



In silico modelling of quantitative structure–activity relationship of multi-target anticancer compounds on k-562 cell line

David Ebuka Arthur¹ · Adamu Uzairu¹ · Paul Mamza¹ · Stephen Eyije Abechi¹ · Gideon Shallangwa¹

Received: 9 October 2017 / Revised: 6 March 2018 / Accepted: 10 April 2018 / Published online: 23 May 2018
© Springer-Verlag GmbH Austria, part of Springer Nature 2018

Abstract

The pGI₅₀ cytotoxicity values of 112 compounds on K-562 cancer cell line were modeled to illustrate the quantitative structure–activity relationship (QSAR) of the compounds. The dataset 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 the genetic functional algorithm (GFA) for selection of descriptor to be modeled. The best QSAR model developed was then subjected to a rigorous statistical test. 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. The model developed was used in designing new anticancer drugs with higher activity or more potent and less toxic in nature when compared to the lead compound. These compounds significant activities were found to correlate to with some of the molecular descriptors such as the number of hydrogen bond acceptors present in the surface of the molecule.

Keywords K-562 cell line · QSAR · GFA-MLR · Anticancer · Williams plot

1 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 defense 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 from heart disease (News 2003). Most cancers 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

et al. 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 et al. 2011). Some tumor cells have been reported to resist the effect of present-day chemotherapeutic agents, given rising 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 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 and 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 et al. 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, the 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.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13721-018-0168-y>) contains supplementary material, which is available to authorized users.

✉ David Ebuka Arthur
eadavid@abu.edu.ng

¹ Department of Chemistry, Ahmadu Bello University (ABU) Zaria, Zaria, Kaduna, Nigeria

2016). These kinases assume a part in the development of spindles, the 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 acknowledgment and popularity. This is due to its tendency to reduce the classical trial and error approach (Roy et al. 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 et al. 2010; Speck-Planche et al. 2012a, b). 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.

2 Experimental section

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

The dataset 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 et al. 2003). The chemical structures of the dataset, 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 et al. (2015) and Roy (Kar and Roy 2012; Roy et al. 2015a).

The counterpart of the ions was optimized at a protonated state, as they should in solution.

The biological activity ($-\text{LogGI}_{50}$) of the studied compounds are presented in Table 1 and the dataset of the activities ranges from 2.2 to 9.3. Further literature (Chopade et al. 2015), showing the wide range of activities dataset used to improve the quality of information gotten from the compounds.

3 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 using the view option on Spartan 14. From the build option on the program, the structures were minimized using molecular mechanic force field (MMFF) 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 et al. 1989). From the setup 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 was 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 was 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.

4 Splitting of dataset into modelling sets and evaluation test sets

The dataset 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 dataset. While the test set which constitutes the remaining 20% of the whole dataset was not used in the construction of the model but to ascertain the predictive ability of the model (Tropsha 2010).

5 Data division

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

Table 1 Chemical names of dataset with NSC numbers and their pGI₅₀ 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 | 606,172 | 5.7 | 4.808 | 0.892 | 1.592 |
| 2 | 11-Hydroxymethyl-20(RS)-camptothecin | 606,173 | 5.6 | 6.165 | -0.565 | -1.009 |
| 3 | 14-Chloro-20(S)-camptothecin hydrate | 643,833 | 5.7 | 6.521 | -0.821 | -1.466 |
| 4 | 2'-Deoxy-5-fluorouridine | 27,640 | 6.1 | 4.809 | 1.291 | 2.305 |
| 5 | 3-HP | 95,678 | 5.7 | 5.888 | -0.188 | -0.336 |
| 6 | 5,6-Dihydro-5-azacytidine | 264,880 | 5.5 | 5.571 | -0.071 | -0.127 |
| 7 | 5-Aza-2'-deoxycytidine | 127,716 | 4 ^a | 4.243 | -0.243 | -0.596 |
| 8 | 5-Azacytidine | 102,816 | 6.1 | 5.289 | 0.811 | 1.448 |
| 9 | 5-HP | 107,392 | 5.3 | 5.530 | -0.230 | -0.411 |
| 10 | 7-Chlorocamptothecin | 249,910 | 7.3 ^s | 6.897 | 0.403 | 0.720 |
| 11 | 9-Amino-20-(R,S)-camptothecin | 629,971 | 7.5 | 7.307 | 0.193 | 0.345 |
| 12 | Acivicin | 163,501 | 5.5 ^a | 4.490 | 1.010 | 2.478 |
| 13 | Allocolchicine | 406,042 | 8 ^a | 6.869 | 1.131 | 2.774 |
| 14 | Alpha-TGDR | 71,851 | 4.1 | 4.996 | -0.896 | -1.599 |
| 15 | Aminopterin derivative 1 | 132,483 | 6.4 ^{a,*} | 8.250 | -1.850 | -4.539 |
| 16 | Aminopterin derivative 2 | 184,692 | 8 | 8.520 | -0.520 | -0.929 |
| 17 | Aminopterin derivative 3 | 134,033 | 7.6 | 8.334 | -0.734 | -1.311 |
| 18 | Amonafide | 308,847 | 5.4 | 5.671 | -0.271 | -0.484 |
| 19 | AN antifol | 623,017 | 7.6 | 7.344 | 0.256 | 0.457 |
| 20 | Anthrapyrazole derivative | 355,644 | 6.7 | 5.929 | 0.771 | 1.377 |
| 21 | Aphidicolin glycinate | 303,812 | 5.3 | 5.744 | -0.444 | -0.793 |
| 22 | ARA-C | 63,878 | 4.6 | 5.422 | -0.822 | -1.467 |
| 23 | Asaley | 167,780 | 5.2 | 5.811 | -0.611 | -1.498 |
| 24 | AZQ | 182,986 | 5.3 | 5.203 | 0.097 | 0.174 |
| 25 | Baker's soluble antifol | 139,105 | 6.8 | 6.653 | 0.147 | 0.262 |
| 26 | BCNU | 409,962 | 4.3 | 3.858 | 0.442 | 0.789 |
| 27 | Beta-TGDR | 71,261 | 6.2 | 5.348 | 0.852 | 1.521 |
| 28 | Bisantrene HCl | 337,766 | 7.3 | 6.931 | 0.369 | 0.659 |
| 29 | Brequinar | 368,390 | 6.9 ^a | 7.050 | -0.150 | -0.368 |
| 30 | Busulfan | 750 | 3.6 ^a | 3.201 | 0.399 | 0.978 |
| 31 | Camptothecin | 94,600 | 7.3 ^a | 6.766 | 0.534 | 1.311 |
| 32 | Camptothecin analog | 295,500 | 6 | 6.655 | -0.655 | -1.169 |
| 33 | Camptothecin analog 2 | 606,985 | 7.5 | 6.622 | 0.878 | 1.567 |
| 34 | Camptothecin analog 3 | 295,501 | 7.5 ^a | 7.019 | 0.481 | 1.179 |
| 35 | Camptothecin butylglycinate ester hydrochloride | 606,499 | 6.3 | 6.528 | -0.228 | -0.408 |
| 36 | Camptothecin ethylglycinate ester hydrochloride | 606,497 | 6.1 | 6.466 | -0.366 | -0.654 |
| 37 | Camptothecin glutamate HCl | 610,459 | 6.5 ^{a,*} | 8.558 | -2.058 | -5.049 |
| 38 | Camptothecin hemisuccinate sodium salt | 610,456 | 6.3 | 6.431 | -0.131 | -0.234 |
| 39 | Camptothecin lysinate HCl | 610,457 | 7.2 ^a | 6.366 | 0.834 | 2.046 |
| 40 | Camptothecin phosphate | 610,458 | 6.2 | 4.868 | 1.332 | 2.379 |
| 41 | Camptothecin, 9-methoxy- | 176,323 | 7.3 | 7.002 | 0.298 | 0.532 |
| 42 | Camptothecin, acetate | 95,382 | 5.5 | 6.050 | -0.550 | -1.349 |
| 43 | Camptothecin, hydroxy- | 107,124 | 7.4 | 7.153 | 0.247 | 0.442 |
| 44 | Camptothecin, NA salt | 100,880 | 7.3 | 7.424 | -0.124 | -0.222 |
| 45 | Camptothecin,20-O-(4-(2-hydroxyethyl)-1-piperazino)OAC | 374,028 | 6.1 ^a | 7.211 | -1.111 | -2.726 |
| 46 | Camptothecin-20-O-(N,N-dimethyl)glycinate HCl | 618,939 | 7.3 ^a | 7.767 | -0.467 | -1.147 |
| 47 | CCNU | 79,037 | 4.6 | 4.393 | 0.207 | 0.370 |
| 48 | Chlorambucil | 3088 | 4 | 4.608 | -0.608 | -1.086 |

Table 1 (continued)

| Serial number (ID) | Name | NSC | K-562 (experimental pGI ₅₀) | K-562 (predicted pGI ₅₀) | Residual | Standardized residual |
|--------------------|--|---------|---|--------------------------------------|----------|-----------------------|
| 49 | Chlorozotocin | 178,248 | 3.2 | 2.824 | 0.376 | 0.671 |
| 50 | Clomesone | 338,947 | 3.3 ^a | 2.779 | 0.521 | 1.277 |
| 51 | Colchicine | 757 | 7.2 | 7.402 | -0.202 | -0.362 |
| 52 | Colchicine derivative | 33,410 | 7.9 ^a | 7.947 | -0.047 | -0.116 |
| 53 | Cyanomorpholinodoxorubicin | 357,704 | 8.3 | 8.023 | 0.277 | 0.494 |
| 54 | Cyclocytidine | 145,668 | 3.4 ^a | 4.465 | -1.065 | -2.612 |
| 55 | Cyclodisone | 348,948 | 4.1 | 3.032 | 1.068 | 1.906 |
| 56 | Daunorubicin | 82,151 | 7 | 6.565 | 0.435 | 0.777 |
| 57 | Deoxydoxorubicin | 267,469 | 7.4 | 7.731 | -0.331 | -0.591 |
| 58 | Dianhydrogalactitol | 132,313 | 3.9 | 4.369 | -0.469 | -0.838 |
| 59 | Dichlorallyl lawsone | 126,771 | 5.7 | 5.962 | -0.262 | -0.468 |
| 60 | Dolastatin 10 | 376,128 | 10.2 | 9.797 | 0.403 | 0.720 |
| 61 | Doxorubicin | 123,127 | 7 | 7.485 | -0.485 | -0.865 |
| 62 | Fluorodopan | 73,754 | 3.4 ^a | 4.587 | -1.187 | -2.912 |
| 63 | Ftorafur (pro-drug) | 148,958 | 3 | 4.029 | -1.029 | -1.838 |
| 64 | Glycinate | 364,830 | 7 | 7.718 | -0.718 | -1.282 |
| 65 | Guanazole | 1895 | 2.2 ^{a,*} | 4.738 | -2.538 | -6.226 |
| 66 | Hepsulfam | 329,680 | 3.4 | 3.245 | 0.155 | 0.276 |
| 67 | Hycanthone | 142,982 | 5.3 | 6.207 | -0.907 | -1.619 |
| 68 | Hydroxyurea | 32,065 | 3 | 3.119 | -0.119 | -0.213 |
| 69 | Inosine glycodialdehyde | 118,994 | 4 ^s | 3.228 | 0.772 | 1.378 |
| 70 | L-Alanosine | 153,353 | 4.8 ^{a,*} | 6.127 | -1.327 | -3.256 |
| 71 | Macbecin II | 330,500 | 7.1 ^a | 8.458 | -1.358 | -3.331 |
| 72 | M-amsa | 249,992 | 6* | 5.616 | 0.384 | 0.686 |
| 73 | Maytansine | 153,858 | 7.8 | 8.709 | -0.909 | -1.624 |
| 74 | Melphalan | 8806 | 4.3 | 4.551 | -0.251 | -0.449 |
| 75 | Menogaril | 269,148 | 5.9 | 5.972 | -0.072 | -0.128 |
| 76 | Methotrexate | 740 | 7.5 | 6.725 | 0.775 | 1.383 |
| 77 | Methotrexate derivative | 174,121 | 9.4 | 9.272 | 0.128 | 0.229 |
| 78 | Methyl CCNU | 95,441 | 4.4 | 4.647 | -0.247 | -0.441 |
| 79 | Mitomycin C | 26,980 | 5.6 | 5.204 | 0.396 | 0.707 |
| 80 | Mitoxantrone | 301,739 | 6.9 | 6.927 | -0.027 | -0.048 |
| 81 | Mitozolamide | 353,451 | 4.1 | 4.052 | 0.048 | 0.086 |
| 82 | Morpholinodoxorubicin | 354,646 | 8.6 | 7.752 | 0.848 | 1.514 |
| 83 | <i>N</i> -(phosphonoacetyl)-L-aspartate (pala) | 224,131 | 4 | 4.822 | -0.822 | -1.468 |
| 84 | <i>N,N</i> -dibenzyl daunomycin | 268,242 | 5.2 ^s | 5.289 | -0.089 | -0.159 |
| 85 | Nitrogen mustard | 762 | 5.2 | 3.963 | 1.237 | 2.209 |
| 86 | Oxanthrazole | 349,174 | 5.9 | 6.560 | -0.660 | -1.178 |
| 87 | PCNU | 95,466 | 3.8 | 4.014 | -0.214 | -0.382 |
| 88 | Piperazine drugmainator | 344,007 | 3.7 | 3.501 | 0.199 | 0.356 |
| 89 | Piperazinedione | 135,758 | 5.6 | 5.637 | -0.037 | -0.067 |
| 90 | Pipobroman | 25,154 | 3.9 | 4.267 | -0.367 | -0.656 |
| 91 | Porfiromycin | 56,410 | 4.8 | 4.962 | -0.162 | -0.290 |
| 92 | Pyrazofurin | 143,095 | 6.3 | 6.082 | 0.218 | 0.389 |
| 93 | Pyrazoloacridine | 366,140 | 6.7 | 5.895 | 0.805 | 1.438 |
| 94 | Pyrazoloimidazole | 51,143 | 2.5 | 2.649 | -0.149 | -0.267 |
| 95 | Rhizoxin | 332,598 | 8 | 7.624 | 0.376 | 0.672 |
| 96 | Rubidazone | 164,011 | 6.4 | 6.730 | -0.330 | -0.589 |

Table 1 (continued)

| Serial number (ID) | Name | NSC | K-562 (experimental pGI ₅₀) | K-562 (predicted pGI ₅₀) | Residual | Standardized residual |
|--------------------|-------------------------|---------|---|--------------------------------------|----------|-----------------------|
| 97 | Spirohydantoin mustard | 172,112 | 3.7 | 4.093 | -0.393 | -0.701 |
| 98 | Taxol | 125,973 | 8.4 | 8.746 | -0.346 | -0.617 |
| 99 | Teroxirone | 296,934 | 4.5 ^a | 3.977 | 0.523 | 1.283 |
| 100 | Tetraplatin | 363,812 | 6 | 5.387 | 0.613 | 1.095 |
| 101 | Thiocolchicine | 361,792 | 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 | 352,122 | 7.6 | 7.460 | 0.140 | 0.250 |
| 106 | Trityl cysteine | 83,265 | 6.2 | 6.105 | 0.095 | 0.169 |
| 107 | Uracil nitrogen mustard | 34,462 | 4.4 | 4.924 | -0.524 | -0.936 |
| 108 | Vinblastine sulfate | 49,842 | 9.3 | 8.995 | 0.305 | 0.544 |
| 109 | Vincristine sulfate | 67,574 | 7 | 6.348 | 0.652 | 1.164 |
| 110 | VM-26 | 122,819 | 6.1 | 6.577 | -0.477 | -0.852 |
| 111 | VP-16 | 141,540 | 4.7 | 5.277 | -0.577 | -1.029 |
| 112 | Yoshi-864 | 102,627 | 2.7 | 3.441 | -0.741 | -1.323 |

*Identifies compounds found outside the applicability domain

[§]Compounds structurally different from all other compounds within the dataset

^aTest sets for K-562 leukaemia cell lines, respectively

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 et al. 2003; Wu et al. 1996; Kennard and Stone 1969).

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

K is the number of inputs and 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

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 Euclidean distance of x -vectors (predictors) and then it sequentially selects a sample to maximize the Euclidean 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 Euclidean distance in x space is defined as (Wu et al. 1996; Saptoro et al. 2012; Kennard and Stone

1969). The algorithm employs Euclidean distance $ED_x(p, q)$, between the x vectors of each pair (p, q) of samples 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.

5.1 Model development

Multiple linear regression was used to show the relationship between the dependent variable Y (pGI₅₀) 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, the contingent mean of dependent variable (pGI₅₀) Y relies on (descriptors) X .

5.2 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).

5.3 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:

$$Q_{CV}^2 = 1 - \text{press}/_{\text{total}} = 1 - \frac{\sum_{i=1}^n (y_{\text{exp}} - y_{\text{pred}})^2}{\sum_{i=1}^n (y_{\text{exp}} - \bar{y})^2}$$

where y_{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 et al. (2015) were utilized. The coefficient of determination for the test set R_{test}^2 , was calculated through the accompanying mathematical statement

$$R_{\text{Test}}^2 = 1 - \frac{\sum (Y_{\text{pred}_{\text{test}}} - Y_{\text{Test}})^2}{\sum (Y_{\text{pred}_{\text{test}}} - \bar{Y}_{\text{Training}})^2}$$

where $Y_{\text{pred}_{\text{test}}}$ and Y_{Test} is the predicted value founded on the QSAR equation (model response) and experimental activity values, respectively, of the external test set compounds. $\bar{Y}_{\text{Training}}$ is the average activity value of the training set compounds (Tropsha et al. 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 rm^2 metric calculator developed by Roy et al. (2013).

5.4 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 et al. 2007), Leverage of a given chemical compound h_i , is defined as: $h_i = x_i(X^T X)^{-1} x_i^T$ ($i = 1, \dots, m$), 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^*) which 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, 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 3α (3 standard deviation units), hence they are accepted as Y outlier. Test set compounds where ($h_i > h^*$) 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 ($h_i > h^*$) in developing the model.

6 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).

7 QSAR on K-562 cell line dataset

7.1 K-562 cell line

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

$$N_{\text{train}} = 90, \quad R_{\text{train}}^2 = 0.915, \quad R_{\text{adjusted}}^2 = 0.902,$$

$$F_{\text{train}} = 69.298, \quad Q_{\text{LOO}}^2 = 0.845, \quad \text{Outliers} > 3.0 = 5,$$

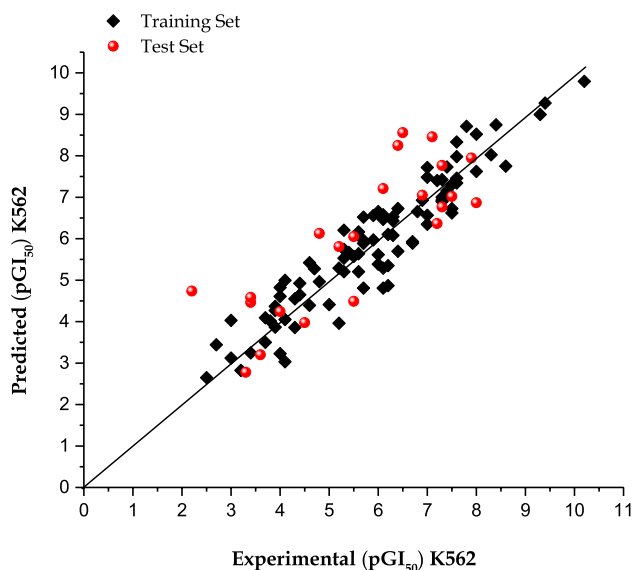
$$N_{\text{test}} = 22$$

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

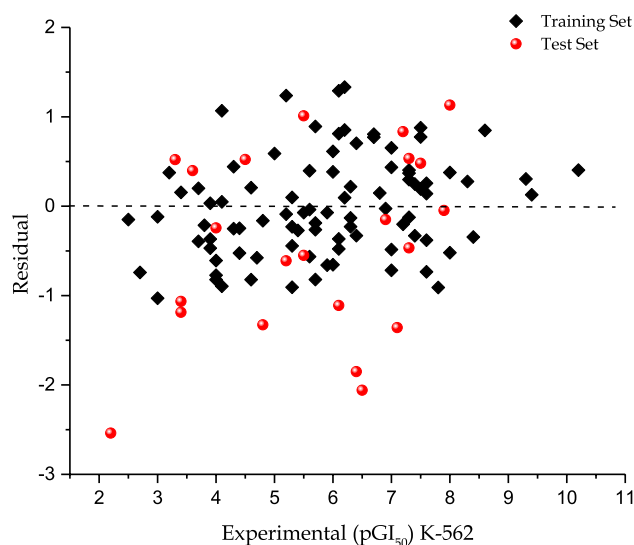
The built model was used to predict the test set data, and the results are presented in Table 1. The predicted pGI_{50} values for the compounds in the training and test sets for K-562 leukaemia cell line were plotted against the experimental pGI_{50} values in Fig. 1. Likewise, the plot of the residuals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

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 ² Test (100% data) | 0.6722 | |
| | R0 ² Test (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 ² Test (95% data) | 0.7390 |
| | | R0 ² Test (95% data) | 0.7205 |
| | | Q2F1 (95% data) | 0.9397 |
| Q2F2 (95% data) | | 0.6862 | |
| ScaledAvgRm2 (95% data) | | 0.6509 | |
| ScaledDeltaRm2 (95% data) | | 0.0601 | |
| Error-based metrics (for 100% data) | CCC (95% data) | 0.8507 | |
| | RMSEP (100% data) | 1.1011 | |
| | SD (100% data) | 0.6363 | |
| | SE (100% data) | 0.1357 | |
| Basic data structure information | MAE (100% data) | 0.9088 | |
| | <i>N</i> compound test | 22 | |
| | Result (MAE-based criteria applied on 95% data) | Prediction quality | Moderate |

**Fig. 1** The predicted pGI₅₀ against the experimental values for the training and test sets of K562 leukaemia cell line

experimental pGI₅₀ estimations is presented in Fig. 2. As can be seen from Table 1 and Figs. 1 and 2, the computed values for the pGI₅₀ are in great concurrence with those of the test set, hence the model did not demonstrate any relative and

**Fig. 2** The residuals against the predicted pGI₅₀ values for the training and test sets of K-562 leukaemia cell line

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

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 | | |

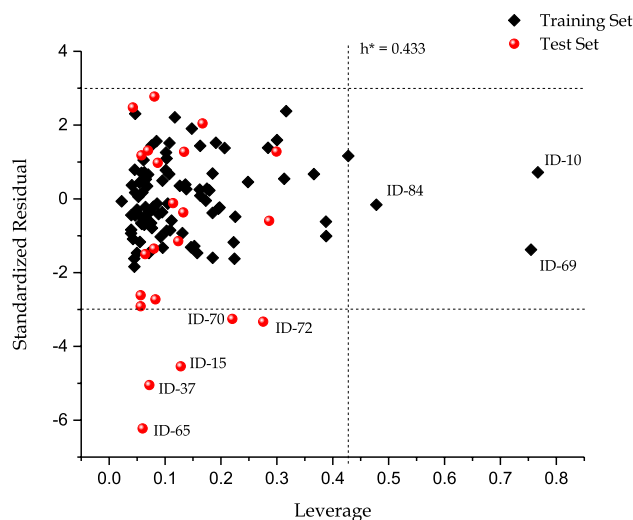
for the external validation R^2_{pred} , Q^2F_1 and Q^2F_2 values were reported in Table 3 as 0.672, 0.916 and 0.581. The result justifies that the classic metric test for 100% developed by Roy et al. (2015a) for a QSAR model biasness test is good and in well agreement with other standards stated by Tropsha and Golbraikh (Golbraikh and Tropsha 2002).

7.2 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 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 dataset 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^* ($h^* = 3p^o/n$), 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 Fig. 3, establishes applicability domain of the model within $\pm 3d$ and a leverage threshold $h^* = 0.433$.

The Williams plot for K-562 dataset 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

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

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 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.

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 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 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 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 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 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 et al. 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 et al. 1998), polar surface area (PSA) and maximum hydrogen electropological state (hmax) (Hall and 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.

7.3 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 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 (2009), which is defined as Broto–Moreau autocorrelation—lag 7/weighted by Sanderson electronegativities.

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

where, w_i and w_j are the weights of the atoms i and j , $w \in (m, p, e, v)$, and δ_{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 and Devillers 1990; Broto et al. 1984; Moreau and Broto 1980a, b).

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

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

descriptor is weighted by the charges on the molecule unlike AT57e 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 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 are negatively correlated to the activity of the molecules contributes the least to the model. Roy and Mitra (Ojha et al. 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 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.

8 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 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 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 (-CH₂-), 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 Tables 5 and 6 shows a correlation between the activity of the newly designed compounds with the mean effect values of *hmax* and PSA. *pGI*₅₀ 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.

9 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,

Table 5 Molecular descriptor values and calculated pGI₅₀ 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 |
| CA5 | 0 | 0.639 | 0.115 | 0.691 | 0 | 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 | 0.319 | 0 | 10.233 |
| CA7 | 0 | 0.639 | 0.155 | 0.679 | 0 | 0 | 0 | 0 | 0.806 | 0.165 | 0.319 | 0 | 7.919 |
| CA8 | 0 | 0.639 | 0.206 | 0.52 | 0 | 0 | 0 | 0 | 0.786 | 0.165 | 0.319 | 0 | 8.056 |
| CA9 | 0 | 0.639 | 0.242 | 0.576 | 0 | 0 | 0 | 0 | 0.769 | 0.165 | 0.319 | 0 | 7.787 |
| CA10 | 0 | 0.639 | 0.298 | 0.626 | 0 | 0 | 0 | 0 | 0.751 | 0.165 | 0.319 | 0 | 7.413 |
| CA11 | 0 | 0.639 | 0.348 | 0.554 | 0 | 0 | 0 | 0 | 0.713 | 0.165 | 0.319 | 0 | 7.441 |
| CA12 | 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

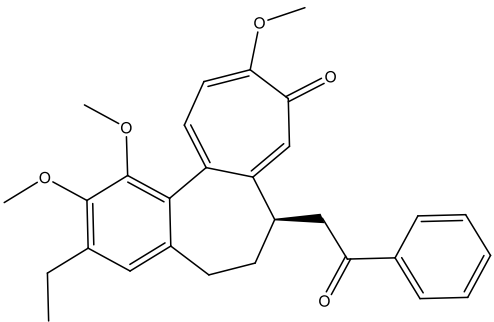
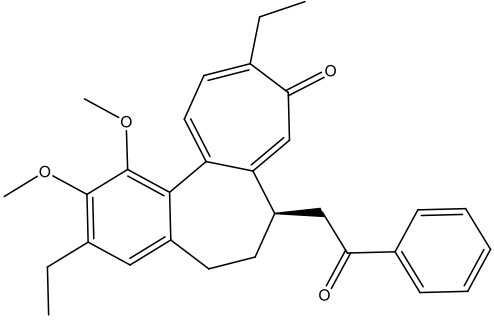
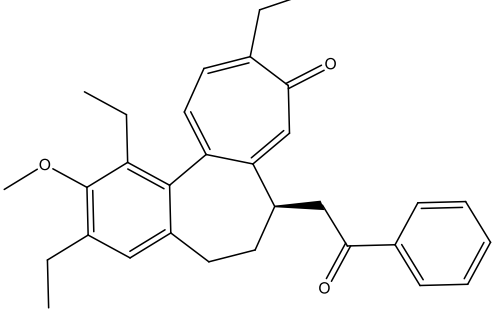
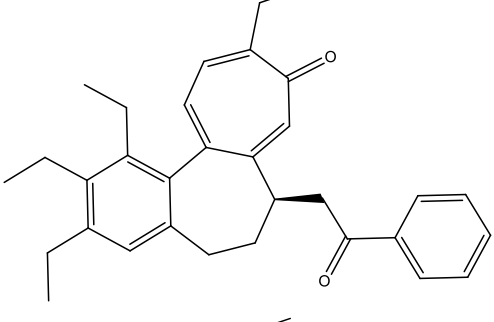
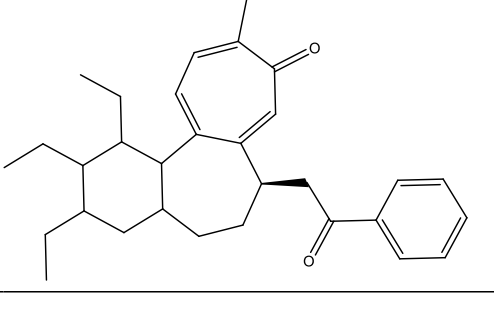
| | Compound ID | Newly designed drugs | Predicted pGI_{50} |
|---|-------------|--|----------------------|
| 1 | CD1 |  | 12.391 |
| 2 | CD2 |  | 12.516 |
| 3 | CD3 |  | 12.972 |
| 4 | CD4 |  | 13.302 |
| 5 | CD5 |  | 12.513 |

Table 6 (continued)

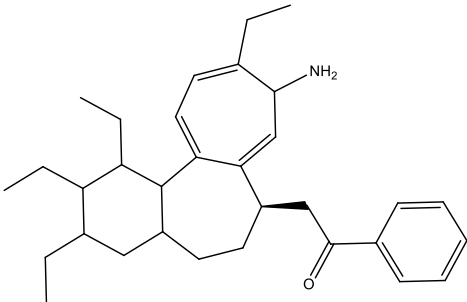
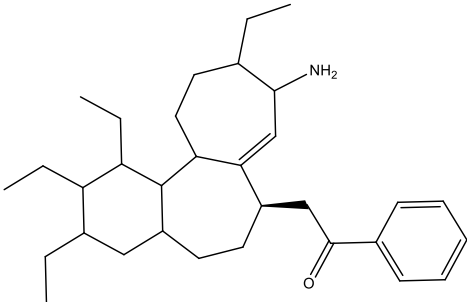
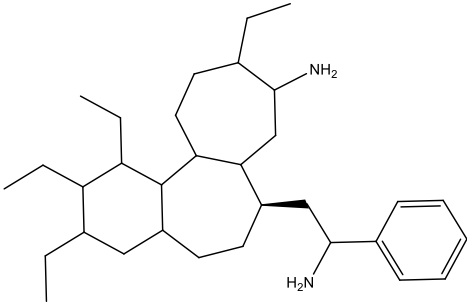
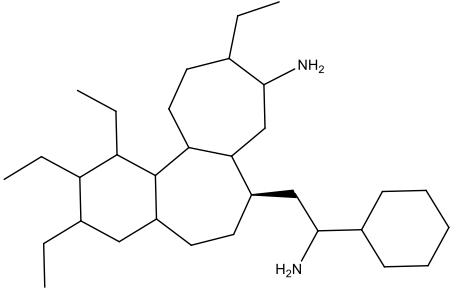
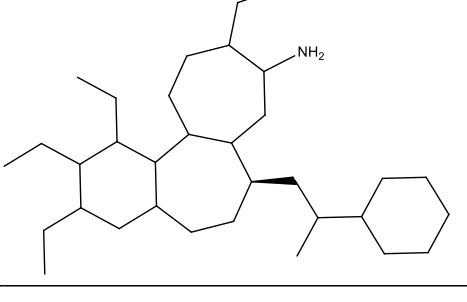
| | Compound ID | Newly designed drugs | Predicted pGI ₅₀ |
|----|-------------|--|-----------------------------|
| 6 | CD6 |  | 10.322 |
| 7 | CD7 |  | 10.560 |
| 8 | CD8 |  | 9.244 |
| 9 | CD9 |  | 13.182 |
| 10 | CD10 |  | 13.242 |

Table 6 (continued)

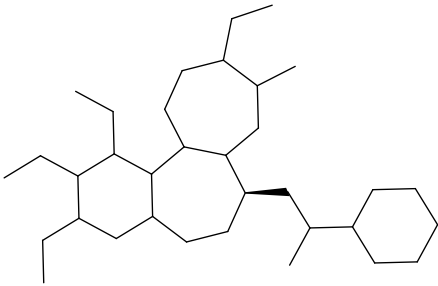
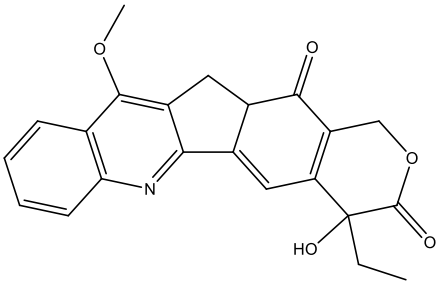
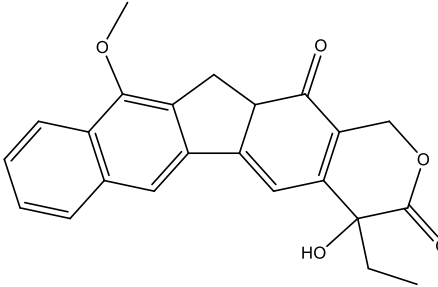
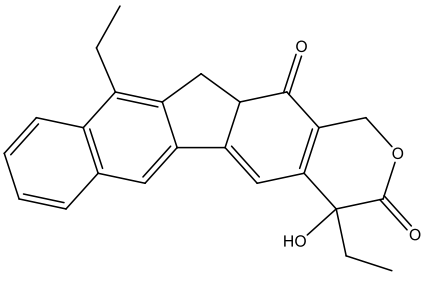
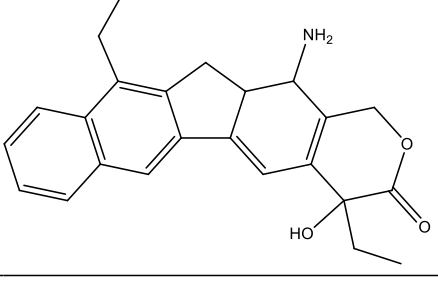
| | Compound ID | Newly designed drugs | Predicted pGI ₅₀ |
|----|-------------|--|-----------------------------|
| 11 | CD11 |  | 14.437 |
| 12 | CA1 |  | 8.263 |
| 13 | CA2 |  | 10.553 |
| 14 | CA3 |  | 9.553 |
| 15 | CA4 |  | 10.314 |

Table 6 (continued)

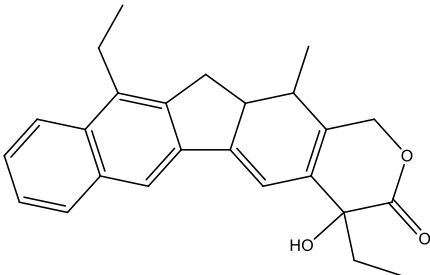
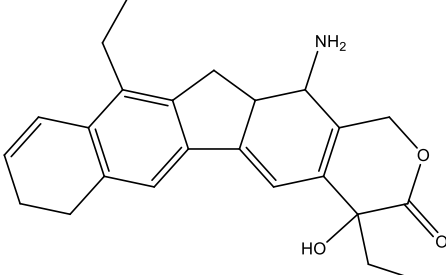
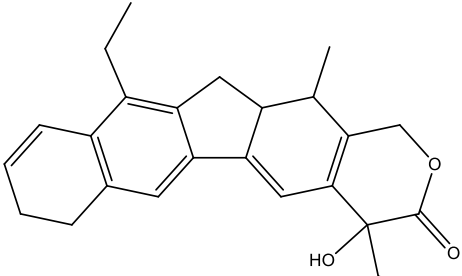
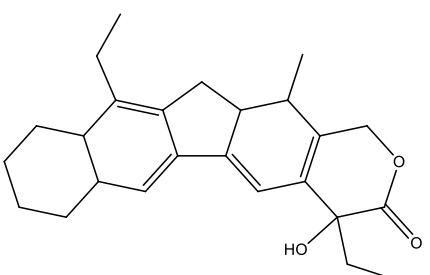
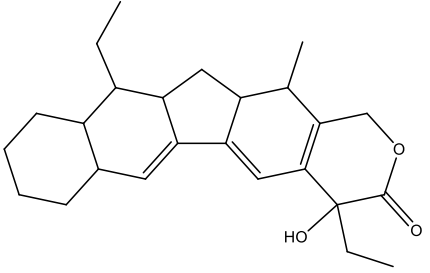
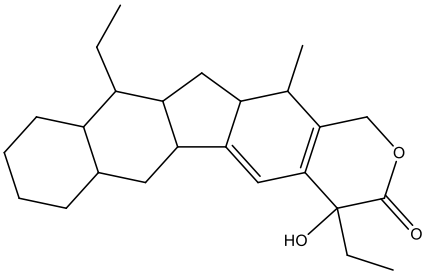
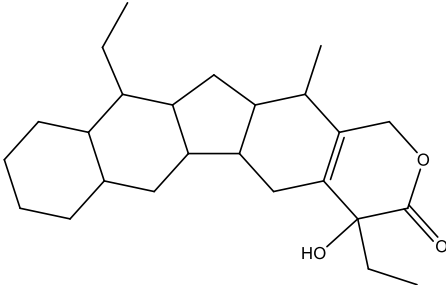
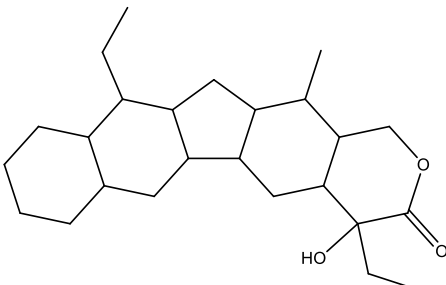
| | Compound ID | Newly designed drugs | Predicted pGI ₅₀ |
|----|-------------|--|-----------------------------|
| 16 | CA5 |  | 8.100 |
| 17 | CA6 |  | 10.233 |
| 18 | CA7 |  | 7.919 |
| 19 | CA8 |  | 8.056 |
| 20 | CA9 |  | 7.787 |

Table 6 (continued)

| | Compound ID | Newly designed drugs | Predicted pGI ₅₀ |
|----|-------------|---|-----------------------------|
| 21 | CA10 |  | 7.413 |
| 22 | CA11 |  | 7.441 |
| 23 | CA12 |  | 7.766 |

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 as a decrease in the total polar surface area (PSA) for any compound is required to improve the pGI₅₀ of anticancer agents.

Acknowledgements 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>).

References

- Andersson LC, Nilsson K, Gahmberg CG (1979) K562—a human erythroleukemic cell line. *Int J Cancer* 23(2):143–147
- Broto P, Devillers J (1990) Autocorrelation of properties distributed on molecular graphs. Kluwer Academic Publishers, Dordrecht
- 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. *Eur J Med Chem* 19(1):71–78
- Chopade SM, Phadnis PP, Hodage AS, Wadawale A, Jain VK (2015) Synthesis, characterization, structures and cytotoxicity of platinum(II) complexes containing dimethylpyrazole based selenium ligands. *Inorg Chim Acta* 427:72–80. <https://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. *Univ Chem Educ* 5(1):1–8
- Dimić D, Mercader AG, Castro EA (2015) Chalcone derivative cytotoxicity activity against MCF-7 human breast cancer cell QSAR study. *Chemometr Intell Lab Syst* 146:378–384. <https://doi.org/10.1016/j.chemolab.2015.06.011>
- Drexler HG (2000) The leukemia-lymphoma cell line factsbook. Academic Press, Cambridge
- Evans DA (2014) History of the Harvard ChemDraw project. *Angew Chem Int Ed* 53(42):11140–11145
- Fan Y, Lu H, An L, Wang C, Zhou Z, Feng F, Zhao Q et al (2016) Effect of active fraction of *Eriocaulon sieboldianum* on human leukemia K562 cells via proliferation inhibition, cell cycle arrest and apoptosis induction. *Environ Toxicol Pharmacol* 43:13–20
- Fatemi MH, Heidari A, Gharaghani S (2015) QSAR prediction of HIV-1 protease inhibitory activities using docking derived

- molecular descriptors. *J Theor Biol* 369:13–22. <https://doi.org/10.1016/j.jtbi.2015.01.008>
- Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J 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 Koc prediction. *J Mol Graph Model* 25(6):755–766
- Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci* 35(6):1039–1045
- Hebre WJ, Huang WW (1995) Chemistry with computation: an introduction to SPARTAN. Wavefunction, Inc, Irvine
- Iuliano A, Strianese D, Uccello G, Diplomatico A, Tebaldi S, Bonavolontà G (2012) Risk factors for orbital exenteration in periocular basal cell carcinoma. *Am J Ophthalmol* 153(2):238–241.e231
- Kar S, Roy K (2012) QSAR of phytochemicals for the design of better drugs. *Expert Opin Drug Discov* 7(10):877–902. <https://doi.org/10.1517/17460441.2012.716420>
- Karimiani EG, Marriage F, Merritt AJ, Burthem J, Byers RJ, Day PJ (2014) Single-cell analysis of K562 cells: an imatinib-resistant subpopulation is adherent and has upregulated expression of BCR-ABL mRNA and protein. *Exp Hematol* 42(3):183–191.e185
- Kennard RW, Stone LA (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 . *J Chem Inf Comput Sci* 38(3):387–394
- Lozzio CB, Lozzio BB (1975) Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* 45(3):321–334
- Marx KA, 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. *J Chem Inf Comput Sci* 43(5):1652–1667
- Moreau G, Broto P (1980a) The auto-correlation of a topological-structure—a new molecular descriptor, vol 4. Gauthier-Villars, Paris Cedex, pp 359–360
- Moreau G, Broto P (1980b) Auto-correlation of molecular-structures, application to SAR studies. *Nouveau J Chim N J Chem* 4(12):757–764
- News B (2003) Cancer number one killer of men. *Health*. <http://news.bbc.co.uk/2/hi/health/3019801.stm>. Accessed 18 Apr 2016
- Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring r m 2 metrics for validation of QSPR models. *Chemometr Intell Lab Syst* 107(1):194–205
- Parkin DM, Boyd L, Walker L (2011) 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *Br J Cancer* 105:S77–S81
- Rajer-Kanduć 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. *Chemometr Intell Lab Syst* 65(2):221–229
- Roy K, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN (2013) Some case studies on application of “ rm^2 ” metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *J Comput Chem* 34(12):1071–1082
- Roy K, Kar S, Das RN (2015a) A primer on QSAR/QSPR modeling: fundamental concepts. Springer, Berlin
- Roy K, Kar S, Das RN (2015b) Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment. Academic Press, Cambridge
- Roy K, Kar S, Ambure P (2015c) On a simple approach for determining applicability domain of QSAR models. *Chemometr Intell Lab Syst* 145:22–29. <https://doi.org/10.1016/j.chemolab.2015.04.013>
- Sabet R, Mohammadpour M, Sadeghi A, Fassihi A (2010) QSAR study of isatin analogues as in vitro anti-cancer agents. *Eur J Med Chem* 45(3):1113–1118
- Saportó A, Tadé MO, Vuthaluru H (2012) A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chem Prod Process Model* 7(1):1–14
- Siegel RL, Miller KD, Jemal A (2015) Cancer statistics, 2015. *CA* 65(1):5–29
- Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MND (2012a) Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents. *Bioorg Med Chem* 20(15):4848–4855
- Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MNDS. (2012b) Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in silico discovery of anti-breast cancer agents. *Eur J Pharm Sci* 47(1):273–279. <https://doi.org/10.1016/j.ejps.2012.04.012>
- Supratik Kar KR (2010) Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. *Indian J Biochem Biophys* 48:111–122
- Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics, vol 41. Wiley, Hoboken
- Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inf* 29(6–7):476–488
- Tropsha A, Gramatica P, Gombar VK (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 VN, Ghose AK, Revankar GR, Robins RK (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. *J Chem Inf Comput Sci* 29(3):163–172
- World Health Organization (2002) National cancer control programmes: policies and managerial guidelines. World Health Organization
- Wu X, Fini P, Keller S, Tarsa E, Heying B, Mishra U, Speck J et al (1996) Morphological and structural transitions in GaN films grown on sapphire by metal-organic chemical vapor deposition. *Jpn J Appl Phys* 35(12B):L1648
- Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32(7):1466–1474