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 and Gideon Shallangwa
Insilico modelling of quantitative structure–activity relationship of pGI50 anticancer compounds on K-562 cell line

David Ebuka Arthur1*, Adamu Uzairu1, Paul Mamza1, Stephen Eyije Abechi1 and Gideon Shallangwa1

Abstract: The pGI50 cytotoxicity values of 112 compounds on K-562 cancer cell line were modelled in order to illustrate the quantitative structure–activity relationship 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 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 GI50 of anticancer compounds on K-562 leukemia cell line.

Subjects: Medicinal & Pharmaceutical Chemistry; Physical Chemistry; Computational and Theoretical Chemistry

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

ABOUT THE AUTHOR

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.

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 fasttrack 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.
1. Introduction

Cancer is one of the 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; World Health 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 immortalized myelogenous leukaemia 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 modelling techniques such as quantitative–structure 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; Speck-Planche, Kleandrova, Luan, & Cordeiro, 2012a, 2012b). The purpose of this study is 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.

2. Experimental section

The computational hardware and software used in this work includes the following: computer (HP pavilion Intel(R) core i5-4200U with 1.63 Hz and 2.3 Hz processors 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) and 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 numbers, were taken from the drug discovery and development arm of the National Cancer
The data contains aminopterin and camptothecin derivatives, colchicine analogues and so on. The anticancer activity results are shown in GI<sub>50</sub>, 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 (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<sub>50</sub>) of the studied compounds are 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 is used to improve the quality of information got from the compounds.

2.1. Generation of molecular descriptors
The two-dimensional (2D) structure of each of the compounds was generated using the sketch option on Spartan 14 and was converted into three-dimensional (3D) structure by using the view option on Spartan 14. From the build option on the program, the structures were minimized using molecular mechanic force field option to remove any strain present in the molecular structure. In addition, this ensures a well-defined 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 descriptors 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 structures 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

2.2. 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 80% of the entire data set, while the test set which constitutes the remaining 20% 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).

2.3. Data division
In order to obtain validated QSAR models, the data set 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% of compounds) are distributed within the whole descriptor space occupied by the entire data set, 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 the data set into a training test 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 \( \mu \) and \( \sigma \) 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 &
Table 1. Chemical names of data set 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-S-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-S-azacytidine                 | 264880  | 5.5                            | 5.571                      | -0.071   | -0.127                 |
| 7                 | 5-aza-2'-deoxycytidine                    | 127716  | 4a                             | 4.243                      | -0.243   | -0.596                 |
| 8                 | 5-Azacytidine                             | 102816  | 6.1                            | 5.289                      | 0.811    | 1.448                  |
| 9                 | 5-hp                                       | 107352  | 5.3                            | 5.350                      | -0.230   | -0.411                 |
| 10                | 7-Chloro-20(s)-camptothecin               | 249910  | 7.3                            | 7.307                      | 0.193    | 0.345                  |
| 11                | 9-Amino-20(rl,s)-camptothecin             | 629971  | 7.5                            | 7.307                      | 0.193    | 0.345                  |
| 12                | Acivicin                                   | 163501  | 5.5                            | 4.490                      | 1.010    | 2.478                  |
| 13                | Allocolchicine                             | 406042  | 8a                             | 6.869                      | 1.131    | 2.774                  |
| 14                | Alpha-tgdr                                 | 71851   | 4.1                           | 4.996                      | -0.896   | -1.599                 |
| 15                | Aminopterin derivative1                   | 132483  | 6.4**                          | 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                      | 303182  | 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                | Beto-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*                           | 7.050                      | -0.150   | -0.368                 |
| 30                | Busulfan                                   | 750     | 3.6*                           | 3.201                      | 0.399    | 0.978                  |
| 31                | Camptothecin                               | 94600   | 7.3*                           | 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*                           | 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**                          | 8.558                      | -2.058   | -5.049                 |
| 38                | Camptothecin hemisuccinate sodium salt      | 610456  | 6.3                            | 6.431                      | -0.131   | -0.234                 |

(Continued)
| Serial number (ID) | Name | NSC | K-562 (experimental pGI<sub>50</sub>) | K-562 (predicted pGI<sub>50</sub>) | Residual | Standardized residual |
|-------------------|------|-----|--------------------------------------|--------------------------------------|----------|-----------------------|
| 39                | Camptothecin lysinate hcl | 610457 | 7.2* | 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 | −0.120 | −0.222 |
| 45                | Camptothecin,20-o-((4-(2-hydroxyethyl)-1-piperazino) oac | 374028 | 6.1 | 7.211 | −1.111 | −2.726 |
| 46                | Camptothecin-20-o-(n,n-dimethyl)glycinate hcl | 618939 | 7.3 | 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 | 7.3 | 7.153 | 0.247 | 0.442 |
| 50                | Clomesone | 100880 | 7.3 | 7.424 | −0.120 | −0.222 |
| 51                | Colchicine | 757 | 7.2 | 7.402 | −0.202 | −0.362 |
| 52                | Colchicine derivative | 33410 | 7.9 | 7.947 | −0.047 | −0.116 |
| 53                | Cyanomorpholinodoxorubicin | 357704 | 8.3 | 8.023 | 0.277 | 0.494 |
| 54                | Cyclocytidine | 145668 | 3.4 | 4.465 | −1.055 | −2.612 |
| 55                | Cyclodisone | 142982 | 5.3 | 6.207 | −0.907 | −1.619 |
| 56                | Daunorubicin | 33410 | 7.9 | 7.947 | −0.047 | −0.116 |
| 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 | 4.587 | −1.187 | −2.912 |
| 63                | Flotrafur (pro-drug) | 148958 | 3 | 4.029 | −1.029 | −1.838 |
| 64                | Glycinate | 364830 | 7 | 7.718 | −0.718 | −1.282 |
| 65                | Guanazole | 82151 | 7 | 6.565 | 0.435 | 0.777 |
| 66                | Hepsulfam | 329680 | 3.4 | 3.245 | 0.155 | 0.276 |
| 67                | Hydrazide | 142982 | 5.3 | 6.207 | −0.907 | −1.619 |
| 68                | Hydroxyurea | 32065 | 3 | 3.119 | −0.119 | −0.213 |
| 69                | Inosine glycodialdehyde | 118994 | 4.6 | 3.228 | 0.772 | 1.378 |
| 70                | L-alanosine | 153353 | 4.8 | 6.127 | −1.327 | −3.256 |
| 71                | Macebcin i | 330500 | 7.1 | 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 | −0.072 | −0.128 |
| 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 | 0.336 | 0.707 |

(Continued)
The algorithm employs Euclidean 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.

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

2.4. Model development

Multiple linear regression was used to show the relationship between the dependent variable \( Y \) (\( \text{pGI}_{50} \)) 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 (\( \text{pGI}_{50} \)) \( Y \) relies on (descriptors) \( X \).

2.5. 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).

2.6. 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 (\( Q^2_{\text{CV}} \)) was calculated with the following equation:

\[
Q^2_{\text{CV}} = 1 - \frac{\sum_{i=1}^{n} (y_{\exp} - y_{\text{pred}})^2}{\sum_{i=1}^{n} (y_{\exp} - \bar{y})^2}
\]

where \( y_{\text{pred}}, y_{\exp} \) and \( \bar{y} \) 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 (2002) and Roy, Kar, and Ambure (2015) were utilized. The coefficient of determination for the test set \( R^2_{\text{test}} \) was calculated through the accompanying mathematical statement:

\[
R^2_{\text{test}} = 1 - \frac{\sum (y_{\text{pred}} - y_{\text{test}})^2}{\sum (y_{\text{pred}} - y_{\text{Training}})^2}
\]

where \( y_{\text{pred}} \) and \( y_{\text{test}} \) are the predicted value founded on the QSAR equation (model response) and experimental activity values, respectively, of the external test set compounds. \( y_{\text{Training}} \) 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 \( (r_m)^2 \), using the \( r_m^2 \) metric calculator developed by Roy et al. (2013).

2.7. 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 \( h_i \) is defined as \( h_i = x_i(X'X)^{-1}x_i^T \) (\( i = 1, \ldots, m \)), where \( x_i \) is the descriptor row-vector of the query compound \( i \) and \( X \) is the \( n \times k \) descriptor matrix of the training set compounds used to develop the model. As a prediction tool, the warning leverage (\( h^* \)) is the limit of normal values for \( X \) outliers and is defined as \( h^* = 3(k + 1)/n \), where \( n \) is the number of training compounds and \( k \) is the number of descriptors in the model. The test compounds with leverages \( h_i < h^* \) 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 (\( h_i < h^* \)) and standardized residuals no greater than 3\( \alpha \) (3 standard deviation units), hence they are accepted as Y outlier. Test set compounds where \( h_i > h^* \) are thought to be unreliablely 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 \((h_i > h^*)\) in developing the model.

3. Results and discussion

A QSAR analysis was performed to explore the SAR 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).

### 3.1. QSAR on K-562 cell line data set

#### 3.1.1. K-562 cell line

\[
\begin{align*}
\text{pgI}_{50} &= -5.524(\text{Methanal}) + 5.514(\text{PSA}) - 6.097(\text{AT57e}) \\
& - 2.255(\text{ATSC5c}) - 1.219(\text{naasN}) - 2.813(\text{minHBint7}) - 2.162(\text{minHBint10}) \\
& + 1.482(\text{maxHBint5}) - 4.484(\text{hmax}) + 7.419(\text{MDEC}) - 11) \\
& + 8.762(\text{MDEC} - 23) - 3.254(\text{RDF155v}) + 6.467
\end{align*}
\]

\[
\begin{align*}
N_{\text{train}} &= 90, \quad R^2_{\text{train}} = 0.915, \quad R^2_{\text{adjusted}} = 0.902, \quad F_{\text{train}} = 69.298, \quad Q^2_{\text{LOO}} = 0.845. \quad \text{Outliers > 3.0 = 5.} \quad N_{\text{test}} = 22
\end{align*}
\]

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

| Test set validation information | Name                  | K-562               |
|--------------------------------|-----------------------|---------------------|
| Model biasness test            | Systematic error result | Absent             |
| \(R^2\) test (100% data)      | 0.6722                |                     |
| \(R^2\) test (100% data)      | 0.6614                |                     |
| Classical metrics              | Q2F1 (100% data)      | 0.9161              |
| (for 100% data)                | Q2F2 (100% data)      | 0.5816              |
| Scaled avg. \(Rm^2\) (100% data) | 0.5591               |                     |
| Scaled Delta \(Rm^2\) (100% data) | 0.1417               |                     |
| CCC (100% data)                | 0.7961                |                     |
| \(R^2\) test (95% data)       | 0.7390                |                     |
| Classical metric               | R^2 test (95% data)   | 0.7205              |
| (after removing 5% data with high residuals) | Q2F1 (95% data) | 0.9397              |
| Q2F2 (95% data)                | 0.6862                |                     |
| ScaledAvgRm^2 (95% data)       | 0.6509                |                     |
| ScaledDeltaRm^2 (95% data)     | 0.0601                |                     |
| CCC (95% data)                 | 0.8507                |                     |
| RMSEP (100% data)              | 1.1011                |                     |
| Error-based metrics            | SD (100% data)        | 0.6363              |
| (for 100% data)                | SE (100% data)        | 0.1357              |
| MAE (100% data)                | 0.9088                |                     |
| Basic data structure information | N compound test     | 22                  |
| Prediction quality             | Prediction quality    | Moderate            |
The built model was used to predict the test set data, and the results are presented in Table 1. The predicted pGI\textsubscript{50} values for the compounds in the training and test sets for K-562 leukaemia cell line were plotted against the experimental pGI\textsubscript{50} values in Figure 1. Likewise, the plot of the residuals values for both the training and test sets against the experimental pGI\textsubscript{50} estimations is presented in Figure 2. As can be seen from Table 1 and Figures 1 and 2, the computed values for the pGI\textsubscript{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 $R^2_{pred}$, $Q^2_F1$ and $Q^2_F2$ 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. (2015a) for a QSAR model biasness test is good and in well agreement with other standards stated by Golbraikh and Tropsha (2002).

### 3.2. 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 data set. The applicability domain is established inside a squared area within ±3 bound for residuals and a leverage threshold $h^*$ ($h^* = \frac{3p^o}{n}$, where $p^o$ 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 $h^* = 0.433.$

---

**Figure 1.** The predicted pGI\textsubscript{50} against the experimental values for the training and test sets of K562 leukaemia cell line.
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. Compounds 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 influence the scope of the model positively are structurally different from other compounds in the model (Roy et al., 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.

Table 3. $R^2_{\text{train}}$ and $Q^2_{\text{LOO}}$ 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|        |
| $cR^2$:     |      | 0.766 |        |
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 SAR to validate the adequacy of the training set molecules.

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

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

$MF_j$ represents the mean effect for the considered descriptor $j$, $\beta_j$ 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 reveal 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 PSA and molecules edge of the type (MEDC-23) were responsible for hindering the potency of these compounds on K-562 cancer cell line.

### 3.3. 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 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 optimization of a drug’s ability to permeate cells. The mean effect of PSA (−3.887) reported in Table 4 is significantly high and it is 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.

### 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|
| ATSC5c        | 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.
ATS7e is a 2D autocorrelation molecular descriptor developed by Todeschini and 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 \(\delta_{ij}\) is Kronecker delta, that is, \(\delta_{ij} = 1\) if the \(ij\)th entry in the topological level matrix is \(= d\) and \(\delta_{ij} = 0\) otherwise (Broto & Devillers, 1990; Broto, Moreau, & Vandycke, 1984; Moreau & Broto, 1980a, 1980b).

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 (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 descriptors 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, maximum H E-state, 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. Ojha, Mitra, Das, and 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. (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 descriptors 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
| ID | Methanal | PSA | ATS7e | ATSC5c | nagsN | minHBind7 | minHBind10 | maxHBind5 | hmax | MDEC-11 | MDEC-23 | RDF155v | pGI50 |
|----|----------|-----|-------|--------|-------|-----------|------------|-----------|------|---------|---------|---------|-------|
| CD1 | 0        | 0.849 | 0.317 | 0.752 | 0     | 0         | 0          | 0.764     | 0.398 | 0.61    | 0       | 12.391  |
| CD2 | 0        | 0.722 | 0.335 | 0.854 | 0     | 0         | 0          | 0.727     | 0.398 | 0.724   | 0       | 12.516  |
| CD3 | 0        | 0.595 | 0.388 | 0.825 | 0     | 0         | 0          | 0.697     | 0.398 | 0.87    | 0       | 12.972  |
| CD4 | 0        | 0.469 | 0.418 | 0.84  | 0     | 0         | 0          | 0.674     | 0.398 | 1       | 0       | 13.302  |
| CD5 | 0        | 0.469 | 0.536 | 1     | 0     | 0         | 0          | 0.609     | 0.398 | 1       | 0       | 12.513  |
| CD6 | 0        | 0.592 | 0.604 | 0.913 | 0     | 1         | 0          | 0.573     | 0.398 | 1       | 0       | 10.322  |
| CD7 | 0        | 0.592 | 0.73  | 0.904 | 0     | 0.756     | 0          | 0.506     | 0.398 | 1       | 0       | 10.560  |
| CD8 | 0        | 0.715 | 0.871 | 0.719 | 0     | 0.269     | 0          | 0.432     | 0.398 | 1       | 1       | 9.244   |
| CD9 | 0        | 0.715 | 0.956 | 0.71  | 0     | 0.222     | 0          | 0.198     | 0.398 | 1       | 0       | 13.182  |
| CD10| 0        | 0.357 | 0.981 | 0.787 | 0     | 0         | 0          | 0.183     | 0.623 | 1       | 0       | 13.242  |
| CD11| 0        | 0     | 1     | 0.937 | 0     | 0         | 0          | 0         | 1     | 1       | 0       | 14.437  |
| CA1 | 0        | 1     | 0.004 | 0.307 | 0     | 0         | 0          | 0         | 1     | 0       | 0       | 8.263   |
| CA2 | 0        | 1     | 0     | 0     | 0     | 0         | 0          | 0.982     | 0.993 | 0.179   | 0       | 10.553  |
| CA3 | 0        | 0.873 | 0.027 | 0.621 | 0     | 0         | 0          | 0.939     | 0.97  | 0.319   | 0       | 9.553   |
| CA4 | 0        | 0.996 | 0.094 | 0.633 | 0     | 0         | 0          | 0.956     | 0.86  | 0.319   | 0       | 10.314  |
| CA5 | 0        | 0.639 | 0.115 | 0.693 | 0     | 0         | 0          | 0.814     | 0.165 | 0.319   | 0       | 8.100   |
| CA6 | 0        | 0.996 | 0.134 | 0.563 | 0     | 0         | 0          | 0.935     | 0.852 | 0.319   | 0       | 10.233  |
| CA7 | 0        | 0.639 | 0.155 | 0.679 | 0     | 0         | 0          | 0.806     | 0.165 | 0.319   | 0       | 7.919   |
| CA8 | 0        | 0.639 | 0.206 | 0.52  | 0     | 0         | 0          | 0.786     | 0.165 | 0.319   | 0       | 8.056   |
| CA9 | 0        | 0.639 | 0.242 | 0.576 | 0     | 0         | 0          | 0.769     | 0.165 | 0.319   | 0       | 7.787   |
| CA10| 0        | 0.639 | 0.298 | 0.626 | 0     | 0         | 0          | 0.751     | 0.165 | 0.319   | 0       | 7.413   |
| CA11| 0        | 0.639 | 0.348 | 0.554 | 0     | 0         | 0          | 0.713     | 0.165 | 0.319   | 0       | 7.441   |
| CA12| 0        | 0.639 | 0.377 | 0.594 | 0     | 0         | 0          | 0.581     | 0.165 | 0.319   | 0       | 7.766   |
| Compound ID | Newly designed drugs | Predicted PGi\textsubscript{50} |
|-------------|----------------------|-----------------------------|
| 1           | CD1                  | 12.391                      |
|             | ![CD1 structure]     |                             |
| 2           | CD2                  | 12.516                      |
|             | ![CD2 structure]     |                             |
| 3           | CD3                  | 12.972                      |
|             | ![CD3 structure]     |                             |
| 4           | CD4                  | 13.302                      |
|             | ![CD4 structure]     |                             |

(Continued)
| Compound ID | Newly designed drugs | Predicted PG150 |
|-------------|----------------------|-----------------|
| 5           | CD5                  | 12.513          |
| 6           | CD6                  | 10.322          |
| 7           | CD7                  | 10.560          |
| 8           | CD8                  | 9.244           |
| 9           | CD9                  | 13.182          |
Table 6. (Continued)

| Compound ID | Newly designed drugs | Predicted PGI₅₀ |
|-------------|----------------------|-----------------|
| 10          | CD10                 | 13.242          |
| 11          | CD11                 | 14.437          |
| 12          | CA1                  | 8.263           |
| 13          | CA2                  | 10.553          |
| 14          | CA3                  | 9.553           |

(Continued)
| Compound ID | Newly designed drugs | Predicted PGlu |
|------------|----------------------|----------------|
| 15         | CA4                  | 10.314         |
| 16         | CA5                  | 8.100          |
| 17         | CA6                  | 10.233         |
| 18         | CA7                  | 7.919          |
| 19         | CA8                  | 8.056          |

(Continued)
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.

3.4. Ligand base drug design
Twenty-three 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 (CD) was used as a lead compound in designing the remaining 11 compounds. The MF value of PSA descriptors suggest the removal of hetero atoms such as oxygen and nitrogen in order to reduce the PSA of the compounds, while hmax supports the conversion of unsaturated carbons to saturated carbons or replacing the (–O–) alkoxy groups with methylene carbons (–CH₂–), thereby making more room for hydrogen atoms and increasing the possibility of hydrogen bond formation with the receptor.

The pGI₅₀ result of the designed analogues of CAMPTOTHECIN ANALogue 3 (CA) and COLCHICINE DERivative (CD) presented in Tables 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 lines.

4. Conclusion
For the robustness and statistical significance of the developed model, an initial division of data set was done for training and test set compounds using Kennard–Stone algorithm, before using genetic functional algorithm (GFA)-MLR tool for building the model. The model is statistically robust both internally (Q²: 0.845) and externally (Q²F₁: 0.9397; Q²F₂: 0.6862, R²pred: 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 PSA for any compound is required to improve the pGI₅₀ of anticancer agents.
induction. *Environmental Toxicology and Pharmacology*, 43, 13–20. doi:10.1016/j.etap.2015.11.001

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: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), 2839–2917. doi:10.1002/ijc.25516

Golbraikh, A., & Tropsha, A. (2002). Beware of q 2! *Journal of Molecular Graphics & Modelling*, 20(4), 269–276.

Gramatica, P., Giani, E., & Papa, E. (2007). Statistical external validation and consensus modeling: A QSAR case study for K oc prediction. *Journal of Molecular Graphics & Modelling*, 25(6), 755–766. doi:10.1016/j.jmgm.2006.06.005

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), 1039-1042.

Hehe, W. J., & Huang, W. W. (1995). Chemistry with computation: An introduction to SPARTAN. Wavefunction, Inc.

Iuliano, A., Strianese, D., Uccelli, G., Diplomatico, A., Tebaldi, S., & Bonavolonta, G. (2012). Risk factors for orbital exenteration in periorbital basal cell carcinoma. *American Journal of Ophthalmology*, 153(2), 238–241.

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

Karimian, E. G., Marriage, F., Merritt, A. J., Buthjem, 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. doi:10.1016/j.exphem.2013.11.006

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, 11(1), 137–148. doi:10.1080/00401706.1969.10490666

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. *Journal of Chemical Information and Computer Sciences*, 44 (5), 1886–1890. doi:10.1021/ci049794h

Liu, S., Cao, C., & Li, Z. (1998). Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on novel molecular distance-edge (MDE) vector, λ. *Journal of Chemical Information and Computer Sciences*, 38(3), 387–394. doi:10.1021/ci970109z

Luzzio, C. B., & Luzzio, 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. doi:10.1021/ci0304050

Moreau, G., & Broto, P. (1980b). The auto-correlation of a topological-structure-a new molecular descriptor (Vol. 4, pp. 359–360). FRANCE: GAUTHIER-VILLARS 120 BLVD SAINT-GERMAIN, 75280 PARIS CEDEX 06.

Moreau, G., & Broto, P. (1980b). Auto-correlation of molecular-structures, application to sar studies. *Nouveau Journal De Chimie-New Journal of Chemistry*, 4(12), 757–764.

News, B. (2005, 12 May). Cancers number one killer of men. *Health*. Retrieved April 18, 2016, from http://news.bbc.co.uk/2/hi/health/3019801.stm

Ohja, 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. doi:10.1016/j.chemolab.2011.03.011

Parkin, D. M., Boyd, L., & Walker, L. (2011). The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *British Journal of Cancer*, 105, 577–581. doi:10.1038/bjc.2011.489

Rajer-Kanduc, Ž., 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. doi:10.1016/S0169-7439(02)00110-7

Roy, K., Chakraborty, P., Mitra, I., Ohja, 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. doi:10.1002/jcc.23231

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:10.1016/j.chemolab.2015.04.013

Roy, K., Kar, S., & Das, R. N. (2015a). A primer on QSAR/ QSAR 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 isatins analogues as in vitro anti-cancer agents. *European Journal of Medicinal Chemistry*, 45(3), 1113–1118. doi:10.1016/j.ejmech.2009.12.010

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), doi:10.1515/CPPM2012-06145

Siegel, R. L., Miller, K. D., & Jemal, A. (2015). Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians*, 65(1), 5–29. doi:10.3322/caac.21254

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. D. (2012a). 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. doi:10.1016/j.bmc.2012.05.071

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. D. S. (2012b). 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: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*, 47, 111–122.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)* (Vol. 4). John Wiley & Sons.
Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29(6-7), 476–488. doi:10.1002/minf.201000061

Tropsha, A., Gromatica, P., & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinaorial Science*, 22(1), 69–77. doi:10.1002/(ISSN)1611-0218

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.

World Health Organization (2002). National cancer control programmes: Policies and managerial guidelines.

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. doi:10.1143/JJAP.35.L1648

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474. doi:10.1002/jcc.21707