Structural Chemistry (2020) 31:1043–1055
https://doi.org/10.1007/s11224-019-01481-z

ORIGINAL RESEARCH

Exploration of nitroimidazoles as radiosensitizers: application of multilayered feature selection approach in QSAR modeling

Priyanka De[1] · Dhananjay Bhattacharyya[2] · Kunal Roy[1]

Received: 25 November 2019 / Accepted: 19 December 2019 / Published online: 10 January 2020

Abstract

Radiosensitizers are aimed to augment tumor cell killing by radiation while having much less effect on normal tissues. Nitroimidazoles and related analogues are efficient radiation sensitivity enhancers, and they particularly work on hypoxic tumor cells. In the current study, we have developed two partial least squares (PLS) regression-based two-dimensional quantitative structure-activity relationship (2D-QSAR) models using a novel class of 84 nitroimidazole compounds to understand their radiosensitization effectiveness ($pC_{1.6}$). Feature selection was done by genetic algorithm along with stepwise regression, while model validation was performed using various stringent validation criteria following the strict rules of OECD guidelines of QSAR validation. The variables included in the models were obtained from Dragon (version 7.0) and simplex representation of molecular structures (SiRMS) (version 4.1.2.270) software. The developed models were robust, externally predictive, and useful tools to predict the radiosensitization effectiveness of nitroimidazole compounds. True external prediction was carried out using a group of six nitroimidazole derivatives and the model reliability was checked using the ***Prediction Reliability Indicator*** tool (http://dtclab.webs.com/software-tools). Furthermore, the developed models will give an insight for development of new radiosensitizers with enhanced radiation sensitivity.

Keywords Radiosensitizers · Radiosensitization effectiveness · QSAR · SiRMS

Introduction

Radiation, surgery, and chemotherapy have been the major approaches of treatment for cancer and malignancies for more than 40 years. Combination therapy including radiation and chemotherapy often termed as chemoradiation has provided promising results in targeting, diagnosis, and treatment of human malignancy. With recent discoveries, newer molecules targeting specific pathophysiology or molecular pathways have come into the forefront. The use of antibodies or hormones labeled with radionuclides to deliver radiation in the systemic circulation has enlarged the concept of radiosensitizers [1]. Nitroimidazoles have proven to be efficient radiation sensitivity enhancer particularly in hypoxic tumor cells [2]. Hypoxia is a particular pathophysiological condition arising due to inefficient vascularization of tumors, causing an alteration in tumor metabolism [3], and metastasis [4], and is associated with poor diagnosis and resistance to therapeutic agents [5]. Nitroimidazole radiosensitizers are relatively non-toxic molecules, and they replace oxygen in oxidizing radiation-induced DNA free radicals to generate cytotoxic DNA strand breakage [6].

A number of studies performed previously have elaborately explained the role of nitroimidazole derivatives in radiation sensitivity enhancement. 1-Methyl-5-sulfonamide-4-nitroimidazole (MJL-1–191-VII) sensitizes hypoxic cells with its electron affinity, but does not affect the radiosensitivity of aerated cells when added to cells 5 min prior to irradiation [7]. 2-nitroimidazoles like misonidazole and etanidazole has ability to kill hypoxic cell by increasing the cells' radiation sensitivity via radiochemical and biochemical means known as "preincubation effect" [8].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11224-019-01481-z) contains supplementary material, which is available to authorized users.

✉ Kunal Roy
kunalroy_in@yahoo.com; kunal.roy@jadavpuruniversity.in

[1] Department of Pharmaceutical Technology, Drug Theoretics and Cheminformatics Laboratory, Jadavpur University, Kolkata 700032, India

[2] Computational Science Division, Saha Institute of Nuclear Physics, Kolkata 700064, India

Molecular modeling studies such as quantitative structure-activity relationships (QSAR) [9] are effective tools in prediction of radiosensitization effectiveness due to lack of data and proper experimental facilities. QSAR studies have found immense applications in the prediction of absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties of drug and other organic biologicals [10–12]. Computational ADMET in combination with in vivo and in vitro predictions helps in reducing the chances of safety related issues [13]. Many pharmaceutical and chemical industries, commercial software developers, and research groups are developing new QSAR models for ADMET properties utilizing large databases or compilation of published data. A wide number of computational research work describing oral absorption and bioavailability [14, 15], metabolism [16], volume of distribution [17], and enzyme inhibition and induction [18, 19] have been carried out in recent years. The theory of QSAR is applied not only to model activity and toxicity, but also properties of materials in the form of quantitative structure-property relationships (QSPR). Radiosensitization effectiveness can be considered as a property of the nitroimidazole compounds and can thus be subjected to QSAR analysis. Many such property based QSAR models for radiopharmaceuticals have been developed previously by different groups of researchers [20–24]. A properly validated QSAR model could generate radiosensitization data for groups of such related chemicals, and such predictions have the ability to substitute experimental evaluation to an extent.

Feature selection is an essential step for unbiased development of QSAR models. The selection of a reduced pool of descriptors by using multilayered variable selection strategy has proven to be an effective method in QSAR model development and easier data handling. Furthermore, feature selection can reduce the chances of intercorrelation among the descriptors [25]. The current study presents QSAR models for predicting the radiosensitization effectiveness of a dataset of 84 nitroimidazole derivatives. Two-dimensional descriptors calculated from Dragon and SiRMS software were capable enough in developing well-validated and predictive models. Simplex representation of molecular structures (SiRMS) descriptors helped in providing a comprehensive understanding of the basic fragments contributing towards the improvement of radiosensitization effectiveness of the nitroimidazole derivatives. The 2D-QSAR models were developed with an intention of producing statistically robust predictions for radiosensitization effectiveness of nitroimidazole derivatives. Furthermore, we have also predicted some related nitroimidazole compounds to prove the validity of the developed models.

Materials and methods

A data of 86 nitroimidazoles possessing radiosensitizing properties are used for two-dimensional QSAR (2D-QSAR) study [26]. Radiosensitization capacities of the compounds can be understood by radiosensitization effectiveness, expressed as $C_{1.6}$, which can be represented as the corresponding concentration of a given compound when its sensitization enhancement ratio (SER) accomplishes 1.6. Higher value of $C_{1.6}$ indicates lower bioactivity of radiosensitization effectiveness. For analysis purpose, the source literature had converted the endpoint $C_{1.6}$ to its negative logarithmic scale ($pC_{1.6,}$ where $pC_{1.6} = -\log(C_{1.6})$). Two compounds (one radical and one salt) were removed, and the final dataset of 84 compounds is used for model development. The structures of the compounds were drawn in MarvinSketch software (version 14.10.27) [27] with proper aromatization and hydrogen bond addition and saved as MDL.mol, a recommended format for further descriptor calculation.

Descriptor calculation

For developing the first 2D-QSAR model, a pool of 270 descriptors was calculated using Dragon version 7 [28] software. This model was developed using specific classes of descriptors including E-state indices, connectivity, constitutional, functional, 2D atom pairs, ring, atom-centered fragments and molecular property descriptors. Additionally, SiRMS descriptors were calculated using SiRMS (version 4.1.2.270) [29] tool. Simplex representations of molecular structure (SiRMS) descriptors symbolize a class of diverse molecular features developed from 1D to 4D molecular structures. These are tetratomic fragments of different simplex descriptors having predefined chirality, composition, and symmetry [29]. SiRMS descriptors consider both connected and unconnected fragments and also take into account not only the nature of atoms but also their different chemical and physical properties like charge, lipophilicity, electronegativity, atomic refraction, donor/acceptor of hydrogen in the potential Hbond, etc. In our study, we have used 2D SiRMS descriptors only in order to avoid conformational complexity and energy minimization requirements for higher dimensional descriptors and to derive reproducible models. The constant (variance < 0.0001), intercorrelated ($|r|$ >0.95) variables and other incompetent data were removed using an in house software available at http://dtclab.webs.com/software-tools before model development.

Dataset splitting

A well-validated QSAR model is the main objective of any QSAR study which can be obtained through proper division of the dataset into training (used for model development) and test (used for model validation) sets. An unbiased external validation with uniform distribution of compounds into training and test sets can be obtained through rational dataset division [30]. For 2D-QSAR modeling, the whole dataset utilized for modeling was divided into training (75%) and test (25%)

sets using modified *k*-Medoids (Modified *k*-medoid GUI 1.3) [31, 32] method of dataset division.

Variable selection and QSAR model development

Development of well-validated QSAR models in order to understand the radiosensitization effectiveness of the dataset compounds was the main aim of the present study. Critical evaluation process helped in the selection of statistically significant models. In this study, we have built two QSAR models; a 2D-QSAR model to deduce a relationship between the molecular properties of the nitroimidazoles and their radiosensitization properties. For the model with Dragon descriptors, a pool of 32 descriptors were selected using Genetic Algorithm (GA) [33, 34] modeling implemented in double cross-validation (DCV) [35] tool (version 1.2). Then, the final model was generated using Partial Least Squares (PLS) regression [33, 36] method using descriptors selected from best subset selection (BSS). In case of SiRMS, the number of descriptors generated was large, i.e., about more than ten thousand. Handling of this large data is very much complicated, and so we have applied stepwise regression on the large pool of SiRMS descriptors to find out the essential descriptors contributing to the radiosensitization properties of the dataset. After descriptor thinning, the obtained pool of 300 descriptors was further subjected to multilayered stepwise regression to obtain a manageable number of descriptors and run best subset selection for development of five descriptors models. From the developed models obtained after best subset selection, we have selected one model based on different validation parameters for the test set. Finally, we have run a partial least squares regression (PLS) using SIMCA-P software [37] and developed a PLS model.

Statistical validation metrics

We have rigorously examined the statistical quality of the derived models to judge the robustness in terms of reliability and predictivity measures using various internal and external validation parameters. In the present work we have computed various statistical parameters like determination coefficient R^2, explained variance R_a^2, variance ratio (F), and standard error of estimate (s). Since these quality parameters are not sufficient to assess the predictive ability of the model, we have further used additional parameters that could properly validate our predictions. For internal predictions, leave-one-out cross-validation ($Q^2_{(LOO)}$) was reported, and for external predictions, parameters like R^2_{pred} or Q^2_{F1},Q^2_{F2} and concordance correlation coefficient (CCC), were calculated [38]. We have also calculated r_m^2 metrics (i.e., $\overline{r_m^2}$ and Δr_m^2) for both training and test set compounds [39]. We have also validated the models using mean absolute error (MAE) based criteria for both external and internal validation [40]. This was done since the Q^2_{ext} based criteria do not always offer the correct indication of the prediction quality because of the influence of the response range as well as the distribution of the values of response in both the training and test set compounds [40].

Results and discussion

Statistically significant 2D-QSAR models using Dragon and simplex (SiRMS) descriptors explaining the chemical features required for good radiosensitization are presented in the following section. The observed versus predicted $pC_{1.6}$ values are plotted for both the models is shown in Fig. 1**.**

2D-QSAR model using dragon descriptors

$$
\begin{aligned}
pC_{1.6} = {} & 3.612 + 0.613 \times (C{-}035) - 0.285 \times nCp - 1.129 \\
& \times (C{-}043) + 0.068 \times (H{-}052) - 1.630 \times (C{-}042) \\
& + 0.295 \times \text{nRNHR}.
\end{aligned}
$$

$$
\begin{aligned}
N_{train} = {} & 63, R^2 = 0.773, R^2_{adj} = 0.757, Q^2_{(LOO)} \\
& = 0.746, r^2_{m(Train)} = 0.647, \Delta r^2_{m(Train)} \\
& = 0.173, MAE(Train) = 0.246, SD(Train) \\
& = 0.195, RMSEC = 0.30, Quality = GoodN_{test} \\
& = 21, Q^2_{F1} = 0.752, Q^2_{F2} = 0.724, r^2_{m(Test)} \\
& = 0.608, \Delta r^2_{m(Test)} = 0.216, CCC\ (Test) \\
& : 0.831, MAE(Test) = 0.240, SD(Test) \\
& = 0.204, RMSEP = 0.31, Quality = Moderate
\end{aligned}
$$

Model 1

The PLS model with 4 latent variables (LVs) could predict 74.6% variance of the training set and 75.2% of the test set. Important internal and external metrics used to determine the quality of the QSAR model are listed in eq. 1. Mechanistic interpretation of the six descriptors obtained in the model would give us an insight about the structural features of the nitroimidazoles which are likely to influence their radiosensitization effectiveness. The obtained descriptors are C-035, nCp, C-043, H-052, C-042, and nRNHR. The model contains four atom-centered fragments **C-035** (R–CX..X; positive contribution), **C-043** (X–CR..X, negative contribution), **H-052** (hydrogen (H^e) attached to sp^3 carbon (C^0) with one X

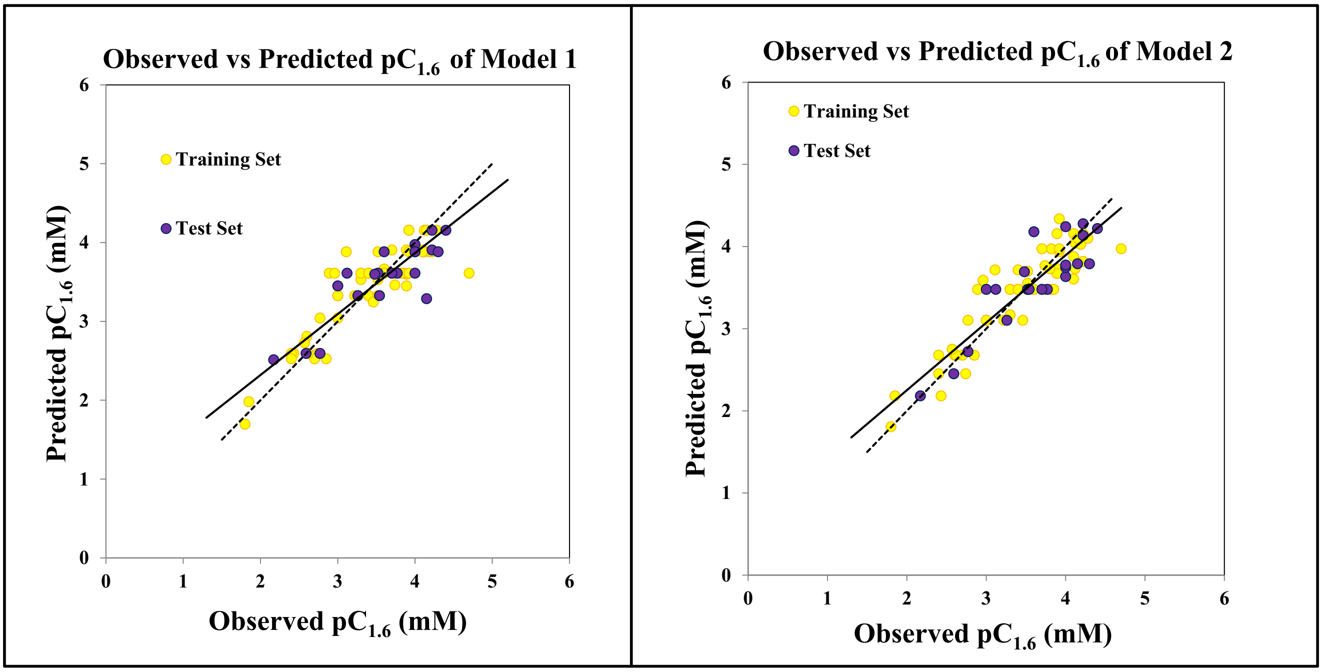


Fig. 1 Scatter plots for observed vs predicted $pC_{1.6}$values for Model 1 and Model 2

attached to next carbon, “e” represents the formal oxidation number; positive contribution) and **C-042** (X–CH..X; negative contribution). These descriptors are further explained with molecular structures from the dataset in Fig. 2. The other two descriptor belonging to functional group counts are nCp (number of terminal primary C (sp^3); negative contribution) and nRNHR (number of secondary amines (aliphatic); positive contribution). The descriptors obtained in the model gives us an idea regarding the vital features essential for better radiosensitization which includes the position of nitro group in the imidazole moiety. Atom-centered fragment-based descriptors like C-042 and C-043 could explain that presence of nitro group at position 4 and position 5 would decrease the $pC_{1.6}$.

The variable importance plot (VIP) [41] analysis gives us a premonition that C-042 and C-035 are the most important descriptors (VIP > 1) and contributing mostly towards the radiation enhancement of the compounds. The loading plot gives the relationship between the *Y* variable ($pC_{1.6}$) and the *X* variables (descriptors). For interpretation of the loading, the distance from the plot origin is considered, where similar types of descriptors with similar properties are located together. The variables which are far away from the plot origin are considered to have stronger impact on the model. This statement is verified by descriptors C-042 and C-035 which are proved to have higher impact from the VIP values also. The closeness of any descriptor to the *Y* variable signifies its higher influence on the response. The VIP and loading plot are shown in Fig. 3.

The 2D-QSAR model with Dragon descriptors gives an insight about the importance of the position of nitro group in the nitroimidazole compounds. Also it is found that the presence of secondary aliphatic amine has significant importance on radiosensitization.

2D-QSAR model using SiRMS descriptors

We have further tried to improve the quality of the model by the use of SiRMS descriptors. The obtained 2D-QSAR model using SiRMS descriptors for radiosensitization effectiveness of nitroimidazoles was highly robust in terms of the statistical parameters as the values of quality metrics were above the recommended threshold as currently practiced [39].

$$\begin{aligned} pC_{1.6} = \quad & 1.381 + 0.802 \times Fr3(elm)/CNN/12s,13a/ \\ & +0.494 \times SA(chg)/ACDD/12s,14a,34s/6 \\ & +0.004 \times SA(chg)/BCCC/14s,34s/4 \\ & +0.377 \times Fr5(type)/C.3C.ARC.ARC.ARN.AR/12s,23a,25a,45a/ \\ & +0.269 \times Fr(en)/CCCCD/15s,23s,25s,34a/ \end{aligned}$$

$$\begin{aligned} N_{train} = \quad & 63, R^2 = 0.82, R^2_{adj} = 0.81, Q^2_{(L00)} = 0.79, r^2_{m(100)} = 0.70, \Delta r^2_{m(l00)} = 0.14, \\ MAE_{train} = \quad & 0.22, SD_{train} = 0.18, RMSEC = 0.26, Quality_{(Train)} = Moderate \\ N_{test} = \quad & 21, Q^2_{F1}\left(orR^2_{pred}\right) = 0.80, Q^2_{F2} = 0.77, r^2_{m(Test)} = 0.70, \Delta r^2_{m(Test)} = 0.05, \\ CCC(Test) = 0.88, & MAE_{test} = 0.23, SD_{test} = 0.16, RMSEP = 0.28, Quality_{(Test)} \\ & = Moderate \end{aligned}$$

Model 2

The PLS equation with 3 LVs is able to predict 79% variance of the training set (Q^2) and 80% of the test set (R^2_{pred}). The various internal and external metric values obtained are given in eq. 2. The observed and predicted radiosensitization effectiveness values of the nitroimidazoles are listed in Table S1 in the Supplementary Section.

From VIP (Fig. 4) the descriptors from highest to lowest order of significance are as follows: Fr3(elm)/C_N_N/1_2s,1_3a/, S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6,

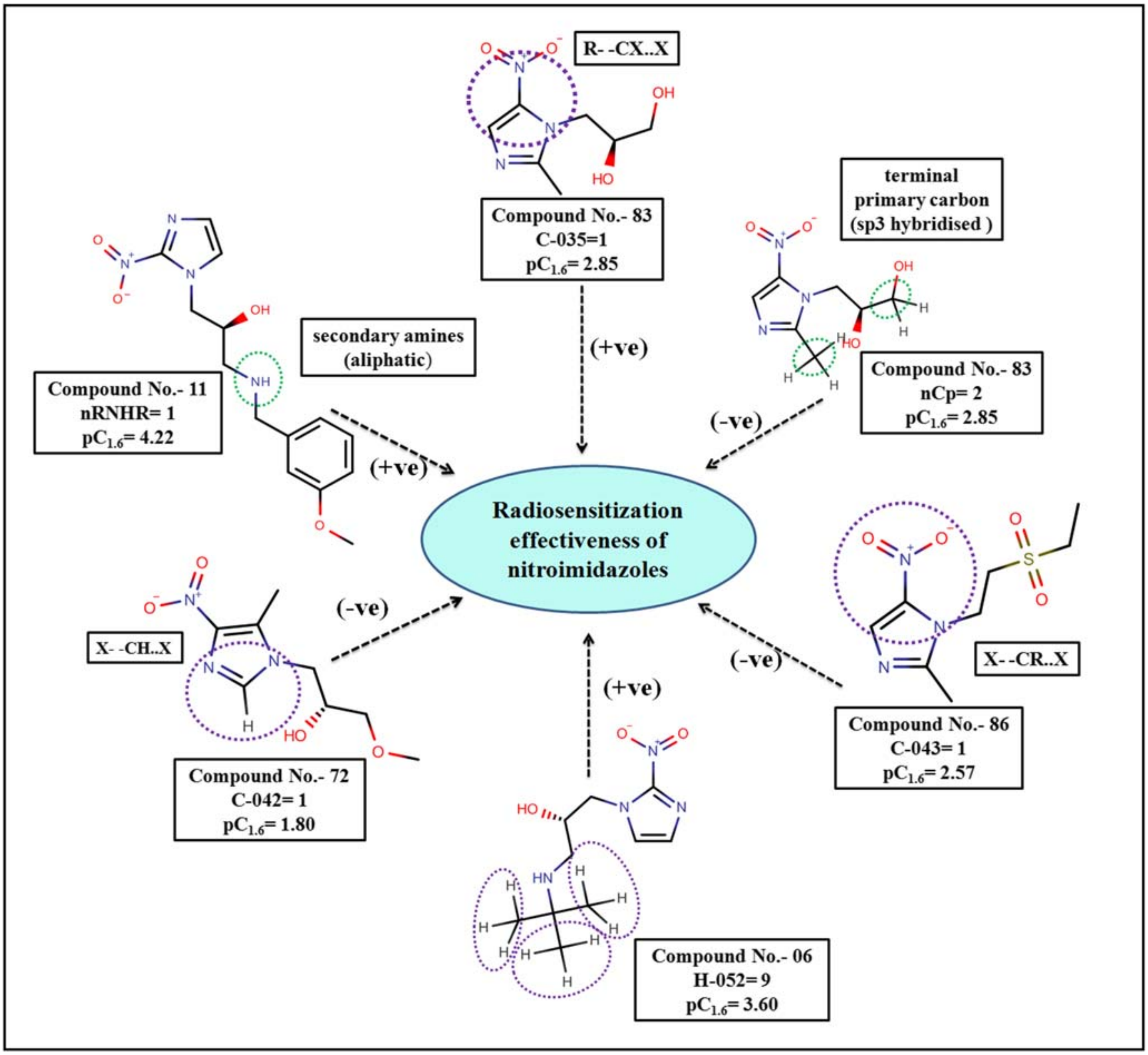


Fig. 2 Descriptor features obtained from Dragon controlling the radiosensitization effectiveness of nitroimidazoles

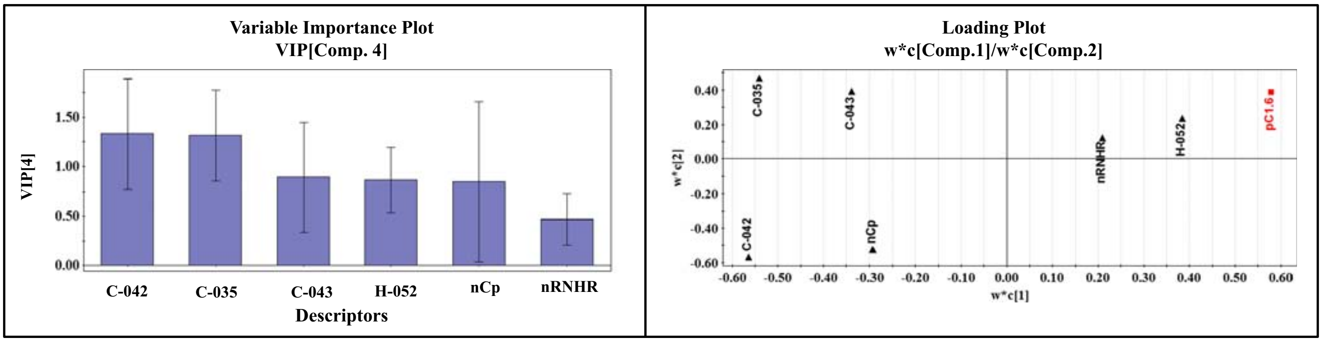


Fig. 3 VIP and loading plot of Model 1

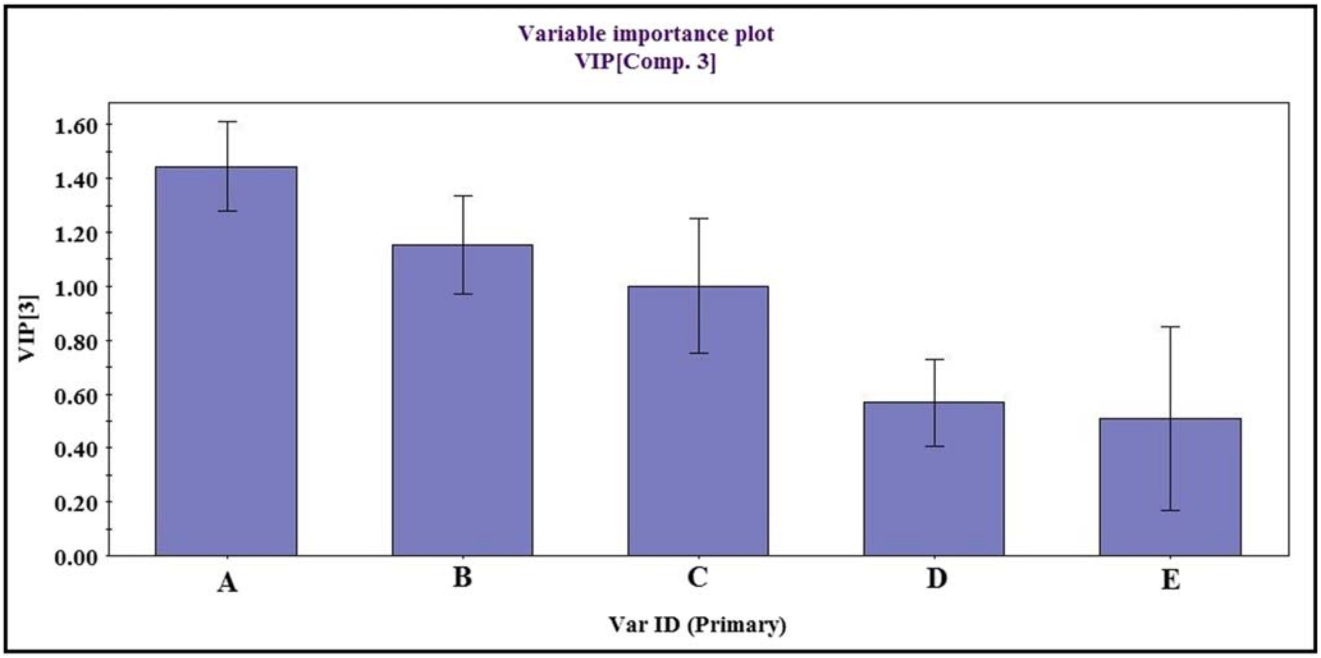


Fig. 4 Variable importance plot of SiRMS model. (A- Fr3(elm)/C_N_N/1_2s,1_3a/, B- S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, C- S_A(chg)/B_C_C_C/1_4s,3_4s/4, D- Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/, E- Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/)

S_A(chg)/B_C_C_C/1_4s,3_4s/4, Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/ and Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/. The loading plot developed using first two components describe the relationship between the *X* variables and *Y* variable is shown in Fig. 5.

The highest contributing descriptor is **Fr3(elm)/C_N_N/1_2s,1_3a/** which is a three atomic fragment depicted by N-C=N (**Box 1**). Here, the unsaturation between carbon and nitrogen takes place within the imidazole moiety and the other nitrogen is from the nitro group. This descriptor has a positive impact on the radiosensitization of the nitroimidazoles thus with higher number of such fragments increases the $pC_{1.6}$ value. All the compounds in the dataset have this particular group once or twice. Compounds with two fragments of this kind has higher pC1.6 values as prominently seen in compounds like **63**, **47**, **11**, **53**, **46**, **51**, **43**, **45**, **10**, **22**, **54**, etc. Compounds with only one fragment have considerably lower $pC_{1.6}$ values as observed in **72**, **71**, **82**, **78**, **75**, **86**, **80**, **81**, **85**,

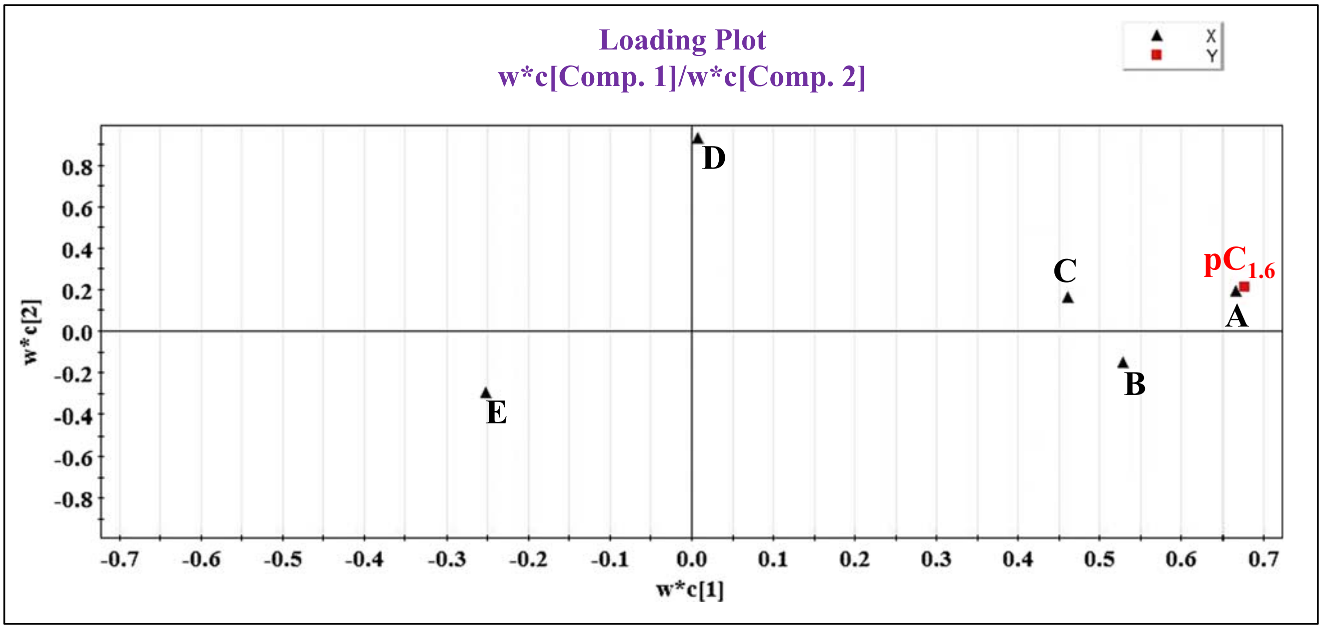


Fig. 5 Loading plot of the SiRMS model. (*A* - Fr3(elm)/C_N_N/1_2s,1_3a/, *B* - S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, *C* - S_A(chg)/B_C_C_C/1_4s,3_4s/4, *D*-Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/, *E*- Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/)

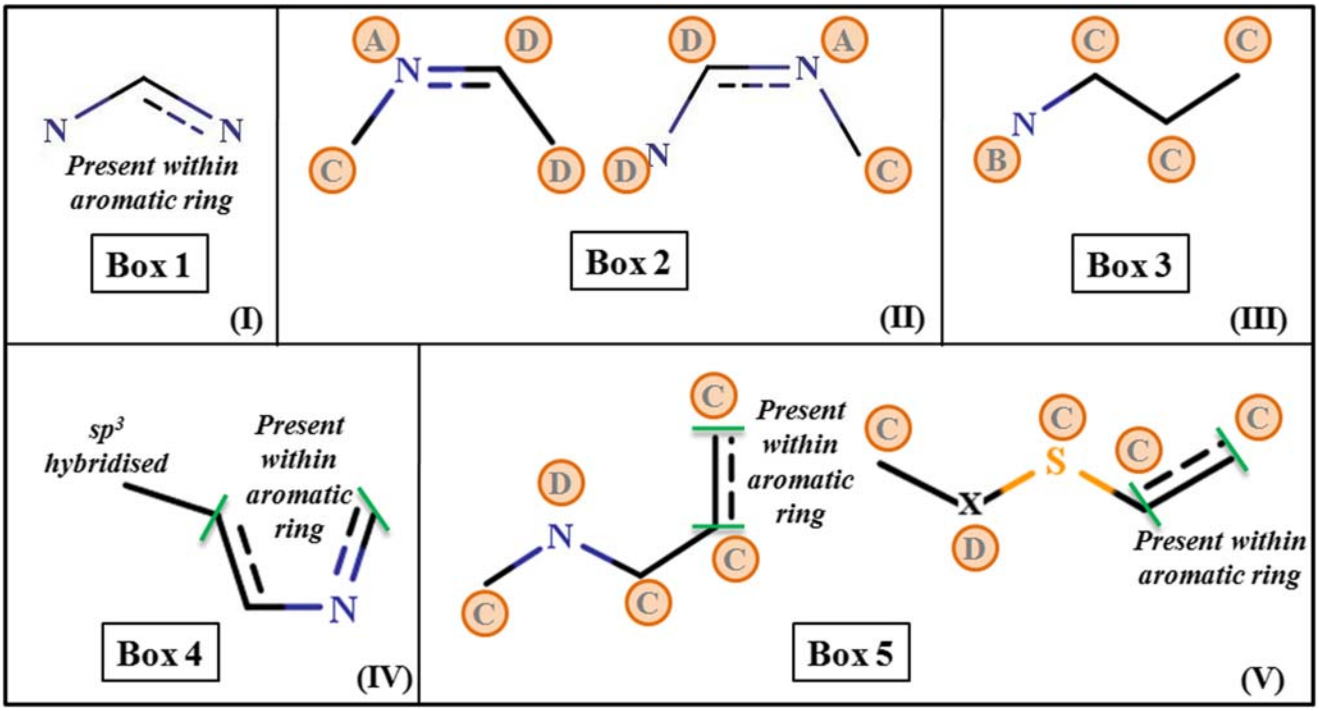

N
N
Present within aromatic ring
Box 1
(I)
A
D
N
C
D
D
A
N
D N
C
Box 2
(II)
C
C
N
B
C
Box 3
(III)
sp3 hybridised
Present within aromatic ring
N
Box 4
(IV)
C
Present within aromatic ring
D
N
C
C
C
C
C
S
C
X
D
Present within aromatic ring
Box 5
(V)

Fig. 6 Simplex representation of molecular structures (SiRMS) fragments appearing in the nitroimidazole dataset. (**I**- Fr3(elm)/C_N_N/1_2s,1_3a/, **II**- S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6, **III**- S_A(chg)/B_C_C_C/1_4s,3_4s/4, **IV**- Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/, **V**- Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/)

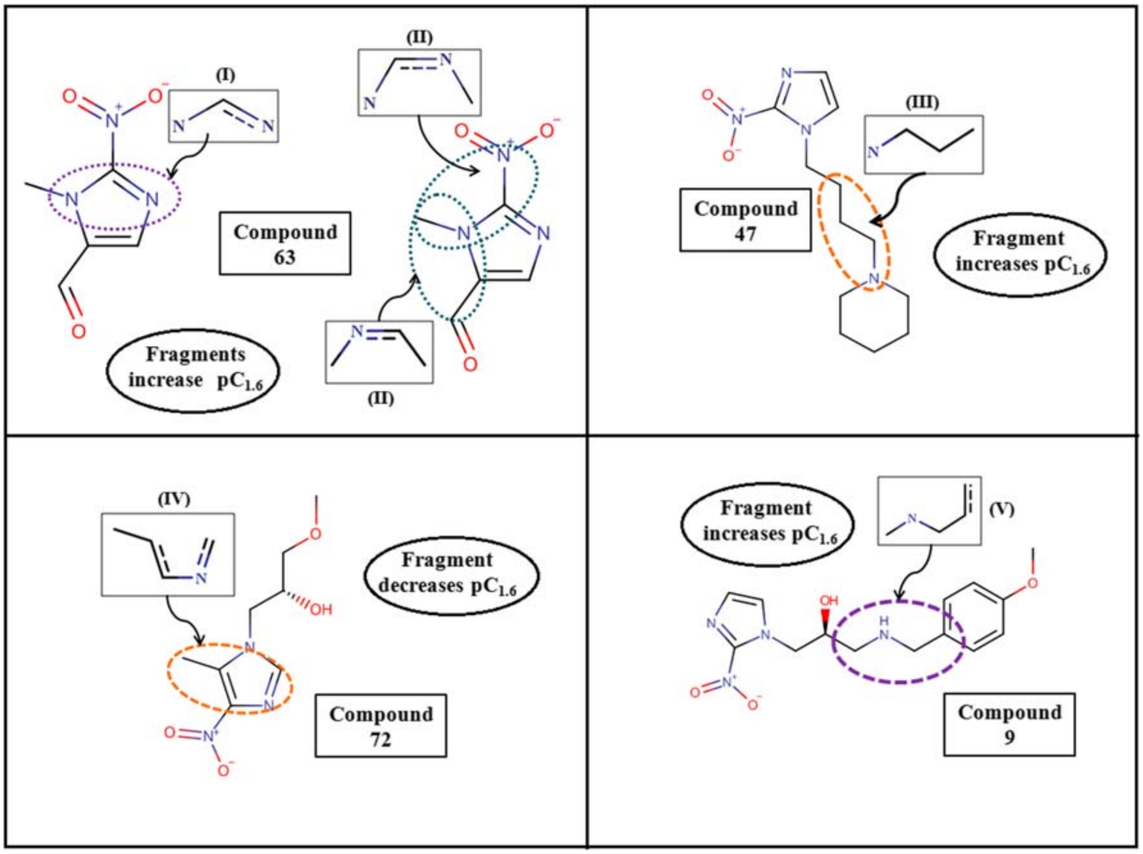


Fig. 7 SiRMS features controlling the increase or decrease in $pC_{1.6}$

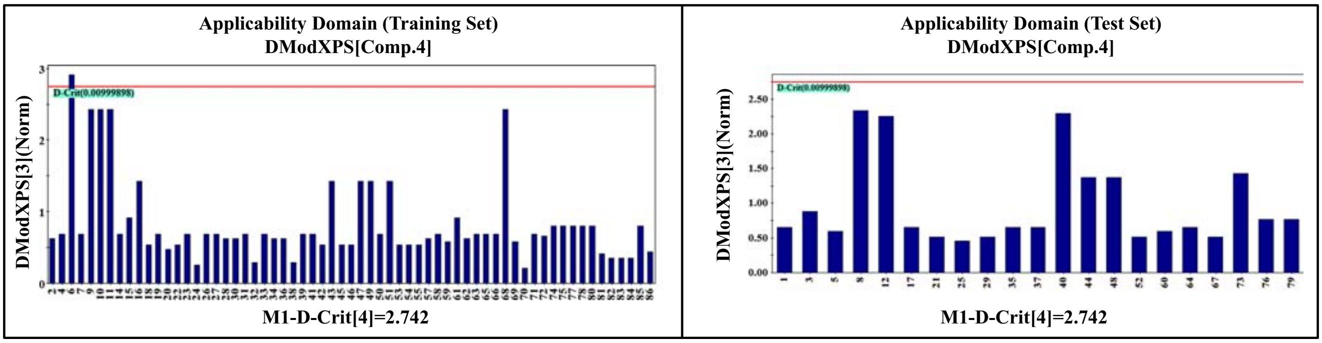


Fig. 8 Applicability Domain of training and test set of Model 1 (with Dragon descriptors) at 99% confidence level

84, etc. Thus, the importance of this fragment leads us to a conclusion that the presence of nitro groups in nitroimidazole should be between N1 and N3 positions of imidazole moiety so as to show better radiosensitization property.

The second important descriptor is **S_A(chg)/A_C_D_D/1_2s,1_4a,3_4s/6** that represents the partial charge of any of the four atom fragment as given in **Box 2**. The fragment here has two possibilities, one with single nitrogen present within the imidazole moiety and another with two nitrogens (one from the imidazole moiety and another from the nitro group) (given in Box 2). Most of the compounds having this fragment have a nitro group attached at position 2 of the imidazole ring. Thus, the position of nitro group plays a vital role in controlling the $pC_{1.6}$ value. This fragment has a positive influence on the radiosensitization effectiveness observed in compounds like **63**, **66**, **65**, **68**, **47**, **11**, and **53**. Compounds which are devoid of these kind of fragments have considerably low pC1.6 value (such as in **74**, **77**, **80**, **75**, **78**, **71**, and **72**) (Figs. 6 and 7).

The next important descriptor is **S_A(chg)/B_C_C_C/1_4s,3_4s/4** which represents the partial charge of a four atom fragments as given in **Box 3**. The presence of the mentioned fragment (i.e., three carbon chain attached to nitrogen from a cyclic nucleus) would increase the radiosensitization effectiveness due to the positive influence of the descriptor. Compounds like **47**, **51**, **43**, **46**, **55**, **49**, **54**, and **53** have higher partial charges due to the presence of the mentioned fragments thereby increasing the radiosensitization effectiveness whereas in compounds with no such fragments (like in **71**, **72**, **82**, **78**, **75**, **80**, and **81**) the effect of such charges is not observed thereby the $pC_{1.6}$ value is less.

The next important descriptor **Fr5(type)/C.3_C.AR_C.AR_C.AR_N.AR/1_2s,2_3a,2_5a,4_5a/** is a five atomic fragment signifying the following formula: C (sp^3)-C (aromatic)-C (aromatic)-C (aromatic)-N (aromatic). The structure of the possible fragment is given in **Box 4.** The presence of this type of fragment reduces the radiosensitization effectiveness as indicated by the negative influence of the descriptor on $pC_{1.6}$ value. This is well observed in compounds like **72**, **59**, **57**, **61**, **69**, **62**, **41**, and **70**. On the other hand, absence of this fragment increases the radiosensitization property as seen in compounds such as **43**, **45**, **51**, **46**, **11**, **53**, **47**, and **63.**

The descriptor with the least significance is **Fr5(en)/C_C_C_C_D/1_5s,2_3s,2_5s,3_4a/** which denotes the electronegativity of the compound due to the presence of a four atomic fragment given in **Box 5**. The positive contribution suggested that the presence of any of the given fragments will influence the electronegativity of the compound thereby increasing the $pC_{1.6}$ value. Compounds **9**, **10**, and **11** have been reported to have two such fragments and thereby increase the radiosensitization effectiveness.

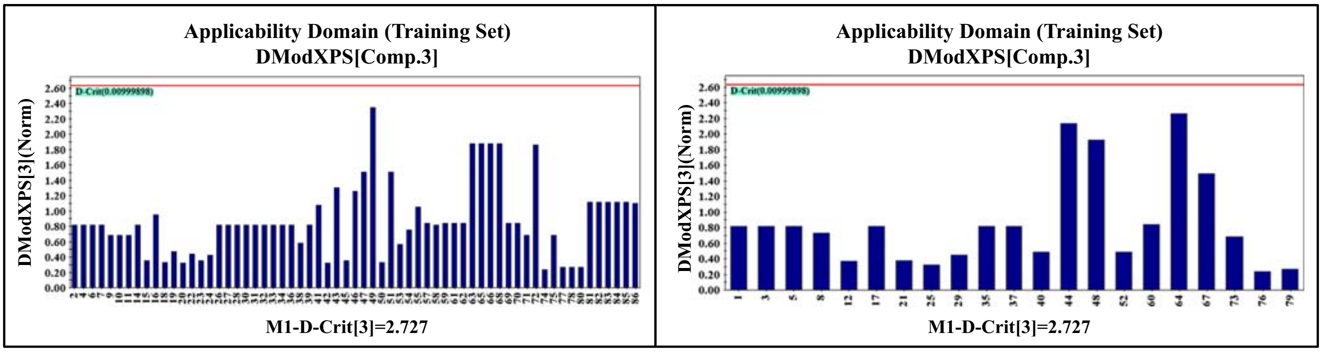


Fig. 9 Applicability Domain of training and test set of Model 2 (with SiRMS descriptor) at 99% confidence level

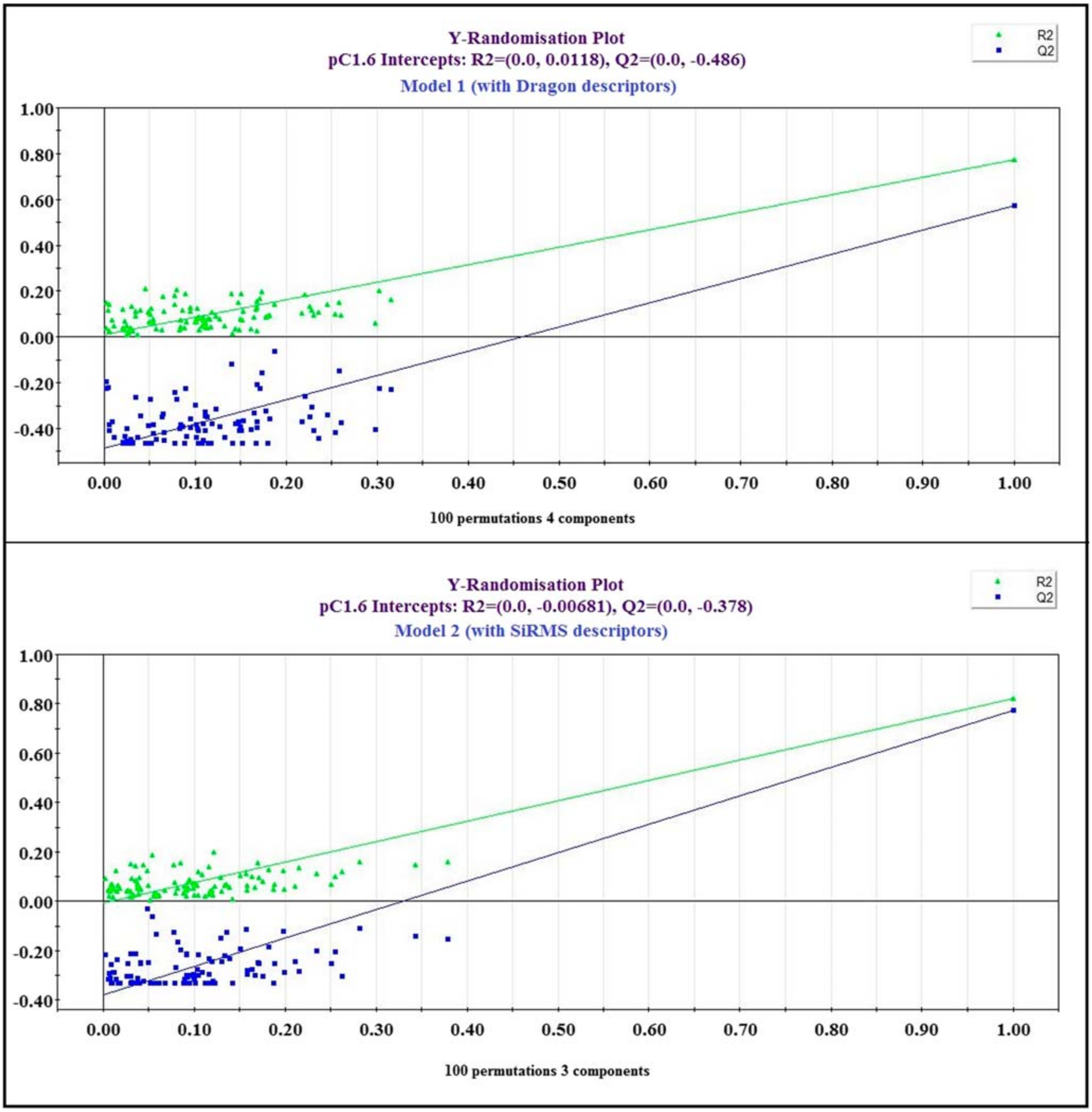


Fig. 10 *Y*-randomization plots for Model 1 and Model 2

Applicability domain assessment

The prediction reliability of both the 2D-QSAR models is determined by the applicability domain (AD) assessment. AD gives a theoretical region in chemical space defined by the respective model descriptors and responses in which the predictions are reliable [42]. AD assessment for both the models was performed using DModX (distance to model in the X-space) approach at 99% confidence level (**Figs. 8** and **9**). Both the models displayed good coverage of domain of applicability showing maximum number of compounds in the AD (only compound **6** is outside the AD in case of Model 1, i.e., 2D-QSAR model with Dragon descriptors). There were no outliers obtained from the test set for both the models. We have also performed AD assessment at 95% confidence level for both the models as given in the Supplementary Materials (Figures S1 and S2) and found that in this case three compounds in the test set were outside AD for the model with Dragon descriptors and two compounds in the test set for the model with SiRMS descriptors.

Table 1 External dataset and their predicted $pC_{1.6}$ values

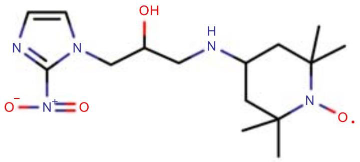
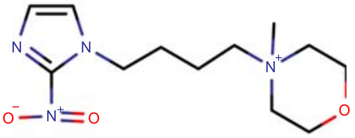
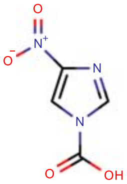
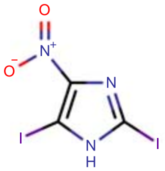
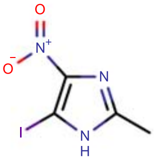
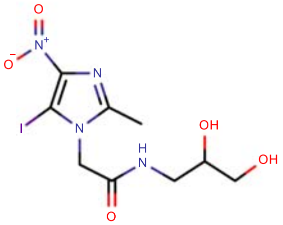
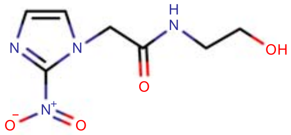
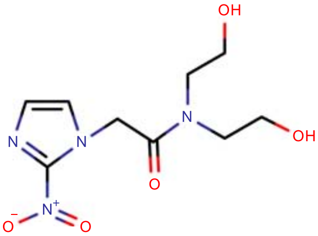
Compound Number	Structure	Observed $pC_{1.6}$	Predicted $pC_{1.6}$ using model 1	Predicted $pC_{1.6}$ using model 2	Reference
P-1		4.05	3.58	3.67	[26]
P-2		2.89	3.88	3.82	[26]
P-3		-	1.98	2.18	[44]
P-4		-	4.22	2.18	[44]
P-5		-	2.81	2.18	[44]
P-6		-	2.53	2.18	[44]
P-7		-	3.33	3.48	[45]
P-8		-	3.04	3.48	[45]

Table 2 Prediction quality [46] for the true external dataset

Compound number	Prediction status of model with Dragon descriptors			Prediction status of model with SiRMS descriptors		
	Composite score	Prediction quality	AD status (using standardization approach)	Composite score	Prediction quality	AD status (using standardization approach)
P-1	3	Good	Outside AD	3	Good	In
P-2	3	Good	In	3	Good	In
P-3	2	Moderate	In	3	Good	In
P-4	3	Good	In	3	Good	In
P-5	3	Good	In	3	Good	In
P-6	3	Good	Outside AD	3	Good	In
P-7	3	Good	In	3	Good	In
P-8	3	Good	In	3	Good	In

Y-randomization

Y-randomization plot analysis helps to understand the statistical significance of the model. The randomization plot confirms that the model is not the result of any chance correlation [43]. In this process, a number of models are generated by shuffling different combinations of *X* or *Y* variables (here *Y* variable only) based on the fit of the reordered model. In our work, we have used 100 permutations for random model generation. A model with no chance correlation would show very poor statistics for the randomized models, i.e., R_Y^2 intercept should not exceed 0.3 and Q_Y^2 intercept should not exceed 0.05 [43]. The randomization plots given in Fig. S8 show that the developed models are non-random and robust (as understood from their R_Y^2 and Q_Y^2 values) and are suitable for prediction of the radiosensitization effectiveness within the AD of the model (Fig. 10).

True external predictions

Prediction of responses for external compounds based on their molecular features using chemometric methods can reduce the experiment costs and animal handling. To verify the predictive power of both the models, we have used a set of eight nitroimidazole derivatives (Table 1) as an external prediction set [26, 44, 45]. The original dataset in the source literature contain 86 nitroimidazoles but we have removed two of them and used the rest 84 for modeling. These two compounds are now used for prediction purpose. In addition to this, the domain of applicability and their predictive reliability are analyzed using ***Prediction Reliability Indicator*** tool [46]. The prediction quality and domain of applicability are given in Table 2. From the prediction status, it can be inferred that model with fragment-based SiRMS descriptors provides better prediction than model with dragon descriptors.

Comparison with the previously published research

In the previously published research by Long and Liu (2010) [26], the authors developed MLR and projection pursuit regression (PPR) [47–49] models using complex descriptors such as geometrical, electrostatic, and quantum chemical descriptors. The models developed by us cannot be critically compared to the previously published since the calibration and validation set compositions are different. However, it can be found that our MLR model developed using SiRMS descriptor is better in terms of both training and test set validation metrics if we consider their MLR model (Table 3). Also the current model comes with an added advantage of presence of lower number of simple descriptors and non-requirement of conformation analysis or energy minimization prior to their calculation. Furthermore, the PPR based model reported in the

Table 3 Comparison of the current SiRMS model with previously developed MLR model

Model	Total no. of compounds used	No. of compounds in the training set	No. of compounds in the test set	Descriptor type	No. of descriptors in final model	Training set			Test set	
						R^2	Q^2	RMSEC	Q^2_{F1}	RMSEP
Current study	84	63	21	2D (fragment-based SiRMS)	5 (3 LVs)	0.82	0.79	0.26	0.80	0.28
Long and Liu, 2010	86	68	18	3D	6	0.80	0.76	0.28	0.76	0.28

previous study is derived from a more complicated process which uses projection based approach to convert high dimensional data to lower dimension. Moreover, 3D descriptors were used in the previous work. MLR or PLS models are more straight-forward and reproducible as used in the current work. In addition, 2D descriptors used in the present work are easy to compute and do not need any conformation analysis or energy minimization process.

Conclusion

This study targets for the development of fragment-based 2D-QSAR models for predicting radiosensitization of nitroimidazole derivatives. The simplex descriptors give an insight about the fragments and their proper position in the nitroimidazole ring that enhance or decline the radiosensitization effectiveness. Also reduction in the large data pool by using multilayered variable selection is shown for better handling of a large pool of descriptors and removing chances of intercorrelation among them. Further, the newly developed models were used for prediction of eight external compounds and their prediction reliability was checked.

Funding information PD thanks Indian Council of Medical Research, New Delhi, for awarding with a Senior Research Fellowship. KR thanks BRNS, Department of Atomic Energy, Govt. of India, for a Major Research Project (36(3)/14/08/2017-BRNS).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Kvols LK (2005) Radiation sensitizers: a selective review of molecules targeting DNA and non-DNA targets. J Nucl Med 46:187S
2. Bonnet M, Hong CR, Gu Y, Anderson RF, Wilson WR, Pruijn FB, Wang J, Hicks KO, Hay MP (2014) Novel nitroimidazolealkylsulfonamides as hypoxic cell radiosensitisers. Bioorg Med Chem 22:2123–2132
3. Cairns RA, Harris IS, Mak TW (2011) Regulation of cancer cell metabolism. Nat Rev Cancer 11:85
4. Chang Q, Jurisica I, Do T, Hedley DW (2011) Hypoxia predicts aggressive growth and spontaneous metastasis formation from orthotopically grown primary xenografts of human pancreatic cancer. Cancer Res 71:3110–3120
5. Rohwer N, Cramer T (2011) Hypoxia-mediated drug resistance: novel insights on the functional interaction of HIFs and cell death pathways. Drug Resist Updat 14:191–201
6. Wilson WR, Hay MP (2011) Targeting hypoxia in cancer therapy. Nat Rev Cancer 11:393
7. Astor M, Hall EJ, Martin J, Flynn M, Biaglow J, Parham JC (1982) Radiosensitizing and cytotoxic properties of ortho-substituted 4- and 5-nitroimidazoles: role of NPSH reactivity. Int J Radiat Oncol Biol Phys 8:409–413
8. Koch CJ, Skov KA (1994) Enhanced radiation-sensitivity by preincubation with nitroimidazoles: effect of glutathione depletion. Int J Radiat Oncol Biol Phys 29:345–349
9. Roy K (2018) Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory. J Indian Chem Soc 95:1497–1502
10. Hansch C, Leo A, Hoekman DH (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. American Chemical Society Washington, DC
11. Hansch C, Leo A, Mekapati SB, Kurup A (2004) Qsar and Adme. Bioorg Med Chem 12:3391–3400
12. Klein C, Kaiser D, Kopp S, Chiba P, Ecker GF (2002) Similarity based SAR (SIBAR) as tool for early ADME profiling. J Comput Aided Mol Des 16:785–793
13. Merlot C (2010) Computational toxicology—a tool for early safety evaluation. Drug Discov Today 15:16–22
14. Xu X, Zhang W, Huang C, Li Y, Yu H, Wang Y, Duan J, Ling Y (2012) A novel chemometric method for the prediction of human oral bioavailability. Int J Mol Sci 13:6964–6982
15. Yoshida F, Topliss JG (2000) QSAR model for drug human oral bioavailability. J Med Chem 43:2575–2585
16. Roy H, Nandi S (2019) In silico modeling in drug metabolism and interaction: current strategies of lead discovery. Bentham Science Publishers, Sharjah
17. Simeon S, Montanari D, Gleeson MP (2019) Investigation of factors affecting the performance of in silico volume distribution QSAR models for human, rat, mouse, dog & monkey. Mol Inform 38:1900059
18. Halder AK, Cordeiro M (2019) Development of multi-target chemometric models for the inhibition of class I PI3K enzyme isoforms: a case study using QSAR-Co tool. Int J Mol Sci 20:4191
19. Dmitriev AV, Lagunin AA, Karasev DЦ, Rudik AV, Pogodin PV, Filimonov DA, Poroikov VV (2019) Prediction of drug-drug interactions related to inhibition or induction of drug-metabolizing enzymes. Curr Top Med Chem 19:319–336
20. Salahinejad M (2015) Quantitative structure property relationships on formation constants of radiometals for radiopharmaceuticals applications. J Radioanal Nucl Chem 303:671–680
21. Singh S, Ojha H, Tiwari AK, Kumar N, Singh B, Mishra AK (2010) Design, synthesis, and in vitro antiproliferative activity of benzimidazole analogues for radiopharmaceutical efficacy. Cancer Biother Radiopharm 25:245–250
22. Yoshizuka K, Pietzsch H-J, Seifert S, Stephan H (2013) Quantitative structure property relationship of logP for radiopharmaceutical technetium and rhenium complexes by using molecular dynamics calculations. Solvent Extr Res Dev, Jpn 20:15–27
23. Santos L, Pilar Cornago M, Izquierdo MC, Consuelo Lopez-Zumel M, Smeyers YG (1989) Electron affinity/radiosensitizing activity relationship for quaternary 5-nitroimidazole derivatives. Quantum chemical QSAR. Quant Struct-Act Rel 8:214–217
24. Wardman P, Clarke ED (1987) Redox properties and rate constants in free-radical mediated damage. Br J Cancer Suppl 8:172
25. De P, Bhattacharyya D, Roy K (2019) Application of multilayered strategy for variable selection in QSAR modeling of PET and SPECT imaging agents as diagnostic agents for Alzheimer's disease. Struct Chem 30:2429–2445
26. Long W, Liu P (2010) Quantitative structure activity relationship modeling for predicting radiosensitization effectiveness of nitroimidazole compounds. J Radiat Res 51:563–572
27. MarvinSketch software, https://www.chemaxon.com. Accessed 26 Aug 2019
28. Dragon version 7, Kodesrl, Milan, Italy, 2016; software available at http://www.talete.mi.it/index.htm. Accessed 26Aug 2019
29. Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) Hierarchic

system of QSAR models (1D–4D) on the base of simplex representation of molecular structure. J Mol Model 11:457–467

30. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253
31. Park H-S, Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341
32. Drug Theoretics and Cheminformatics (DTC) laboratory software tools https://dtclab.webs.com/software-tools Accessed 28 Aug 2019
33. Khan PM, Roy K (2018) Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). Expert Opin Drug Discov 13:1075–1089
34. Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, Cornwall, Great Britain
35. Roy K, Ambure P (2016) The "double cross-validation" software tool for MLR QSAR model development. Chemom Intell Lab Syst 159:108–126
36. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130
37. U. Simca-P, 10.0, info@umetrics.com, www.umetrics.com, Umea, Sweden, 2002. Accessed 30 Aug 2019
38. Roy K, Mitra I (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Comb Chem High Throughput Screen 14: 450–474
39. Ojha PK, Mitra I, Das RN, Roy K (2011) Further exploring rm^2 metrics for validation of QSPR models. Chemom Intell Lab Syst 107:194–205
40. Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom Intell Lab Syst 152:18–33
41. Akarachantachote N, Chadcham S, Saithanu K (2014) Cutoff threshold of variable importance in projection for variable selection. Int J Pure Appl Math 94:307–322
42. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. IJQSPR 1:45–63
43. Rücker C, Rücker G, Meringer M (2007) Y-randomization and its variants in QSPR/QSAR. J Chem Inf Model 47:2345–2357
44. Krause W, Jordan A, Scholz R, Jimenez J-LM (2005) Iodinated nitroimidazoles as radiosensitizers. Anticancer Res 25:2145–2151
45. Brown JM, Ning YY, Brown DM, Lee WW (1981) SR-2508: a 2-nitroimidazole amide which should be superior to misonidazole as a radiosensitizer for clinical use. Int J Radiat Oncol Biol Phys 7:695–703
46. Roy K, Ambure P, Kar S (2018) How precise are our quantitative structure–activity relationship derived predictions for new query chemicals? ACS Omega 3:11392–11406
47. Friedman JH, Stuetzle W (1981) Projection pursuit regression. J Am Stat Assoc 76:817–823
48. Du Y, Liang Y, Yun D (2002) Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. J Chem Inf Comput Sci 42:1283–1292
49. Liu H, Yao X, Liu M, Hu Z, Fan B (2007) Prediction of gas-phase reduced ion mobility constants (K0) based on the multiple linear regression and projection pursuit regression. Talanta 71:258–263