.738	-2.538	-6.226
66	HEPSULFAM	329680	3.4	3.245	0.155	0.276
67	HYCANTHONE	142982	5.3	6.207	-0.907	-1.619
68	HYDROXYUREA	32065	3	3.119	-0.119	-0.213
69	INOSINE GLYCODIALDEHYDE	118994	4 [$]	3.228	0.772	1.378
70	L-ALANOSINE	153353	4.8 *[a]	6.127	-1.327	-3.256
71	MACBECIN II	330500	7.1 [a]	8.458	-1.358	-3.331
72	M-AMSA	249992	6*	5.616	0.384	0.686
73	MAYTANSINE	153858	7.8	8.709	-0.909	-1.624
74	MELPHALAN	8806	4.3	4.551	-0.251	-0.449
75	MENOGARIL	269148	5.9	5.972	-0.07	-0.128

					2	
76	METHOTREXATE	740	7.5	6.725	0.775	1.383
77	METHOTREXATE DERIVATIVE	174121	9.4	9.272	0.128	0.229
78	METHYL CCNU	95441	4.4	4.647	-0.247	-0.441
79	MITOMYCIN C	26980	5.6	5.204	0.396	0.707
80	MITOXANTRONE	301739	6.9	6.927	-0.027	-0.048
81	MITOZOLAMIDE	353451	4.1	4.052	0.048	0.086
82	MORPHOLINODOXORUBICIN	354646	8.6	7.752	0.848	1.514
83	N-(PHOSPHONOACETYL)-L-ASPARTATE (PALA)	224131	4	4.822	-0.822	-1.468
84	N,N-DIBENZYL DAUNOMYCIN	268242	5.2 $	5.289	-0.089	-0.159
85	NITROGEN MUSTARD	762	5.2	3.963	1.237	2.209
86	OXANTHRAZOLE	349174	5.9	6.560	-0.660	-1.178
87	PCNU	95466	3.8	4.014	-0.214	-0.382
88	PIPERAZINE DRUGSMAINATOR	344007	3.7	3.501	0.199	0.356
89	PIPERAZINEDIONE	135758	5.6	5.637	-0.037	-0.067
90	PIPOBROMAN	25154	3.9	4.267	-0.367	-0.656
91	PORFIROMYCIN	56410	4.8	4.962	-0.162	-0.290

92	PYRAZOFURIN	143095	6.3	6.082	0.218	0.389
93	PYRAZOLOACRIDINE	366140	6.7	5.895	0.805	1.438
94	PYRAZOLOIMIDAZOLE	51143	2.5	2.649	-0.149	-0.267
95	RHIZOXIN	332598	8	7.624	0.376	0.672
96	RUBIDAZONE	164011	6.4	6.730	-0.330	-0.589
97	SPIROHYDANTOIN MUSTARD	172112	3.7	4.093	-0.393	-0.701
98	TAXOL	125973	8.4	8.746	-0.346	-0.617
99	TEROXIRONE	296934	4.5 [a]	3.977	0.523	1.283
100	TETRAPLATIN	363812	6	5.387	0.613	1.095
101	THIOCOLCHICINE	361792	7.6	7.981	-0.381	-0.680
102	THIOGUANINE	752	6.4	5.696	0.704	1.256
103	THIO-TEPA	6396	3.9	3.867	0.033	0.059
104	TRIETHYLENEMELAMINE	9706	5	4.411	0.589	1.051
105	TRIMETREXATE	352122	7.6	7.460	0.140	0.250
106	TRITYL CYSTEINE	83265	6.2	6.105	0.095	0.169
107	URACIL NITROGEN MUSTARD	34462	4.4	4.924	-0.524	-0.936
108	VINBLASTINE SULFATE	49842	9.3	8.995	0.305	0.544
109	VINCRISTINE SULFATE	67574	7	6.348	0.65	1.164

110	VM-26	122819	6.1	6.577	-0.4772	-0.852
111	VP-16	141540	4.7	5.277	-0.5777	-1.029
112	YOSHI-864	102627	2.7	3.441	-0.7411	-1.323

Where superscript **a** represent test sets for **K-562** leukaemia cell lines respectively, **$** represents compounds structurally different from all other compounds within the data set and * identifies compounds found outside the applicability domain

Table 2: Model External Validation Statistics

Test set Validation Information	**Name**	K-562
Model biasness test	**Systematic Error Result**	**Absent**
Classical Metrics (for 100% data)	**R^2Test(100% data)**	0.6722
	R0^2Test(100% data)	0.6614
	Q2F1(100% data)	0.9161
	Q2F2(100% data)	0.5816
	Scaled Avg.Rm^2(100% data)	0.5591
	Scaled DeltaRm^2(100% data)	0.1417
	CCC(100% data)	0.7961
Classical Metric (after removing 5% data with high residuals)	**R^2Test(95% data)**	0.7390
	R0^2Test(95% data)	0.7205
	Q2F1(95% data)	0.9397
	Q2F2(95% data)	0.6862
	ScaledAvgRm2(95% data)	0.6509
	ScaledDeltaRm2(95% data)	0.0601
	CCC(95% data)	0.8507
Error-based metrics (for 100% data)	**RMSEP(100% data)**	1.1011
	SD(100% data)	0.6363
	SE(100% data)	0.1357
	MAE(100% data)	0.9088
BASIC DATA STRUCTURE INFORMATION		
	N Compound Test	**22**
RESULT (MAE-based criteria applied on 95% data)	**Prediction Quality**	**MODERATE**

Table 3. R^2_{Train} and Q^2_{LOO} values after several Y-randomization tests for K-562 cell line.

Iteration	R	R^2	Q^2
Random 1	0.287	0.082	-0.434
Random 2	0.359	0.129	-0.176
Random 3	0.313	0.098	-0.161
Random 4	0.256	0.065	-0.325
Random 5	0.375	0.141	-0.049
Random 6	0.164	0.027	-0.221
Random 7	0.357	0.127	-0.218
Random 8	0.317	0.100	-0.326
Random 9	0.255	0.065	-0.173
Random 10	0.381	0.145	-0.169
Random Models Parameters			
Average R :	0.306		
Average R^2 :	0.098		
Average Q^2 :	-0.225		
cRp^2 :	0.766		

Table 4. Specification of entered descriptors in genetic algorithm multiple regression model of K-562.

Descriptors	Definition	Descriptor Type	P-value	VIF	MF
Methanal	Number of Methanal group	2D	1.09E-14	1.345	0.184
PSA	Polar Surface Area	2D	2.01E-12	4.847	-3.887
ATS7e	Broto-Moreau autocorrelation - lag 7 / weighted by Sanderson electronegativities	2D	7.24E-08	11.141	1.837
ATSC5c	Centered Broto-Moreau autocorrelation - lag 5 / weighted by charges	2D	9.63E-06	1.362	1.427
naasN	Count of atom-type E-State: :N:-	2D	4.20E-06	1.217	0.162
minHBint7	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 7	2D	4.60E-06	1.848	1.658
minHBint10	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 10	2D	0.000499	1.097	1.286
maxHBint5	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 5	2D	3.32E-05	2.61	-0.641
hmax	Maximum H E-State	2D	4.42E-11	2.342	2.978
MDEC-11	Molecular distance edge between all primary carbons	2D	2.26E-13	2.857	-0.459
MDEC-23	Molecular distance edge between all secondary and tertiary carbons	2D	3.81E-20	6.158	-3.918
RDF155v	Radial distribution function - 155 / weighted by relative van der Waals volumes	3D	9.30E-09	2.141	0.373

VIF: variance inflation factor

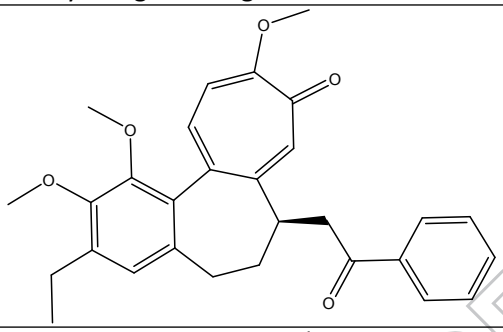
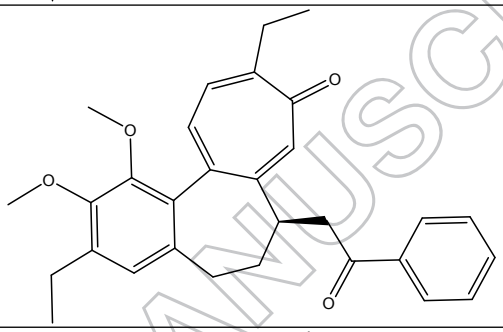
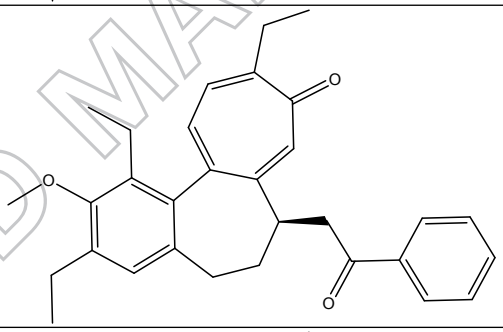
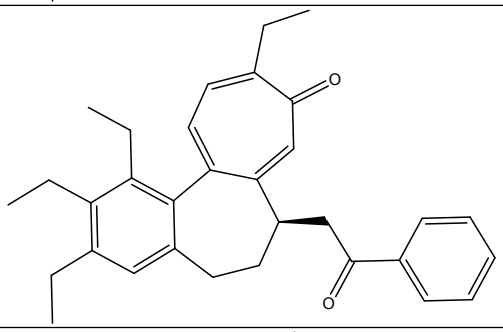
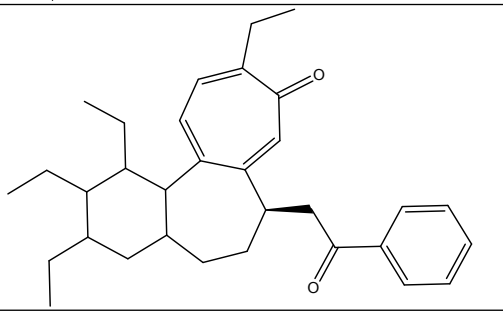
MF: mean effect

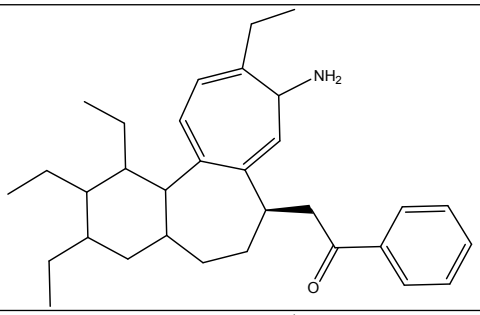
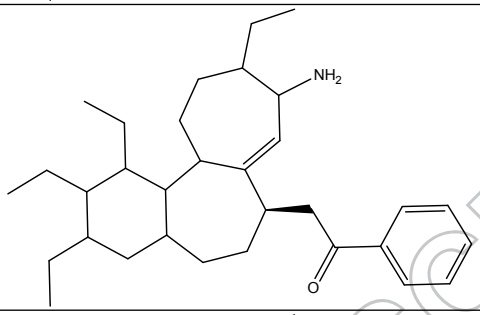
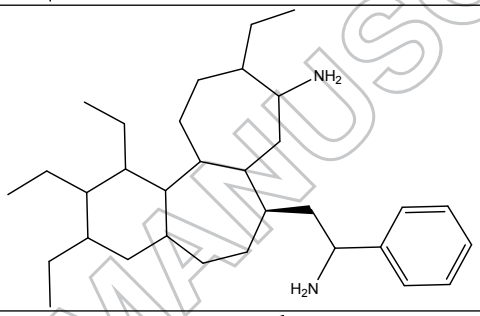
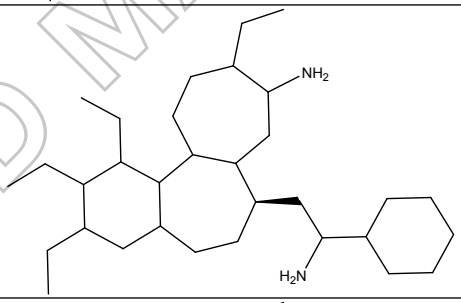
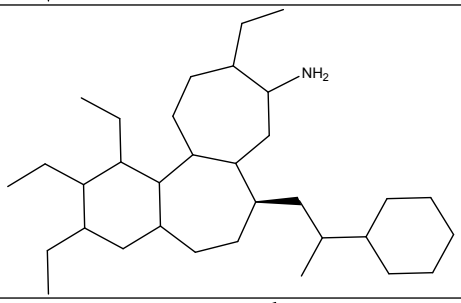
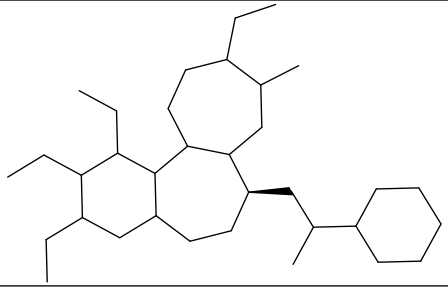
Table 5: Molecular Descriptor Values and Calculated pGI_{50} Values of The Newly Designed CAMPTOTHECIN ANALOGUE 3 (**CA**) and COLCHICINE DERIVATIVE (**CD**) Analogues.

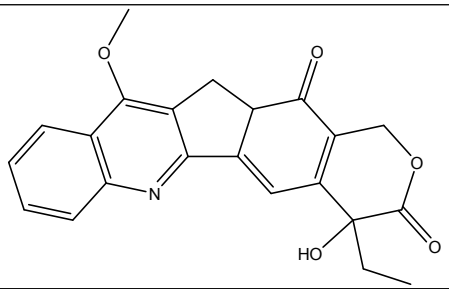
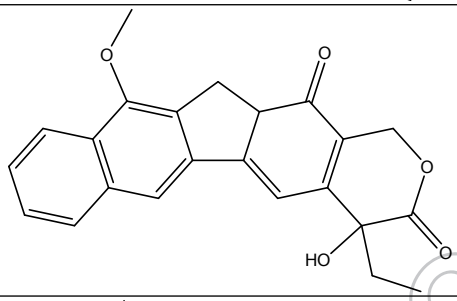
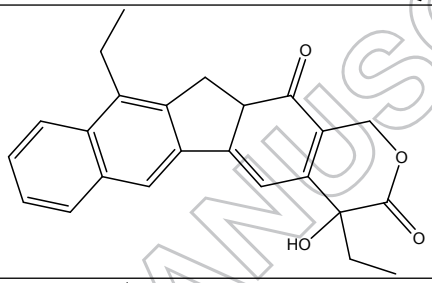
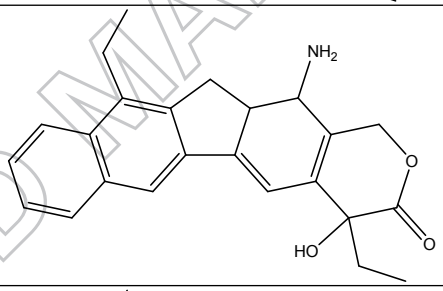
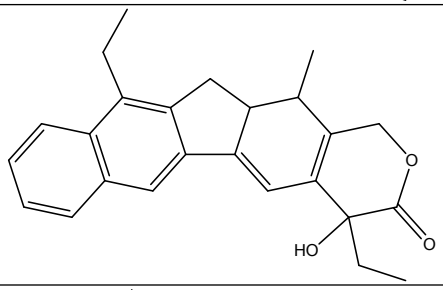
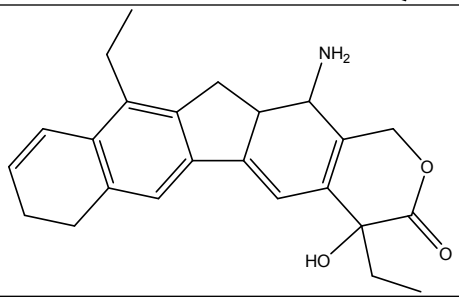
ID	Methanal	PSA	ATS7e	ATSC5c	naasN	minHBint7	minHBint10	maxHBint5	hmax	MDEC-11	MDEC-23	RDF155v	pGI50
CD1	0	0.849	0.317	0.752	0	0	0	0	0.764	0.398	0.61	0	12.391
CD2	0	0.722	0.335	0.854	0	0	0	0	0.727	0.398	0.724	0	12.516
CD3	0	0.595	0.388	0.825	0	0	0	0	0.697	0.398	0.87	0	12.972
CD4	0	0.469	0.418	0.84	0	0	0	0	0.674	0.398	1	0	13.302
CD5	0	0.469	0.536	1	0	0	0	0	0.609	0.398	1	0	12.513
CD6	0	0.592	0.604	0.913	0	1	0	0	0.573	0.398	1	0	10.322
CD7	0	0.592	0.73	0.904	0	0.756	0	0	0.506	0.398	1	0	10.560
CD8	0	0.715	0.871	0.719	0	0.269	0	0	0.432	0.398	1	1	9.244
CD9	0	0.715	0.956	0.71	0	0.222	0	0	0.198	0.398	1	0	13.182
CD10	0	0.357	0.981	0.787	0	0	0	0	0.183	0.623	1	0	13.242
CD11	0	0	1	0.937	0	0	0	0	0	1	1	0	14.437
CA1	0	1	0.004	0.307	0	0	0	1	1	0	0	0	8.263
CA2	0	1	0	0	0	0	0	0.982	0.993	0	0.179	0	10.553
CA3	0	0.873	0.027	0.621	0	0	0	0.939	0.97	0	0.319	0	9.553
CA4	0	0.996	0.094	0.633	0	0	0	0.956	0.86	0	0.319	0	10.314

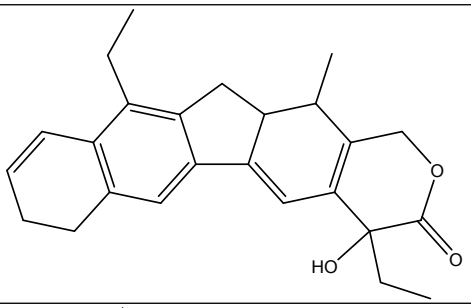
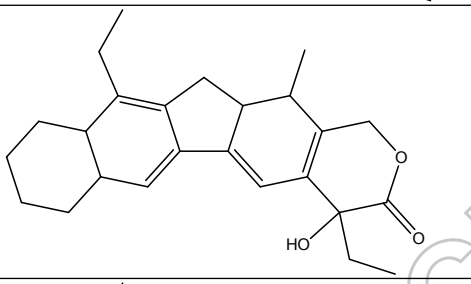
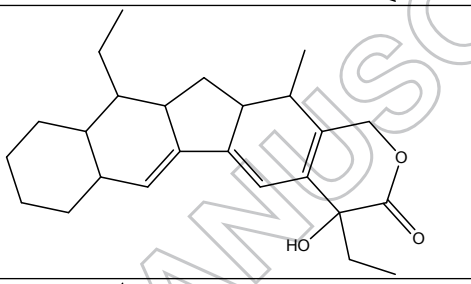
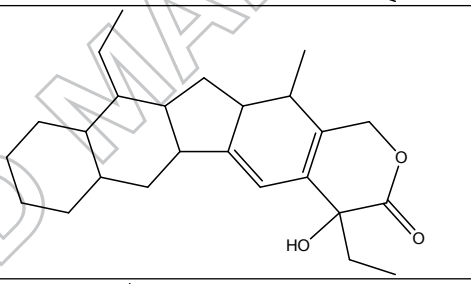
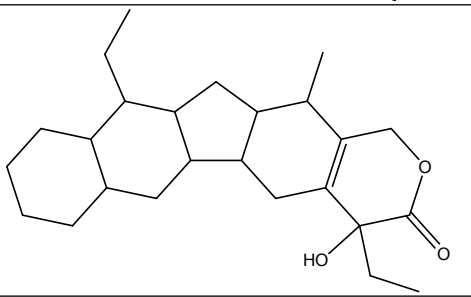
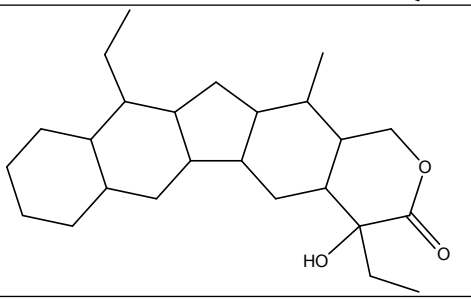
CA 5	0	0.639	0.115	0.691	0	0	0	0	0.814	0.165	0.319	0	8.100
CA 6	0	0.996	0.134	0.563	0	0	0	0.935	0.852	0	0.319	0	10.233
CA 7	0	0.639	0.155	0.679	0	0	0	0	0.806	0.165	0.319	0	7.919
CA 8	0	0.639	0.206	0.52	0	0	0	0	0.786	0.165	0.319	0	8.056
CA 9	0	0.639	0.242	0.576	0	0	0	0	0.769	0.165	0.319	0	7.787
CA 10	0	0.639	0.298	0.626	0	0	0	0	0.751	0.165	0.319	0	7.413
CA 11	0	0.639	0.348	0.554	0	0	0	0	0.713	0.165	0.319	0	7.441
CA 12	0	0.639	0.377	0.594	0	0	0	0	0.581	0.165	0.319	0	7.766

Table 6: 2D structure and the predicted pGI_{50} Values of The Newly Designed CAMPTOTHECIN ANALOGUE 3 (**CA**) and COLCHICINE DERIVATIVE (**CD**) Analogues.

	Compound ID	Newly designed drugs	Predicted PGI_{50}
1	**CD1**		12.391
2	**CD2**		12.516
3	**CD3**		12.972
4	**CD4**		13.302
5	**CD5**		12.513

6	**CD6**		10.322
7	**CD7**		10.560
8	**CD8**		9.244
9	**CD9**		13.182
10	**CD10**		13.242
11	**CD11**		14.437

12	**CA1**		8.263
13	**CA2**		10.553
14	**CA3**		9.553
15	**CA4**		10.314
16	**CA5**		8.100
17	**CA6**		10.233

18	**CA7**		7.919
19	**CA8**		8.056
20	**CA9**		7.787
21	**CA10**		7.413
22	**CA11**		7.441
23	**CA12**		7.766

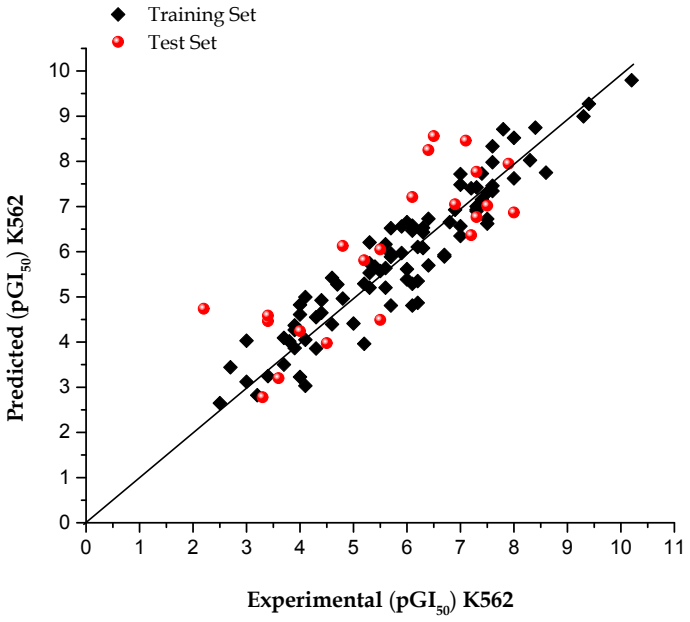


Figure 1: The predicted pGI_{50} against the experimental values for the training and test sets of K562 leukaemia cell line.

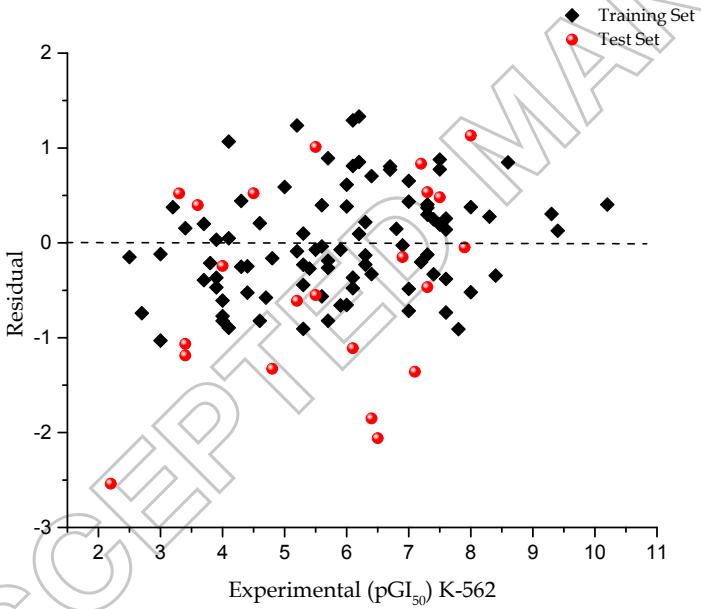


Figure 2: The Residuals against the Predicted pGI_{50} values for the training and test sets of K-562 leukaemia cell line.

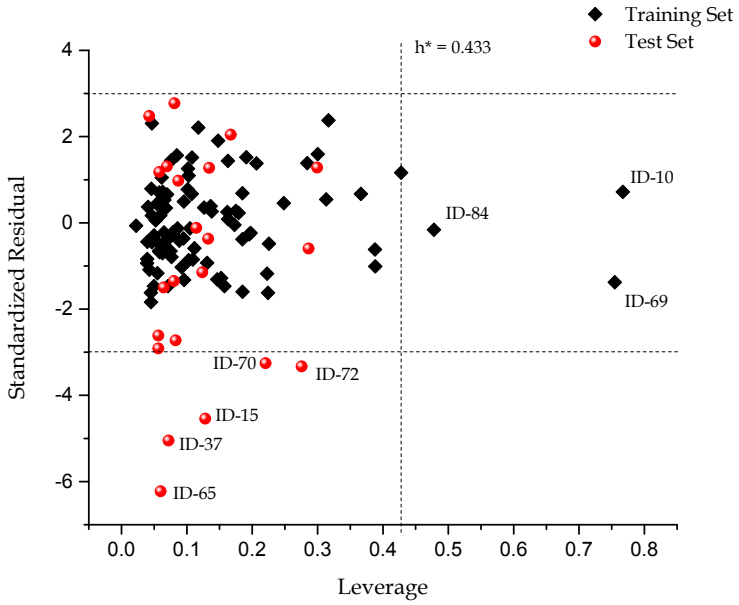


Figure 3: The Williams plot, the plot of the standardized residuals versus the activity (pGI_{50}) leverage value for K562 dataset

References

Andersson, L. C., Nilsson, K., & Gahmberg, C. G. (1979). K562—a human erythroleukemic cell line. *International journal of cancer, 23*(2), 143-147.

Broto, P., & Devillers, J. (1990). *Autocorrelation of properties distributed on molecular graphs*: Kluwer Academic Publishers: Dordrecht, The Netherlands.

Broto, P., Moreau, G., & Vandycke, C. (1984). Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *European Journal of Medicinal Chemistry, 19*(1), 71-78.

Chopade, S. M., Phadnis, P. P., Hodage, A. S., Wadawale, A., & Jain, V. K. (2015). Synthesis, characterization, structures and cytotoxicity of platinum(II) complexes containing dimethylpyrazole based selenium ligands. *Inorganica Chimica Acta, 427*, 72-80. doi: http://dx.doi.org/10.1016/j.ica.2014.11.017

Denton, P. (2001). Generating coursework feedback for large groups of students using MS Excel and MS Word. *University Chemistry Education, 5*(1), 1-8.

Dimić, D., Mercader, A. G., & Castro, E. A. (2015). Chalcone derivative cytotoxicity activity against MCF-7 human breast cancer cell QSAR study. *Chemometr. Intell. Lab. Syst., 146*, 378-384. doi: http://dx.doi.org/10.1016/j.chemolab.2015.06.011

Drexler, H. G. (2000). *The leukemia-lymphoma cell line factsbook*: Academic Press.

Evans, D. A. (2014). History of the Harvard ChemDraw project. *Angewandte Chemie International Edition, 53*(42), 11140-11145.

Fan, Y., Lu, H., An, L., Wang, C., Zhou, Z., Feng, F., . . . Zhao, Q. (2016). 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, 43*, 13-20.

Fatemi, M. H., Heidari, A., & Gharaghani, S. (2015). QSAR prediction of HIV-1 protease inhibitory activities using docking derived molecular descriptors. *Journal of Theoretical Biology, 369*, 13-22. doi: http://dx.doi.org/10.1016/j.jtbi.2015.01.008

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International journal of cancer, 127*(12), 2893-2917.

Golbraikh, A., & Tropsha, A. (2002). Beware of q 2! *J Mol Graph Model, 20*(4), 269-276.

Gramatica, P., Giani, E., & Papa, E. (2007). Statistical external validation and consensus modeling: A QSPR case study for K oc prediction. *J Mol Graph Model, 25*(6), 755-766.

Hall, L. H., & Kier, L. B. (1995). Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of chemical information and computer sciences, 35*(6), 1039-1045.

Hehre, W. J., & Huang, W. W. (1995). *Chemistry with Computation: An introduction to SPARTAN*: Wavefunction, Inc.

Iuliano, A., Strianese, D., Uccello, G., Diplomatico, A., Tebaldi, S., & Bonavolontà, G. (2012). Risk factors for orbital exenteration in periocular basal cell carcinoma. *American journal of ophthalmology, 153*(2), 238-241. e231.

Kar, S., & Roy, K. (2012). QSAR of phytochemicals for the design of better drugs. *Expert Opinion on Drug Discovery, 7*(10), 877-902. doi: 10.1517/17460441.2012.716420

Karimiani, E. G., Marriage, F., Merritt, A. J., Burthem, J., Byers, R. J., & Day, P. J. (2014). Single-cell analysis of K562 cells: an imatinib-resistant subpopulation is adherent and has upregulated expression of BCR-ABL mRNA and protein. *Experimental hematology, 42*(3), 183-191. e185.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics, 11*(1), 137-148.

Li, Z., Wan, H., Shi, Y., & Ouyang, P. (2004). Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and ChemSketch. *J Chem Inf Comput Sci, 44*(5), 1886-1890.

Liu, S., Cao, C., & Li, Z. (1998). Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, λ. *Journal of chemical information and computer sciences, 38*(3), 387-394.

Lozzio, C. B., & Lozzio, B. B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood, 45*(3), 321-334.

Marx, K. A., O'Neil, P., Hoffman, P., & Ujwal, M. (2003). Data mining the NCI cancer cell line compound GI50 values: identifying quinone subtypes effective against melanoma and leukemia cell classes. *Journal of chemical information and computer sciences, 43*(5), 1652-1667.

Moreau, G., & Broto, P. (1980). The auto-correlation of a topological-structure-a new Molecular Descriptor (Vol. 4, pp. 359-360): GAUTHIER-VILLARS 120 BLVD SAINT-GERMAIN, 75280 PARIS CEDEX O6, FRANCE.

Moreau, G., & Broto, P. (1980). AUTO-CORRELATION OF MOLECULAR-STRUCTURES, APPLICATION TO SAR STUDIES. *Nouveau Journal de Chimie-New Journal of Chemistry, 4*(12), 757-764.

News, B. (2003, 12 May, 2003). Cancer number one killer of men. *Health.* Retrieved 18-04, 2016, from http://news.bbc.co.uk/2/hi/health/3019801.stm

Ojha, P. K., Mitra, I., Das, R. N., & Roy, K. (2011). Further exploring r m 2 metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 194-205.

Organization, W. H. (2002). National cancer control programmes: policies and managerial guidelines.

Parkin, D. M., Boyd, L., & Walker, L. (2011). 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *British journal of cancer, 105*, S77-S81.

Rajer-Kanduč, K., Zupan, J., & Majcen, N. (2003). Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemometrics and Intelligent Laboratory Systems, 65*(2), 221-229.

Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., & Das, R. N. (2013). Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *Journal of computational chemistry, 34*(12), 1071-1082.

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems, 145*, 22-29. doi: http://dx.doi.org/10.1016/j.chemolab.2015.04.013

Roy, K., Kar, S., & Das, R. N. (2015a). *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*: Springer.

Roy, K., Kar, S., & Das, R. N. (2015b). *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*: Academic Press.

Sabet, R., Mohammadpour, M., Sadeghi, A., & Fassihi, A. (2010). QSAR study of isatin analogues as in vitro anti-cancer agents. *European Journal of Medicinal Chemistry, 45*(3), 1113-1118.

Saptoro, A., Tadé, M. O., & Vuthaluru, H. (2012). A modified Kennard-Stone Algorithm for optimal division of data for developing artificial neural network models. *Chemical Product and Process Modeling, 7*(1).

Siegel, R. L., Miller, K. D., & Jemal, A. (2015). Cancer statistics, 2015. *CA: a cancer journal for clinicians, 65*(1), 5-29.

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. D. (2012). Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorganic & Medicinal Chemistry, 20*(15), 4848-4855.

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. D. S. (2012). Chemoinformatics in anti-cancer chemotherapy: Multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *European Journal of Pharmaceutical Sciences, 47*(1), 273-279. doi: http://dx.doi.org/10.1016/j.ejps.2012.04.012

Supratik Kar, K. R. (2010). Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs *Indian Journal of Biochemistry & Biophysics, 48*, 111-122.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)* (Vol. 41): John Wiley & Sons.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics, 29*(6-7), 476-488.

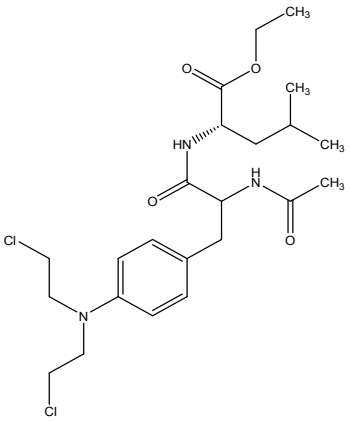
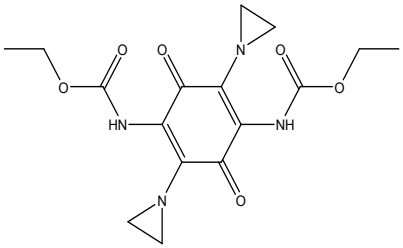
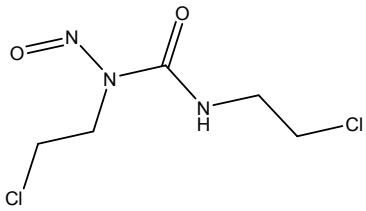
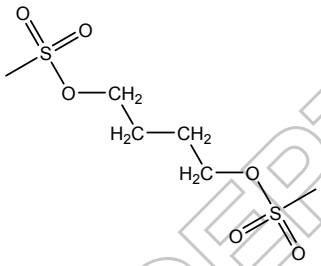
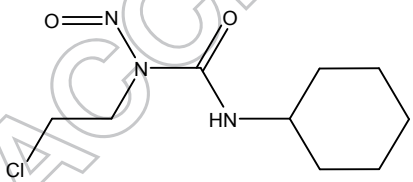
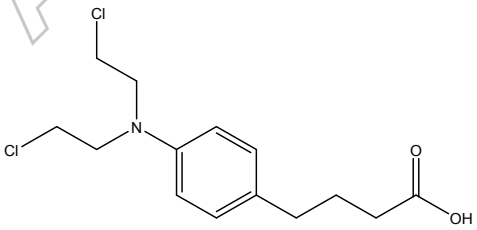
Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci, 22*(1), 69-77.

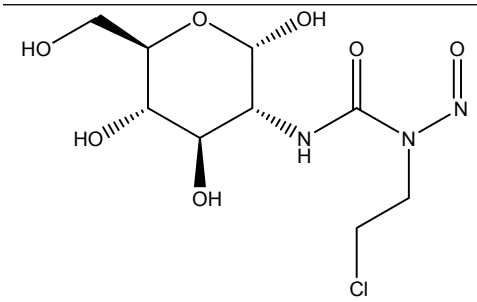
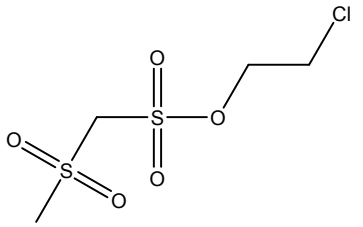
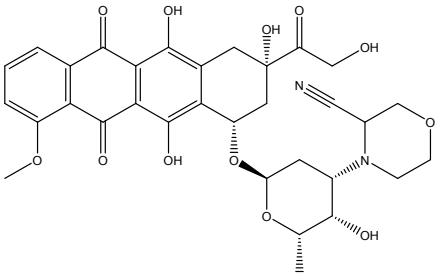
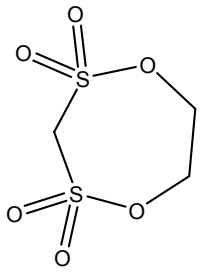
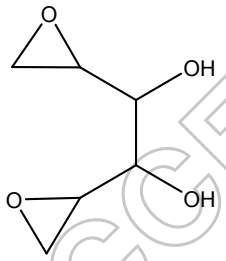
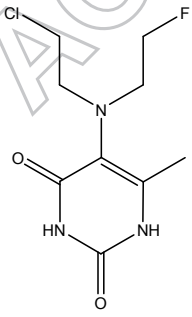
Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., & Robins, R. K. (1989). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of chemical information and computer sciences, 29*(3), 163-172.

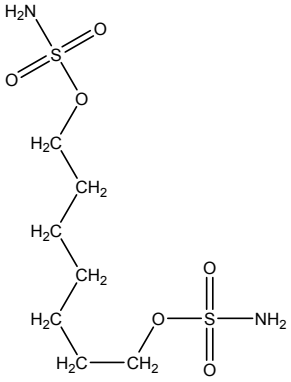
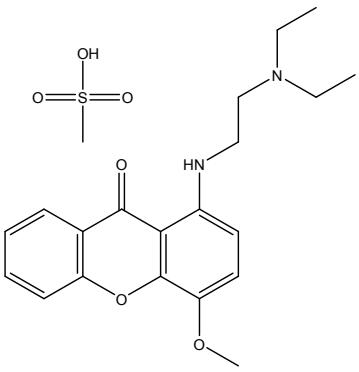
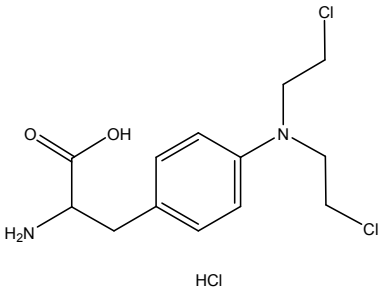
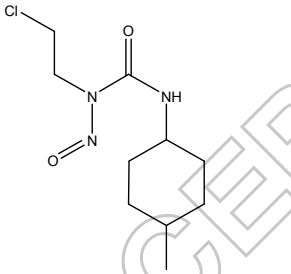
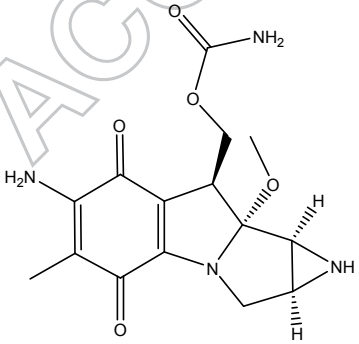
Wu, X., Fini, P., Keller, S., Tarsa, E., Heying, B., Mishra, U., . . . Speck, J. (1996). Morphological and structural transitions in GaN films grown on sapphire by metal-organic chemical vapor deposition. *Japanese journal of applied physics, 35*(12B), L1648.

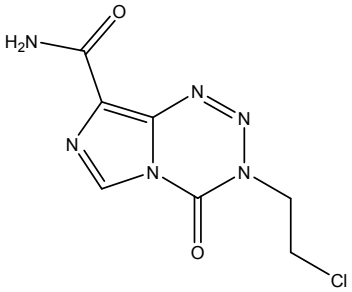
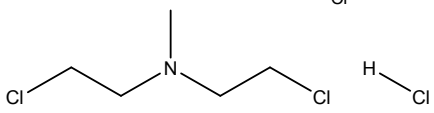
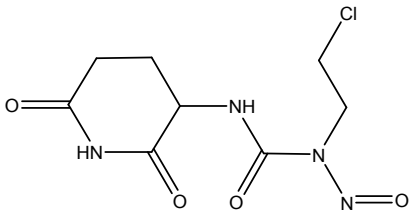
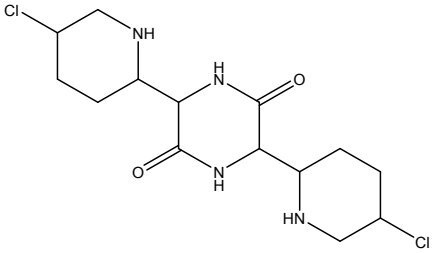
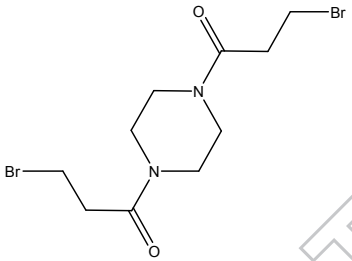
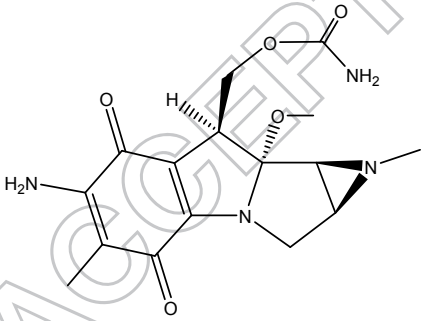
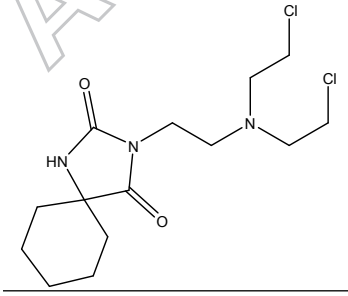
Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem, 32*(7), 1466-1474.

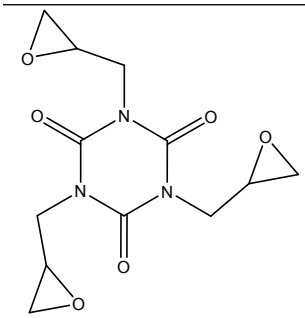
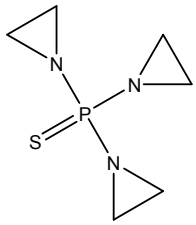
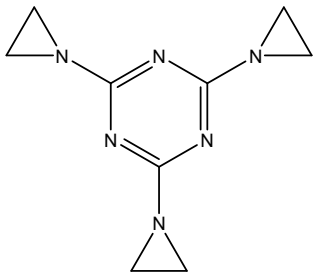
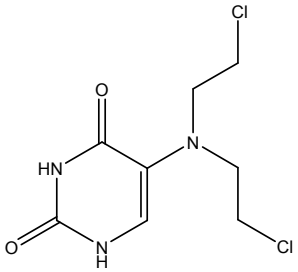
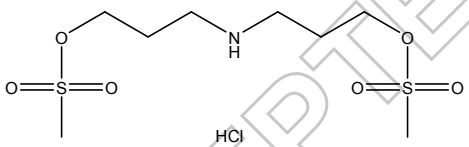
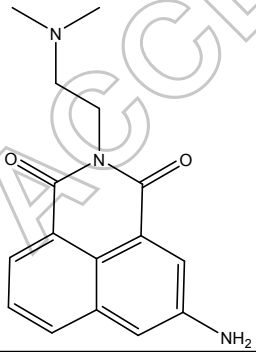
Table 1. Structure, Name, CAS and NSC number of the complete dataset

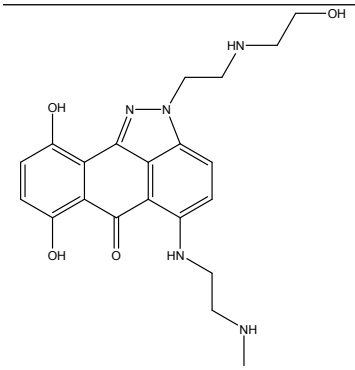
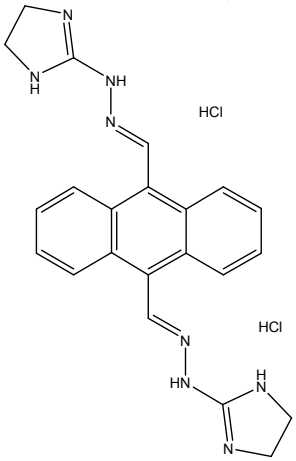
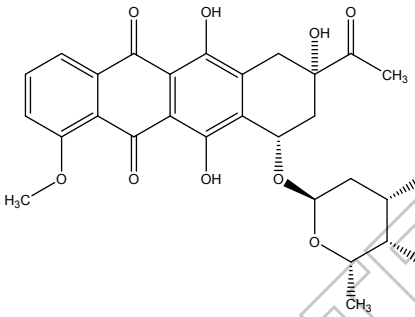
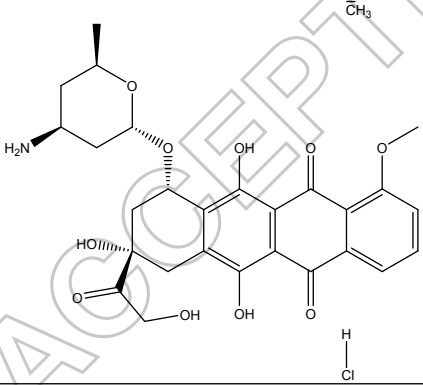
STRUCTURE	NAME	CAS	NSC number
	ASALEY	42228-92-2	167780
	AZQ		182986
	BCNU		409962
	BUSULFEX (BUSULFAN)	55-98-1	750
	CCNU	13010-47-4	79037
	CHLORAMBUCIL	305-03-3	3088

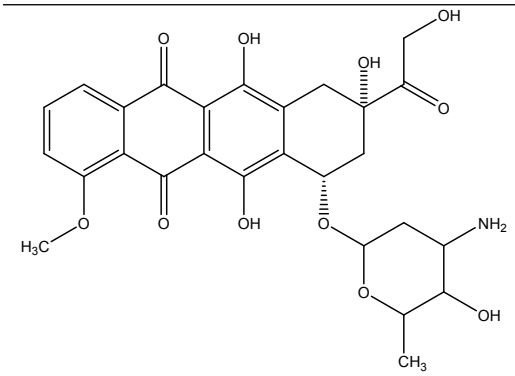
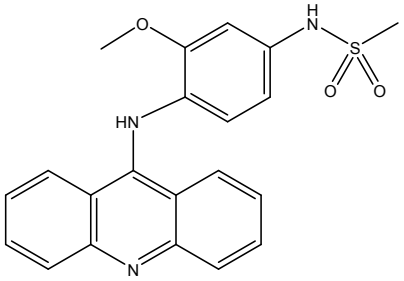
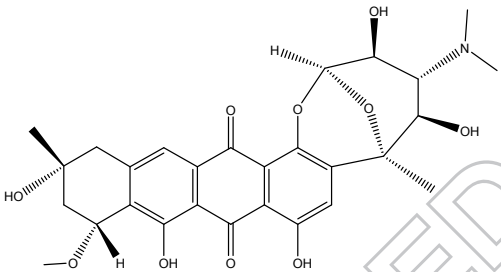
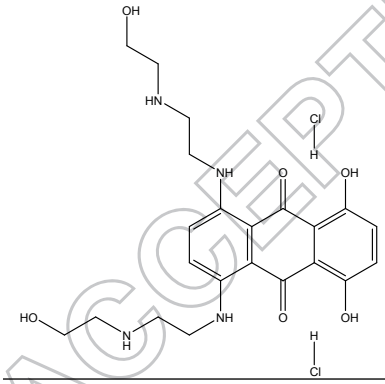
	CHLOROZOTOCIN	54749-90-5	178248
	CLOMESONE	88343-72-0	338947
	CYANOMORPHOLINO ADR	88254-07-3	357704
	CYCLODISONE	99591-73-8	348948
	DIANHYDROGALACTI-TOL	23261-20-3	132313
	FLUORODOPAN	834-91-3	73754

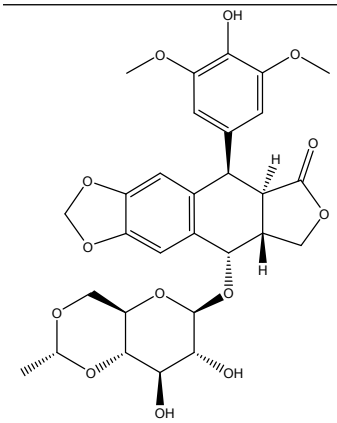
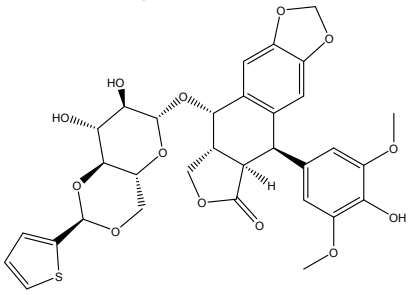
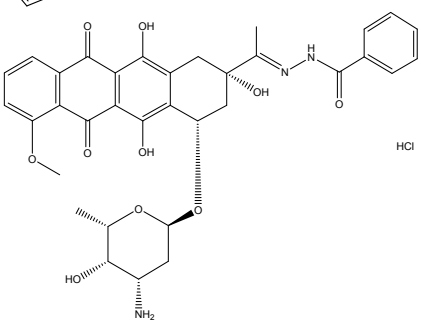
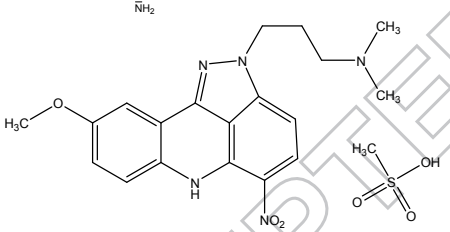
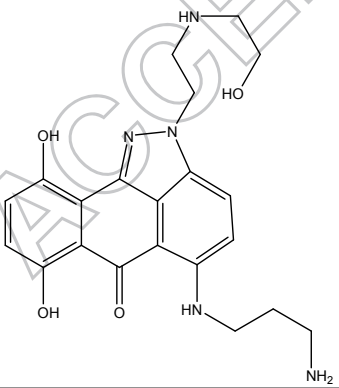
	HEPSULFAM	96892-57-8	329680
	HYCANTHONE	23255-93-8	142982
	MELPHALAN	3223-07-2	8806
	METHYL CCNU	13909-09-6	95441
	MITOMYCIN C	50-07-7	26980

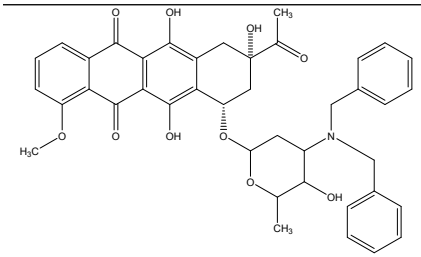
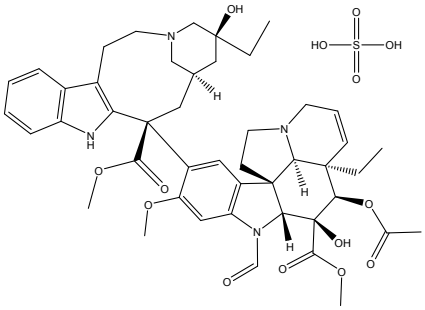
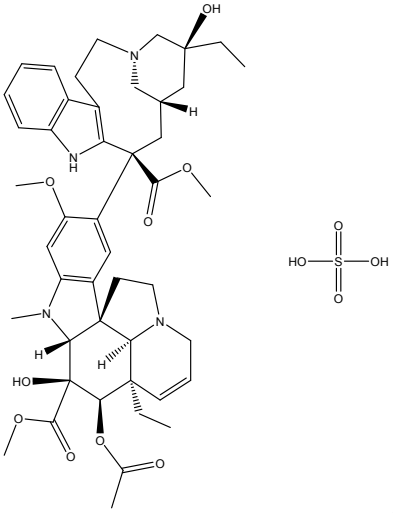
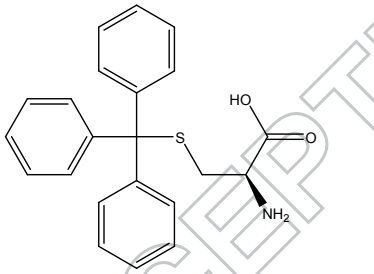
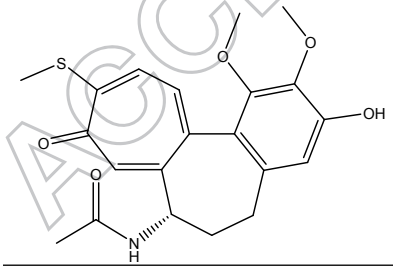
	MITOZOLOMIDE	85622-95-3	353451
	CARYOLYSINE (NITROGEN MUSTARD)	55-86-7	762
	PCNU	13909-02-9	349174
	PIPERAZINEDIONE	41109-80-2	135758
	PIPOBROMAN	54-91-1	25154
	PORFIROMYCIN	801-52-5	56410
	SPIROHYDANTOIN MUSTARD	56605-16-4	172112

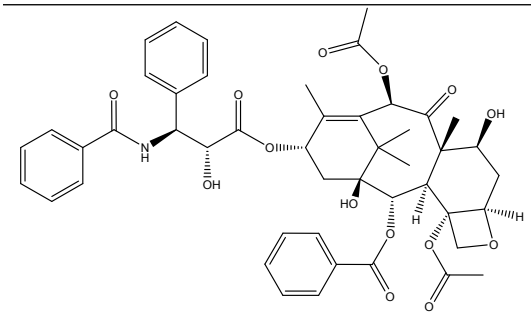
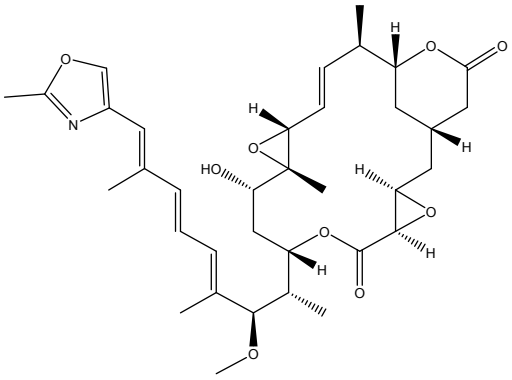
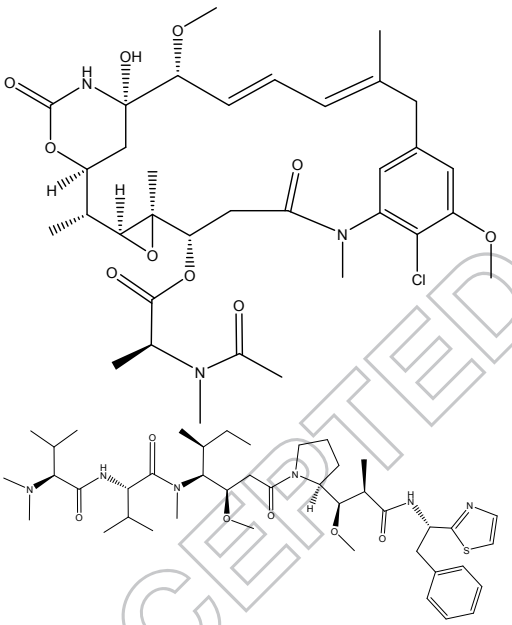
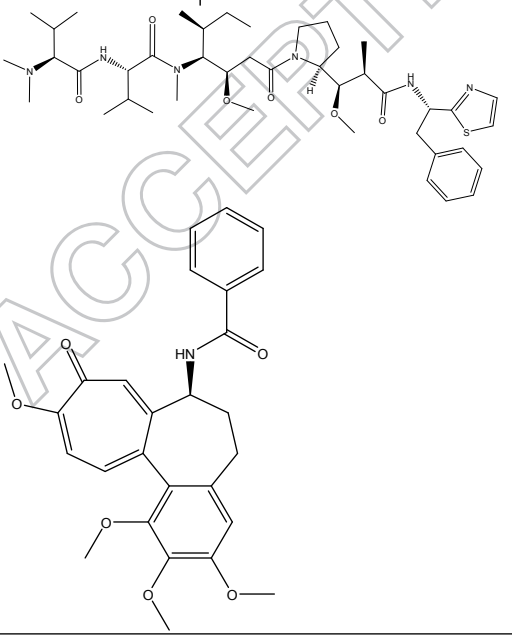
	TEROXIRONE	2451-62-9	296934
	THIOTEPA	52-24-4	6396
	TRIETHYLENEMELAMINE	51-18-3	9706
	URACIL NITROGEN MUSTARD	66-75-1	34462
	YOSHI 864	3458-22-8	102627
	AMONAFIDE	69408-81-7	308847

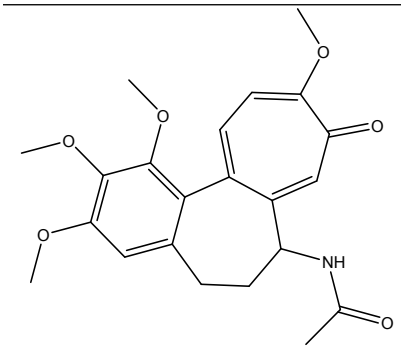
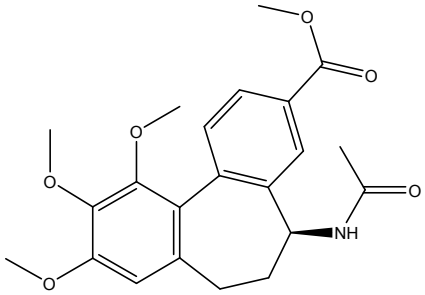
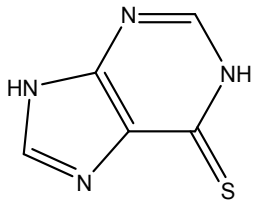
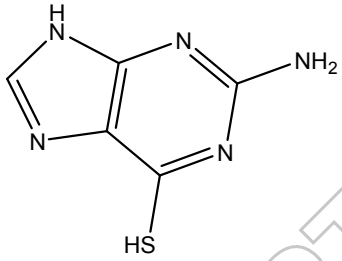
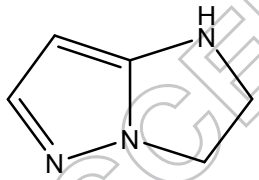
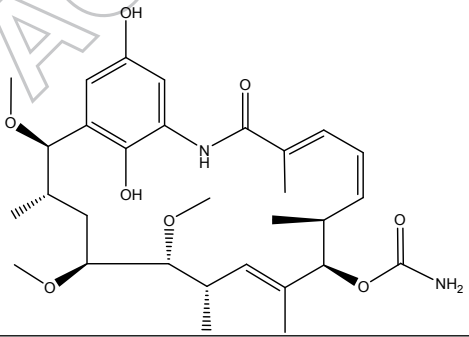
	ANTHRAPYRAZOLE	91440-30-1	355644
	BISANTRENE HYDROCHLORIDE	71439684	337766
	DAUNORUBICIN	23541-50-6	82151
	DEOXEODOXORUBIC-IN	63950-06-1	267469

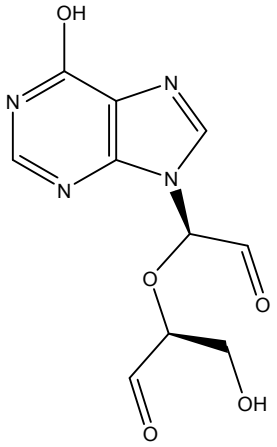
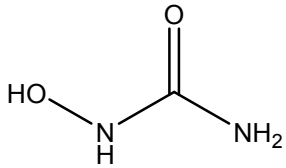
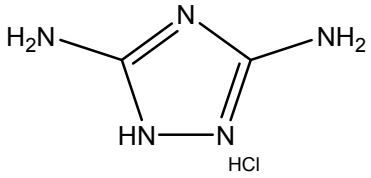
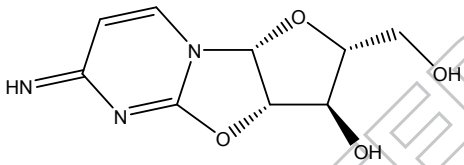
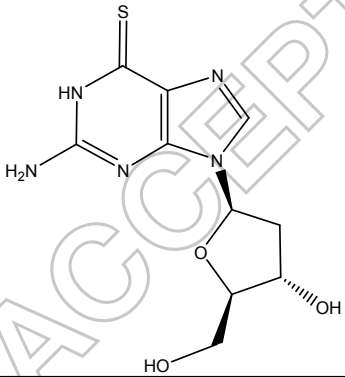
	DOXORUBICIN HYDROCHLORIDE	25316-40-9	357704
	MAMSA (AMSACRINE)	51264-14-3	249992
	MENOGARIL	71628-96-1	269148
	MITOXANTRON	70476-82-3	301739

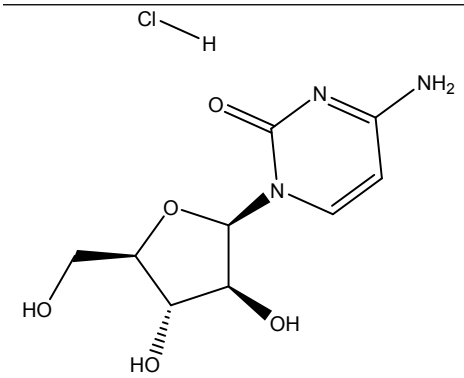
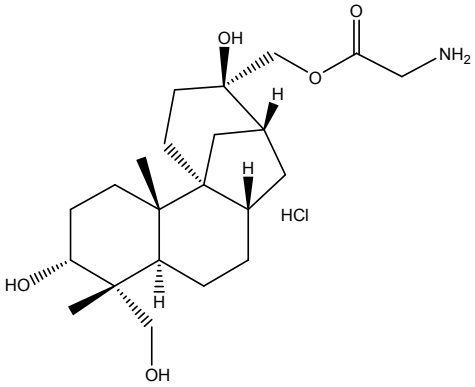
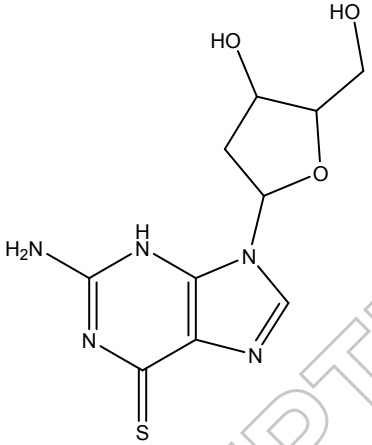
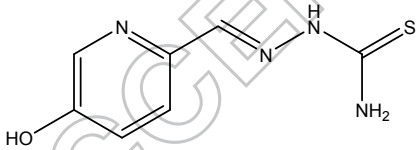
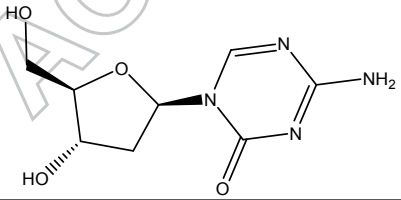
	VP-16 (ETOPOSIDE)	33419-42-0	141540
	VM-26 (TENIPOSIDE)	29767-20-2	122819
	RUBIDAZONE	36508-71-1	164011
	PYRAZOLOACRIDINE	99009-21-9	366140
	OXANTHRAZOLE	105118-12-5	349174

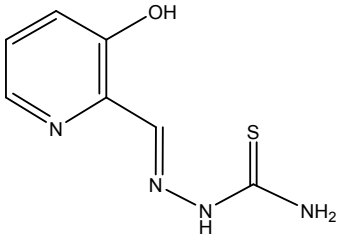
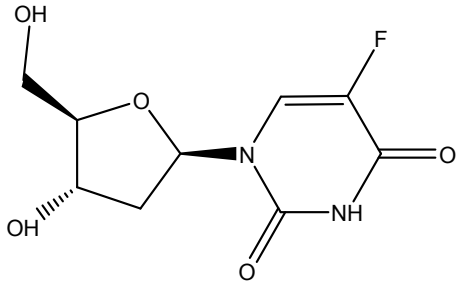
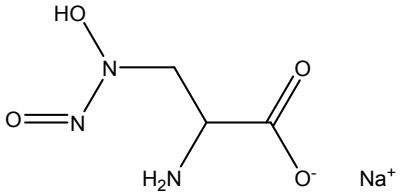
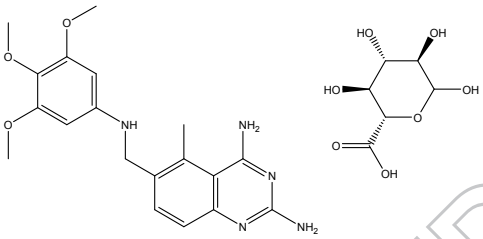
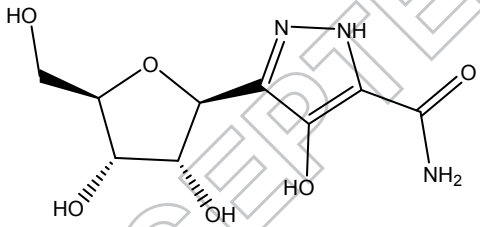
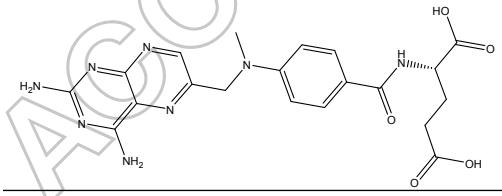
	N,NDIBENZYLDAUNORUBICIN	70878-51-2	268242
	VINCRISTINE SULFATE	2068-78-2	67574
	VINBLASTINE SULFATE	143-67-9	49842
	TRITYL LCYSTEINE	2799-07-7	83265
	THIOCOLCHICINE	87424-25-7	361792

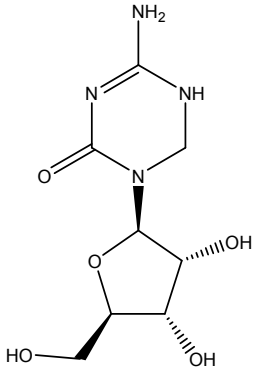
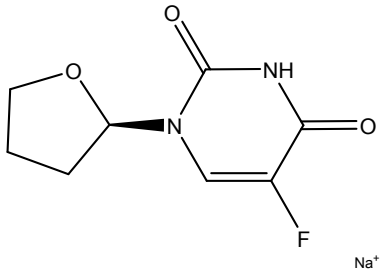
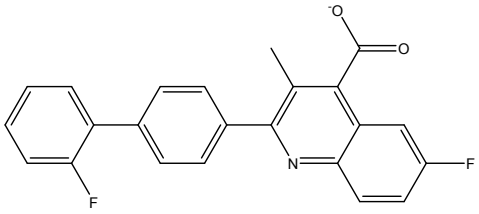
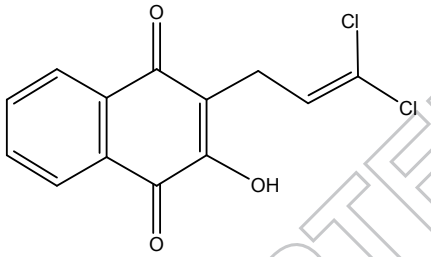
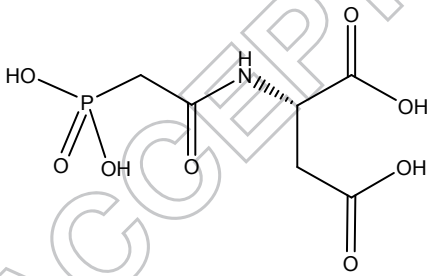
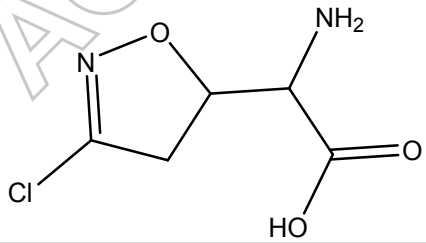
	PACLITAXEL (TAXOL)	33069-62-4	125973
	RHIZOXIN	90996-54-6	332598
	MAYTANSINE	35846-53-8	153858
	DOLASTATIN 10	110417-88-4	376128
	COLCHICINE DERIVATIVE	63989-75-3	33410

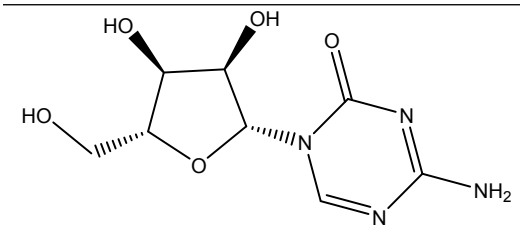
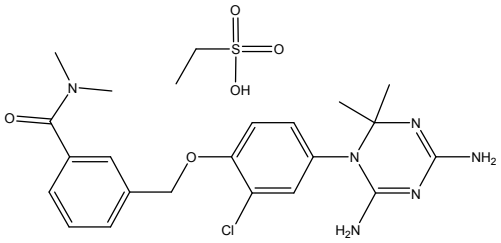
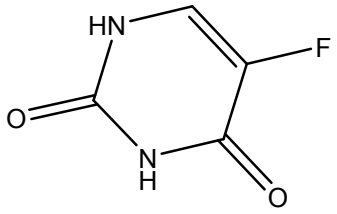
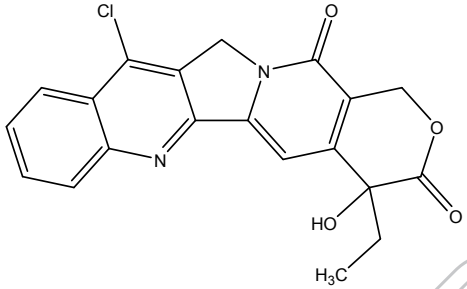
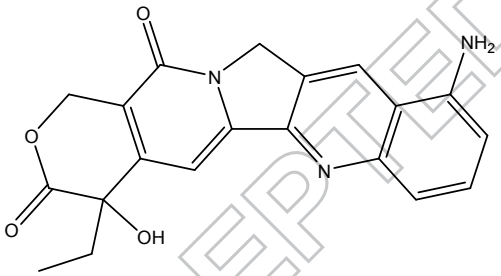
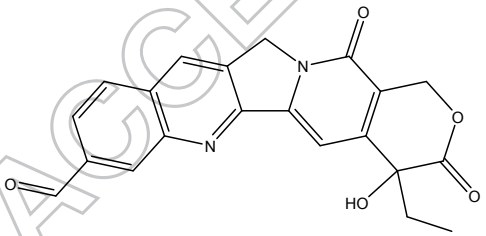
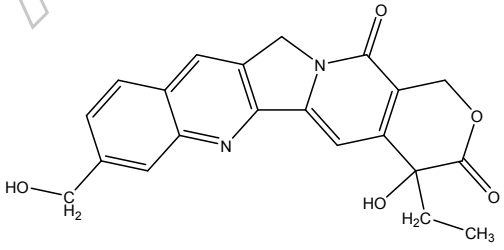
	COLCHICINE	64-86-8	757
	ALLOCOLCHICINE	641-28-1	406042
	THIOPURINE	50-44-2	755
	THIOGUANINE	154-42-7	752
	PYRAZOLOIMIDAZOLE	6714-29-0	51143
	MACBECIN II	73341-73-8	330500

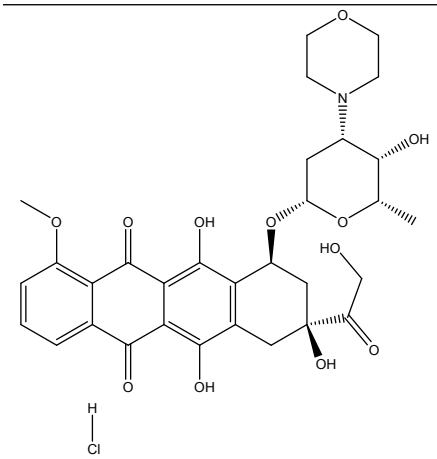
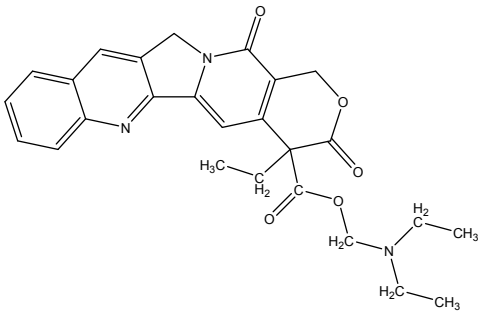
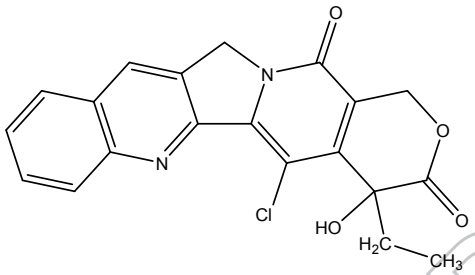
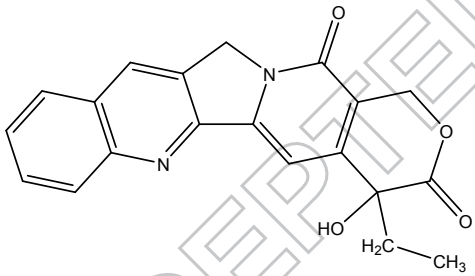
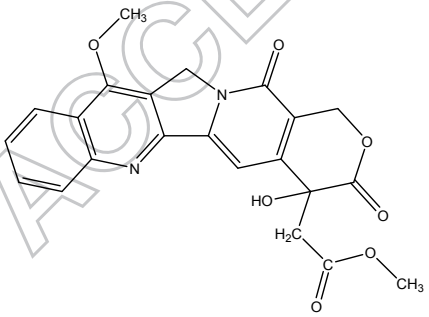
	INOSINE DIALDEHYDE	23590-99-0	118994
	HYDROXYLUREA	127-07-1	32065
	GUANAZOLE	1455-77-2	1895
	CYCLOCYTIDINE	10212-25-6	145668
	B-TGDR	789-61-7	71261

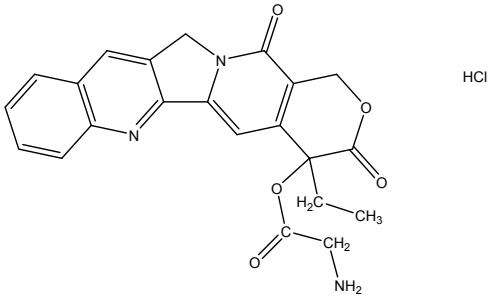
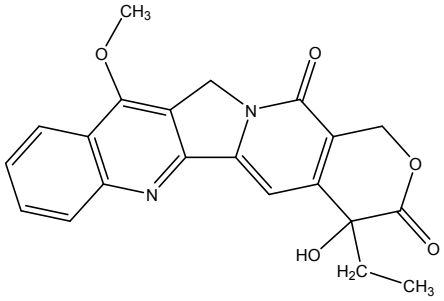
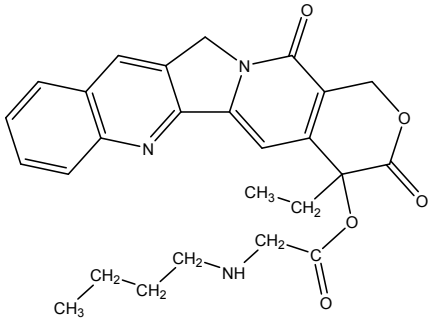
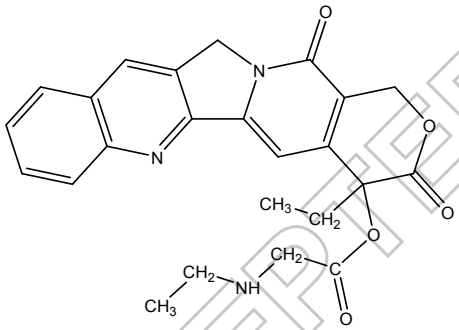
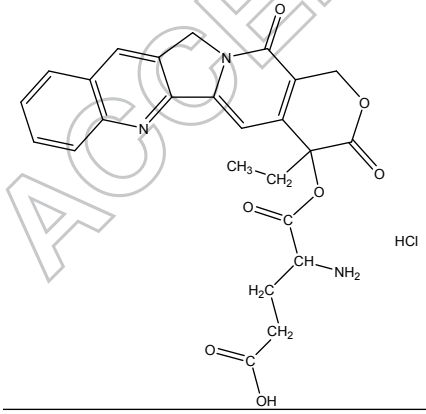
	ARA C	69-74-9	63878
	APHIDICOLIN GLYCINATE	92803-82-2	303812
	A-TGDR	2133-81-5	71851
	5 HP	19494-89-4	107392
	5-AZADEOXYCYTIDINE	2353-33-5	127716

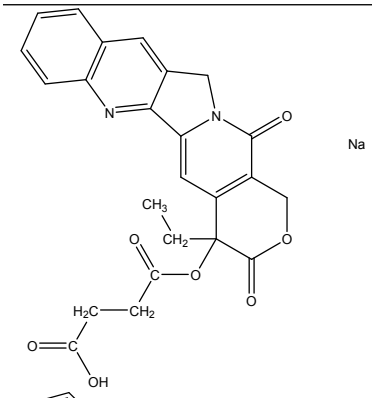
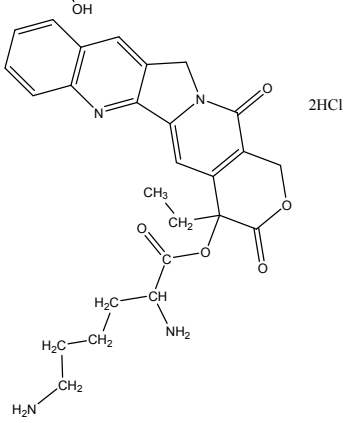
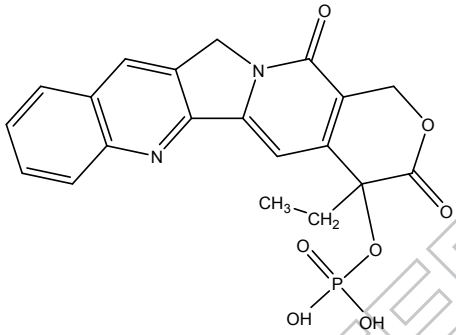
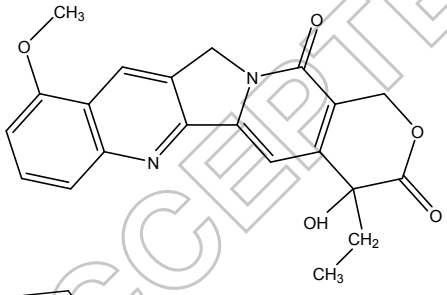
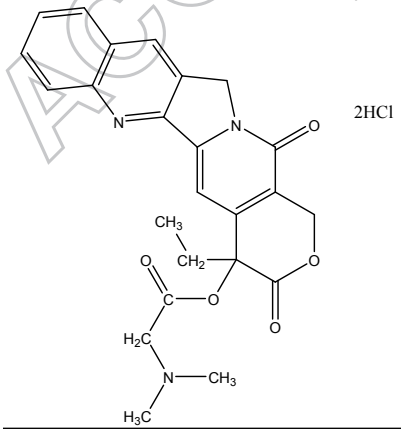
	3-HP	3814-79-7	95678
	2'DEOXY5FLUOROURIDINE	50-91-9	27640
	L-ALANOSINE	59163-41-6	153353
	TRIMETREXATE	82952-64-5	352122
	PYRAZOFURIN	30868-30-5	143095
	METHOTREXATE	59-05-2	740

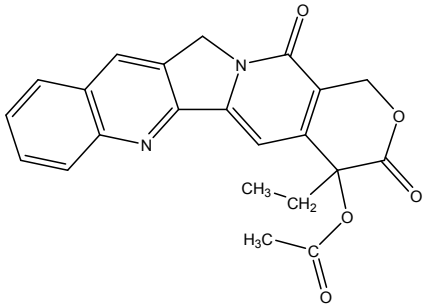
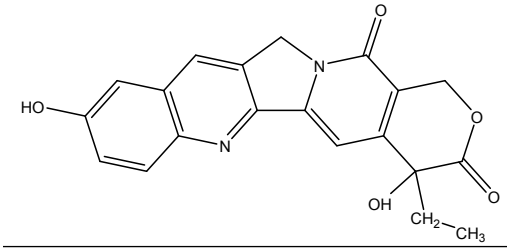
	DIHYDROAZACYTIDINE	62488-57-7	264880
	FTORAFUR	37076-68-9	148958
	DUP 785 (BREQUINAR)	96201-88-6	368390
	DICHLOROALLYL LAWSONE	36417-16-0	126771
	PALA	60342-56-5	224131
	ACIVICIN	42228-92-2	163501

	5-AZACYTIDINE	320-67-2	102816	
	BAKER'S ANTIFOL	41191-04-2	139105	
	5-FLUOROURACIL	51-21-8	19893	
	7-CHLOROCAMPTOTHECIN	41646-05-3	0	24991
	9-AMINO-20 (RS)-CAMPTOTHECIN	130194-90-0	629971	
	11 FORMYL 20 (RS) CAMPTOTHECIN	109466-93-5	606172	
	11-HYDROXYMETHYL20(RS) CAMPTOTHECIN		606173	

	MORPHOLINO-ADR	89196-04-3	354646
	GLYCINATE		364830
	14CHLORO20(S) CAMPTOTHECIN HYDRATE		643833
	CAMPTOTHECIN		94600
	CAMPTOTHECIN ANALOGUE 1		295500

	CAMPTOTHECIN ANALOGUE 2	606985
	CAMPTOTHECIN ANALOGUE 3	295501
	CAMPTOTHECIN BUTYLGLYCINATE ESTER HYDROCHLORIDE	606499
	CAMPTOTHECIN ETHYLGLYCINATE ESTER HYDROCHLORIDE	606497
	CAMPTOTHECIN GLUTAMATE HCL	610459

	CAMPTOTHECIN HEMISUCCINATE SODIUM SALT	100880
	CAMPTOTHECIN LYSINATE HCL	610457
	CAMPTOTHECIN PHOSPHATE	610458
	CAMPTOTHECIN, 9-METHOXY	176323
	CAMPTOTHECIN-20-O-(N,NDIMETHYL) GLYCINATE HCL	618939

	CAMPTOTHECIN, ACETATE	95382
	CAMPTOTHECIN, 10-HYDROXY	107124

PUBLIC INTEREST STATEMENT

Cancer at the present is considered as one of the most deadly disease in the world. Statistics from WHO indicates that one in every five people will die of cancer, this was attributed to the recent rise in chemical carcinogenic agents present in our treated waters, processed foods and non-food chemicals, normally found in homes. This paper aims to fastrack the discovery of anticancer drugs, by applying a well validated mathematical model. The model contains important chemical properties responsible for mitigating the growth of cancerous cells, which can be applied in designing and screening of potential anticancer drugs with high biological activity.

ABOUT THE AUTHORS

David Ebuka Arthur is a scientist with a keen interest in the areas of computational chemistry and drug design, whose desire for developing chemical space in identifying compounds with improved bioactivity is only surpassed by his unflinching pursuit in searching for the relationship between molecular structure and activities of lead compounds. He has published more than 30 scientific research papers. D.E. Arthur amongst other awe-inspiring scientists who authored this paper belongs to a special Nigerian Physical Chemistry Team, whose base is stationed at Ahmadu Bello University Zaria, presently known as the best University in Nigeria. The research team comprises of the Research Head Professor Adamu Uzairu, and other members such as Professor Paul P.A. Mamza, Gideon A. Shallangwa (PhD), Stephen E. Abechi (PhD) and David Ebuka Arthur (PhD) who have collectively spearheaded a lot of groundbreaking research in the area of Medicinal, Inorganic and Physical Chemistry. Furthermore, their efforts have been notably recognized by the numerous grant and publications owed to their names.