Toxicology in Vitro 65 (2020) 104768

Contents lists available at ScienceDirect

Toxicology in Vitro

journal homepage: www.elsevier.com/locate/toxinvit

Application of QSARs in identification of mutagenicity mechanisms of nitro and amino aromatic compounds against *Salmonella typhimurium* species

Gopala Krishna Jillella[a], Kabiruddin Khan[b], Kunal Roy[b,*]

[a] Department of Pharmacoinformatics, National Institute of Pharmaceutical Educational and Research (NIPER), Chunilal Bhawan, 168, Manikata Main Road, 700054 Kolkata, India

[b] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, 188 Raja S C Mullick Road, 700032 Kolkata, India

ARTICLE INFO

Keywords:
OECD
Mutagenicity
QSAR
Salmonella typhimurium
Validation

ABSTRACT

In an attempt to describe the underlying causes of mutagenicity mainly due to organic chemicals, quantitative structure-activity relationship (QSAR) models have been developed using two different *Salmonella typhimurium* mutagenicity endpoints with or without presence of liver metabolic microsomal enzymes (S9) namely TA98-S9 and TA98 + S9. The models were developed using simple 2D variables having definite physicochemical meaning calculated from Dragon, SiRMS, and PaDEL-descriptor software tools. Stepwise regression followed by partial least squares (PLS) regression was used in model development following the strict OECD guidelines for QSAR model development and validation. The models were validated using coefficient of determination R^2, cross-validation coefficient Q^2_{LOO} (leave one out) while the test set predictions were analyzed using Q^2F_1 (coefficient of determination for the test set). Several other internationally accepted validation metrics like $MAE_{95\%train}$, average $r_{m(LOO)}{}^2$ and $\Delta r_{m(LOO)}{}^2$ (for the training set) were used to check model robustness while predictive efficiency was evaluated using $MAE_{95\%test}$, average $r_m{}^2$ and $\Delta r_m{}^2$ (for the test set). The scope of predictions was defined by applicability domain analysis using the DModX approach, a recommended tool for PLS models. The major contributing features related to mutagenicity include lipophilicity, electronegativity, branching and unsaturation, etc. The present manuscript is the first attempt to undertake modeling of two different endpoints (TA98-S9 and TA98 + S9) in order to explore major contributing molecular features linked directly or indirectly to mutagenicity.

1. Introduction

Organic chemicals (OCs) constitute a large collection of chemicals employed in several spheres of life including dyes, polymers, pesticides, textiles, explosives and several food substances, etc. With the gradual rise of chemical consumption in diaspora, there arises a need to check for adversity that follows. The exponential rise in the consumption of OCs has compelled scientists to explore the probable toxicity caused to the flora and fauna at the genetic level starting from the building block of life, i.e., deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) (Moretti et al., 2002; Tsuboy et al., 2007). OCs have long been identified as carcinogenic or mutagenic materials in several ecologically important species including humans along with various bacterial cell lines such as *Salmonella typhimurium* (Chung and Cerniglia, 1992; Chung et al., 1997; Rashid et al., 1987; Sabbioni, 1994; Tomita et al., 1982). Mutagens are distinct from the carcinogens where the former leads to one or the other types of cancer due to mutation while carcinogens need not cause mutation in genetic material to develop cancer (Maisanaba et al., 2015). Among the various functionalities present in the OCs, nitroaromatics and aromatic amines have proven to have greater mutagenic potential compared to other classes of chemicals.

For many decades, scientists have relied upon the Ames test (Stead et al., 1981) for the mutagenicity study. It is estimated from the Ames test that the TA98 strain consists of a full complement of nitroreductases required to activate the reduction reaction of nitro aromatics, whereas aromatic and hetero aromatic amines demand the presence of an exogenous metabolic activation system, i.e. S9, to initiate the oxidation reaction, suggesting that mutagenicity can only take place at the nitro moiety in nitroaromatic amines in *S. typhimurium* strain TA98 without S9 mix (Fu, 1990; Fu and Herreno-Saenz, 1999). Table 1 and Fig. 1 depict the experimentally proved mechanisms leading to mutagenicity.

The major drawbacks of using the Ames test in mutagenicity

* Corresponding author.
E-mail address: kunal.roy@jadavpuruniversity.in (K. Roy).

https://doi.org/10.1016/j.tiv.2020.104768
Received 17 November 2019; Received in revised form 19 December 2019; Accepted 6 January 2020
Available online 08 January 2020
0887-2333/

Table 1
Some of the reported mechanisms leading to mutagenicity.

Sr. no.	Modes of mutagenicity	Mechanism
1	Alkylation	Most of the Alkylating agents damage the DNA with the formation of N2-alkylG (where G stands for guanine) and other lesions; for example formaldehyde reacts with the exocyclic amino group of deoxyguanosine to produce N2-methylG (Yasui et al., 2001). Some other examples of alkylating agents include ethanol which enzymatically get oxidized to acetaldehyde and thus forms N2-ethylG found in liver DNA and urine of alcoholic patients (Cheng et al., 2008). Other alkylating agents having sufficient mutagenic potential include polycyclic aromatic hydrocarbons (PAH-DNA adducts), nitrosamines (formation of the O6-methylG (O6-MeG)) and bis-electrophilic agents like 1, 3-Butadiene (BD can be oxidized to 1, 2, 3, 4-diepoxybutane (DEB), a prominent bis-electrophilic carcinogenic metabolite).
2	Oxidation	Oxidizing agents can produce 7, 8-dihydro-8-oxo-2′-oxodeoxyguanosine (8-oxodG) lesions. 8-oxodG is a ubiquitous lesion arising from the oxidation of the C8 atom of G to form a hydroxyl group by free radical intermediates of oxygen that are produced by chemical oxidation, ionizing radiation, or UV irradiation (Degan et al., 1991; Fraga et al., 1990). The enol (a lactim) at the C8-N7 position of G is converted to the more stable 8-oxodG lactam form.
3	Amination	Several aminating agents like aryl amines and *N*-acetyl aryl amines possess higher propensity to act like potential mutagenic substances. This group is extensively studied for their mutagenic activity and also implemented in several in silico studies due to their presence in various occupational settings like tobacco smoke, chemical dyes etc. These chemicals go on to form adducts like 2-aminofluorene (AF-dG) and *N*-acetyl-2-aminofluorene (AAF-dG) through amination of the C8 atom of guanine (via an initial N7 reaction, linking the amine group of the aryl amine) (Vrtis et al., 2013).
4	Co-ordination	Heavy metal ions also produce mutagenicity with the formation of DNA-DNA intra-strand and inter-strand cross-links via coordination bonds. For example, chromium (VI) complex permeates the cell membranes and gets reduced to form chromium (III) complex, following which it then coordinates with oxygen atoms of phosphate backbone of two adjacent nucleotides within one DNA strand or between two DNA strands, yielding chromium (III)-DNA intra-strand thus yielding an inter-strand cross-links (O'Brien et al., 2002).
5	Photo-addition	The ultra violet (UV) radiation leads to the formation of photoproducts (e.g., CPD) by cycloaddition of the C5-C6 double bonds with adjacent pyrimidine bases; thus it behaves like a non-chemical mutagenic agent. Six diastereomers are generated, depending on the position of pyrimidine moieties with respect to the cyclobutane ring (cis/trans stereochemistry) and on the relative orientation of the two C5-C6 bonds (syn/anti regiochemistry) (Cadet et al., 1985). The cis-syn form is formed preferentially to the trans-syn diastereomers within double-stranded DNA. The trans-anti and trans-syn photoproducts are only present within single-strand or denatured DNA (Ravanat et al., 2001).
6	Hydrolysis	The final proposed mechanism for mutagenesis is by hydrolysis, where AP (apurinic/apyrimidinic) sites are generated by spontaneous reactions, chemical induction or by enzyme-catalyzed hydrolysis of the N-glycosyl bond (Wilson Iii and Barsky, 2001) resulting in the loss of genetic information. In mammalian cells, it has been estimated that approximately 12,000 purines are lost spontaneously per genome per cell generation (20*h*) in the absence of any protective effects of chromatin packaging. It was subsequently shown that depyrimidination occurs at a rate about 100 times more slowly than depurination (Wilson Iii and Barsky, 2001). Damaging chemicals, e.g., free radicals and alkylating agents, promote base release, mostly by generating base structures that destabilize the N-glycosyl linkage due to positively-charged leaving groups (Wilson Iii and Barsky, 2001).

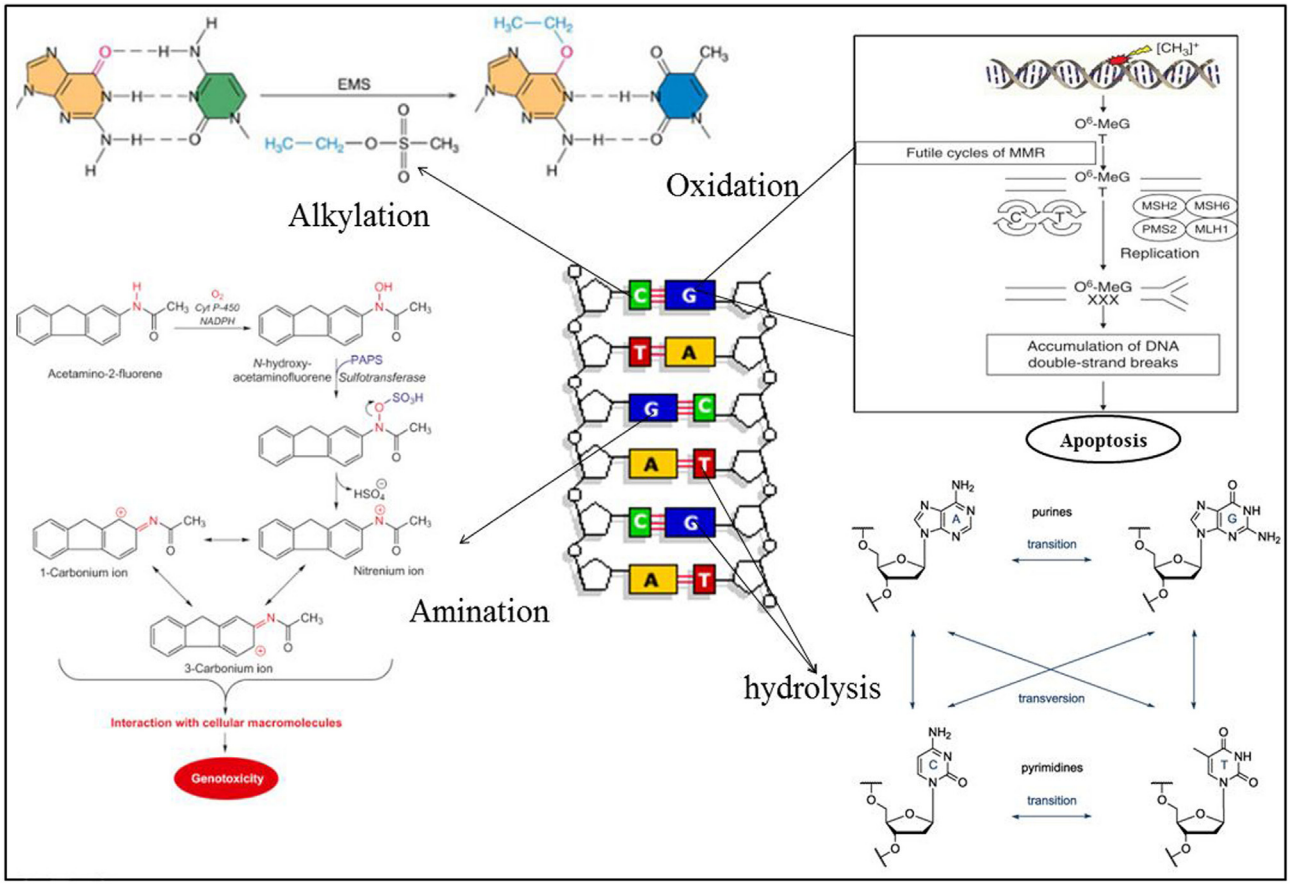

Fig. 1. Known mechanisms of mutagenicity proposed by several experts (Cadet et al., 1985; Cheng et al., 2008; Degan et al., 1991; Fraga et al., 1990; O'Brien et al., 2002; Ravanat et al., 2001; Vrtis et al., 2013; Wilson Iii and Barsky, 2001; Yasui et al., 2001).

estimation lies in its associated time, cost and man power. Finally, a number of existing organic chemicals (more than several millions) make it impossible to undertake experimental evaluation of all chemicals through the Ames test against enormous number of endpoints. Several computational (in silico) techniques like QSAR (quantitative structure-activity relationship) and pharmacophore modeling can help to fill the data gap. QSAR offers several advantages over other techniques as it utilizes limited experimental resources, cost and offers time efficient outcome (Dearden, 2016). Due to these encouraging features, QSAR is recommended for use in early detection of various toxic chemicals (Khan et al., 2019b; Khan et al., 2019c).

A number of scientific groups have attempted to identify the mutagenic features present in the various subclasses of organic compounds, some of them are discussed here. Garg et al. developed a QSAR model correlating the experimental mutagenicity of 43 aminoazobenzenes using several molecular descriptors calculated from quantum-chemical semi empirical approach (Garg et al., 2002). Gramatica and colleagues (Gramatica et al., 2007) used a dataset of 48 nitro-PAHs (polycyclic aromatic hydrocarbons) mutagenicity data against TA100 (*S. typhimurium*) strain to analyze the possible toxicophores using QSAR study, while Wang et al. (2005) performed comparative molecular field analysis (CoMFA) and molecular orbital theory based classic structure-activity relationship analysis in order to explore the structural fragments responsible for mutagenicity using 219 nitro aromatics compounds (Wang et al., 2005).

The current manuscript is the first report giving in silico QSAR models correlating the mutagenicity of nitro aromatics and aromatic/hetero aromatic amines against *Salmonella typhimurium* species (TA98) with or without presence of liver metabolic microsomal enzymes (S9). The previous reports on this topic were solely consisting of either TA98 + S9 or TA98-S9 toxicity endpoints. Only 2D descriptors with definite physicochemical meaning were employed here in model development in order to derive models of relatively less complexity from the interpretation perspective. The models were validated using some of the very stringent validation metrics. The applicability domain study was checked in order to give models a definite zone for reliable prediction for unknown or untested chemicals.

2. Methods and materials

2.1. Dataset

A reliable QSAR model can only be obtained from trusted sources of experimental data. To achieve this, the authors have compiled two sets of mutagenicity data against *Salmonella typhimurium* (TA98) bacterial species with (TA98 + S9) or without (TA98-S9) implementing microsomal activating enzyme named S9, solely collected from literatures (Bhat et al., 2005; Ding et al., 2017; Leong et al., 2010). For the ease of acceptability, the collected data were selectively filtered for uniform experimental procedures, conditions and protocol. The mutagenic endpoint TA98-S9 consists of 295 nitro aromatic compounds along with their derivatives, while the TA98 + S9 endpoint data was derived of 309 chemicals with reported acute mutagenicity against *S. typhimurium* and composed of aromatic amines and hetero aromatic amines along with their derivatives (See Sheet 2-3 in SI-1). The mutagenicity in all cases was expressed as the logarithm of the number of revertants per nmol and was used as such. Out of 295 TA98-S9 endpoints, four compounds (**191**, **192** and **197**, **198**) were identified as stereoisomers having contrast mutagenicity values, thus excluded from the initial analysis. The dataset chemicals in both the endpoints are commercially employed to make several important products or product mixtures such as dyes, personal care products, organic reactants, product intermediate, resins and also for research purposes, thus making them indispensible from day to day life and necessitating their evaluation for mutagenic potential. All the structures were manually drawn in MarvinSketch version 4.0 (available on https://chemaxon.com/products/marvin) and cross verified from the source paper in order to avoid any miscalculation of molecular descriptors at the later stage. Finally, the structures were cleaned in 2D, aromatic bonds and explicit hydrogens added and saved as MDL.mol format, a recommended input format for Dragon (Mauri et al., 2006), SiRMS (Kuz'min et al., 2005) and PaDEL-descriptor (Yap, 2011) softwares.

2.2. Descriptor calculations

For the ease of interpretation, simple 2D descriptors with definite physicochemical meaning were used during the QSAR modeling. A total of 43 ETA indices (extended topochemical indices) (Khan and Roy, 2019) were calculated from PaDEL-descriptor (version 2.21) (Yap, 2011) software while from Dragon (version 7.0) (Mauri et al., 2006) eight different classes of descriptors were generated including constitutional indices, *E*-state indices, 2D atom pairs, molecular property descriptors, connectivity indices, functional group counts, ring indices and atom-centered fragments giving a total of 467 different variables (Mauri et al., 2006). Additionally, nearly 15,000 molecular fragments as 2D descriptors were calculated from simplex representation of molecular structure (SiRMS) software (Kuz'min et al., 2005). The descriptors with correlation more than > 0.9 (R^2) were excluded from the analysis in order to avoid problems of over fitting (Khan et al., 2019a). The endpoints are separately modeled with SiRMS descriptors in order to identify the most contributing features present in the respective datasets.

2.3. QSAR modeling and validation

The initial datasets for the both endpoints were split into training and validation set using defined algorithms. Various data division techniques (Sorted response, Euclidean distance and Kennard-Stone method) partitioned approximately 75% molecules into the training set while remaining 25% were placed in the test set. However, the best division which gave the most reliable models was obtained by the Euclidean distance based partitioning (Golmohammadi et al., 2012) using a software tool (available at http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab/). The training set was solely employed for feature selection and model development while the test set was used to analyze predictivity of the developed models. For feature selection, stepwise regression with stepping criteria also known as “Fisher criteria” was used with specified threshold of $F = 4.0$ for the inclusion and $F = 3.9$ for exclusion (Hossain and Roy, 2018). The process was repeated several times (after removing selected descriptors in the previous runs) in order to identify the significant descriptors. Finally, a set of best 25 variables were collected at the end of stepwise analysis and were subjected to best subset selection. The best two models for both the endpoints were then subjected for partial least squares (PLS) analysis in order to reduce noise form the models, as PLS regression calculates latent variables (LVs) from the original variables and can also handle a lot of noisy descriptors. As a part of the PLS model development, the software internally performs input data scaling (standardization) followed by computation of latent variables scores (the actual regressing variables), although the final regression coefficients are presented in terms of the original un-scaled variables (similar to multiple linear regression or MLR equations). Unlike MLR models, determination of standard errors of regression coefficients for PLS models is not straight-forward; however, the relative importance of different variables can be presented in terms variable importance plot (VIP). For validation, various quantitative validation metrics were used for defining the quality of the developed models which are evaluated in terms of stability, robustness, fitness and predictivity. The coefficient of determination (R^2), internal predictivity metrics like leave-one-out cross-validated R^2 or Q^2_{LOO} and external predictivity metrics like R^2_{pred} or Q^2_{ext} and Q^2F_2 were calculated (Roy and Mitra, 2011). We have also checked the MAE (mean absolute error) based criteria for both internal and external validation sets (Roy

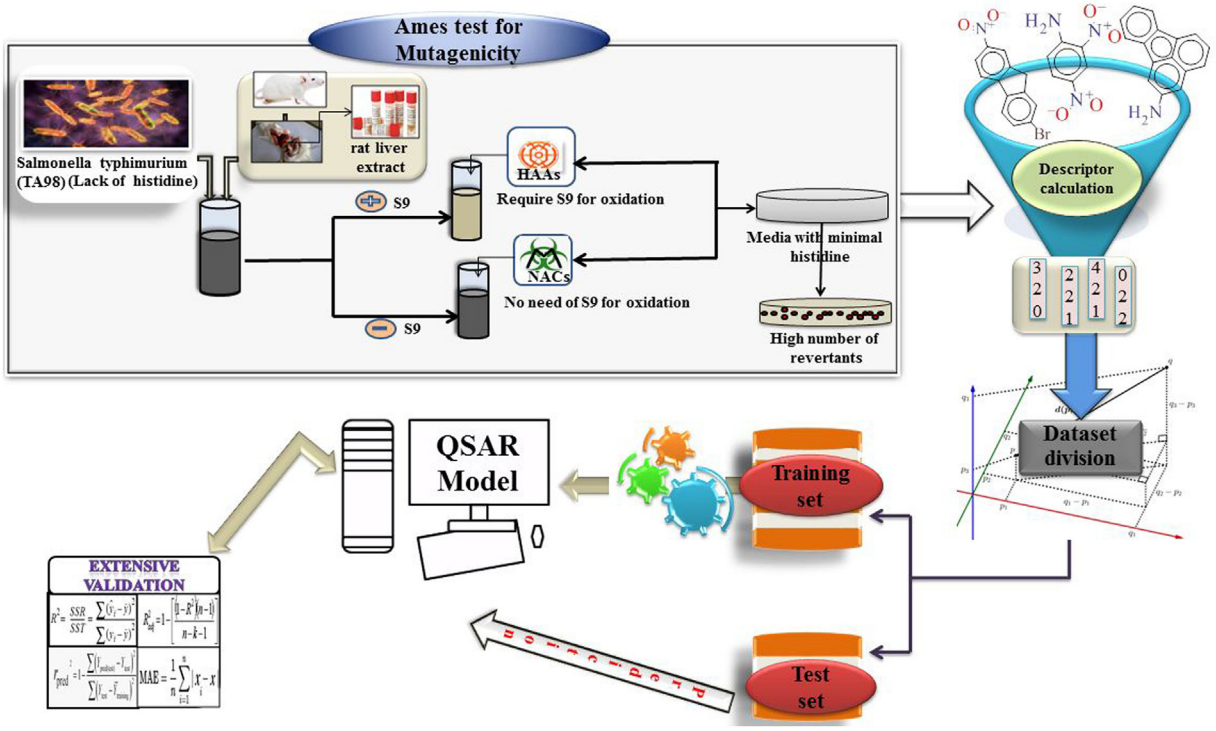

Fig. 2. The detailed methodology followed in the present work.

et al., 2016). Additionally, highly stringent rm^2 parameters for validations were also calculated to check for model robustness (Roy et al., 2012). The prediction error of the developed QSTR model was evaluated by the statistical parameters of the model for the external set which should be within the response and chemical domains of the internal or training set. Model randomization was performed to substantiate the robustness of developed PLS model. Additionally, a variable importance plot was constructed for each model depicting individual contributions of the modeled descriptors. The performed methodology is summarized in Fig. 2.

2.4. Applicability domain (AD)

The applicability domain is a concept of chemical space explained by the model descriptors and the modeled response. In the present study, we have implemented DModX (distance to model in X-space) approach using SIMCA-P software (Umetrics, 2013) to scrutinize the applicability domain of the developed QSTR model. The DModX approach was implemented at 99% confidence level to examine whether the test compounds used in modeling study lie in the chemical space or outside the chemical space of the training compounds used for developing the model (Wold et al., 2001).

2.5. Modeling mutagenicity using molecular fragments

Lastly, QSAR modeling with fragmental variables was performed in order to explore major contributing features embedded in the structures of organic molecules. To achieve this, a series of 2D molecular fragments were computed using simplex representations of molecular structure (SiRMS) software (Version 4.1.2.270). The molecular fragments were then analyzed for their probable cause leading to mutagenicity in *S. typhimurium* as per the recommended OECD guidelines for model interpretations. The remaining procedures like data division, features selection and model validation were performed as per the protocol specified above.

2.6. Software used

Marvin sketch (version 14.10.27) software (http://www.chemaxon.com/) was used to draw the chemical structures. SiRMS (Version 4.1.2.270), Dragon version 7 (http://www.talete.mi.it/products/dragondescription.htm) and PaDEL-Descriptor (http://www.yapcwsoft.com/dd/padeldescriptor) software tools were used to calculate the molecular descriptors. For data division, freely available DatasetDivisionGUIv1.2_9May2017 tool at http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab was used. The stepwise regression analysis was performed by using MINITAB Software (version 14.13) (http://www.minitab.com/en-US/default.aspx). Best subset selection was performed using freely available QSAR tools at http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab. Simca-p (Version 10.0) was used for the generation of various PLS plots.

3. Result and discussion

The current work reports QSAR modeling of mutagenicity potential of selected organic chemicals mainly against *Salmonella typhimurium* species (TA98) employing two different strategies. In the first approach, the data comprise mutagenicity of 291 nitroaromatic compounds measured without using microsomal activation (Rat S9) while the second dataset consists of 309 aromatic and hetero-aromatic amines having mutagenicity data with the employment of microsomal activation. In order to get reliable models, a great care was taken in data collection, curation, and validation of QSAR models in order to obey the strict OECD guidelines for QSAR model development as described in the methods and materials section (Khan et al., 2018). The developed models were rigorously validated using various internal and external validation parameters in order to prove their robustness. Additionally, the MAE based criteria for the test set were checked to enhance acceptance of the final models (Roy et al., 2016). The models developed with employing S9 passed the MAE criteria (Roy et al., 2016). The individual equations depicting their validation metrics are given below with descriptors elaborated thoroughly.

3.1. Modeling nitro-aromatic compounds against Salmonella typhimurium strain (TA98) without microsomal activation (S9)

From the total of 291 nitro-aromatic compounds, 219 molecules were utilized for the model development while the remaining 72 molecules were used to validate the final models. Initially, a number of QSAR models were generated using the best subset selection; however the best two models were finally selected from the large pool of generated models based on the most interpretable chemical features or toxicophores. Both the models gave a moderate level of robustness with coefficient of determination showing values ranging from 0.71–0.73 (0.68–0.71 for leave-one-out cross validation), while the predictivity of the test set was measured using predictive R^2 covering 74–76% variance of the test set molecules. Eqs. (1)–(2) detail the validation metrics obtained for both the models along with regression coefficients of individual variables arranged in a descending order of contribution to the mutagenicity as per the VIP plots. The VIP values rank model descriptors based on their significance to the response and are calculated using classical regression coefficient, weight vector and t-statistic, etc. (Akarachantachote et al., 2014). A VIP score of > 1 corresponds to significant descriptors whereas a score < 1 is considered as insignificant. Model 1 consists of three highly significant variables namely RDCHI (reciprocal distance sum Randic-like index), ETA_dBeta (a measure of relative unsaturation content in a molecule) and nCIR (number of circuits or loops) with VIP scores of 1.45, 1.4 and 1.2 respectively. Model 2 for TA98-S9 endpoint contained relatively more significant variables such as ETA_Beta_ns (a measure of electron-richness of the molecule), X5A (average connectivity index of order 5), ETA_dBeta, RCI (ring complexity index or size) and RDCHI with VIP scores of 1.25, 1.25, 1.10, 1.10 and 1.05 respectively. The remaining variables in both the cases were less significant. The loading plot in both models hinted towards a greater positive contribution from ETA indices and RDCHI due to their closeness to the dependent variable, i.e., mutagenicity. Finally, Y-randomization implemented in SIMCA-P was performed at 100 permutations to check for the non-randomness of the obtained models. Interestingly both the QSAR models on TA98-S9 gave intercepts for determinant coefficients of 0.003 and 0.004 and those for cross validated determinant coefficients of −0.28 and −0.17, which are far below then expected cut off; for a good model, the intercept for R^2 should be < 0.4 and that for Q^2 should be < 0.05. Additionally a stringent correlation (r) cut off of 0.9 was implemented at every stage of modeling to avoid problems of overfitting. The obtained modeled descriptors with highest influence on mutagenicity are grouped in four sub classes such as descriptors depicting unsaturation in molecules like RDCHI, ETA_dBeta and ETA_Beta_ns, descriptors designating presence of various rings like NRS, nCIR, and RCI, descriptors with more hydrophobic influence which include X5Av, X5A, F07[C−C], and finally, descriptors with more electronegative element content such as F09[N−N], B08[Cl−Cl] and B09[N−N]. Various qualitative plots generated by SIMCA tool are given in Figs. S1-S2 in Supplementary Materials (SI-2).

3.1.1. Mechanistic interpretation of the TA98-S9 models

$$\begin{aligned} Log(TA98 - S9)rev/nmol &= -17.487 + 4.310 \times \text{RDCHI} + 0.326 \times \text{ETA\_dBeta} - 0.040 \times \text{nCIR} + 24.010 \times \text{X5Av} - 0.206 \times \text{F07[C - C]} - 1.892 \times \text{NRS} + 1.390 \times \text{F09[N - N]} - 2.220 \times \text{B08[Cl - Cl]} \end{aligned} \quad (1)$$

$n_{training} = 219$, LV = 6, $R^2 = 0.731$, $Q^2 = 0.710$, $r^2_{m(LOO)} = 0.603$, $\Delta r^2_{m(LOO)} = 0.196$, MAE = 0.852

$n_{test} = 72$, $R^2_{pred} = 0.756$, $Q^2_{F2} = 0.751$, $r^2_{m(test)} = 0.649$, $\Delta r^2_{m(test)} = 0.194$, MAE = 0.728

$$\begin{aligned} Log(TA98 - S9)rev/nmol &= -17.4872 - 0.11793 \times \text{ETA\_Beta\_ns} + 37.18461 \times \text{X5A} + 0.35196 \times \text{ETA\_dBeta} + 2.77124 \times \text{RCI} + 4.866 \times \text{RDCHI} - 0.25617 \times \text{F07[C - C]} + 1.48697 \times \text{B09[N - N]} - 1.06744 \times \text{NRS} \end{aligned} \quad (2)$$

$n_{training} = 219$, LV = 6, $R^2 = 0.707$, $Q^2 = 0.679$, $r^2_{m(LOO)} = 0.564$, $\Delta r^2_{m(LOO)} = 0.202$, MAE = 0.883

$n_{test} = 72$, $R^2_{pred} = 0.736$, $Q^2_{F2} = 0.727$, $r^2_{m(test)} = 0.628$, $\Delta r^2_{m(test)} = 0.145$, MAE = 0.781

Both the models for TA98-S9 consist of 8 descriptors as shown in above equations; however, 4 descriptors were common in both cases; i.e. RDCHI, ETA_dBeta, NRS and F07[C−C] thus making a total of 12 descriptors. The descriptor RDCHI is defined by analogy with Randic connectivity index (X1), where the simple vertex degrees are replaced by the row sums of the reciprocal distance matrix as depicted in Eq. (3).

$$RDCHI = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij}.\,(RDS_i.\,RDS_j)^{-1/2} \quad (3)$$

Here, A is the number of vertices and a_{ij} is equal to 1 only for pairs of adjacent vertices and zero; R is the topological radius, D is the topological diameter and 'S' means sum. The RDCHI descriptor values increases with the size of the molecules whereas it decreases with branching. The positive regression coefficient of this parameter suggests that the presence of this fragment in nitro-aromatics enhances mutagenic potency of a molecule. This effect can be seen in compounds **17** (presence of pyrene or polycyclic aromatic hydrocarbon rings and four nitro groups), **176** (consists of 6 aromatic rings (benzene) and 2 nitro groups) and **212** (presence of a nitro-phenol ring) where a higher magnitude of this descriptor produces more toxic, potentially more mutagenic nitro-aromatic compounds. In contrast, molecules with lower RDCHI values were relatively less toxic and less mutagenic as seen in compounds **43** (having only one naphthalene ring), **80** and **177** (a single benzene ring).

The second most influential descriptor enhancing mutagenicity value was ETA_dBeta which gives the relative unsaturation content (Δβ) in the studied compounds. The index of unsaturation can be calculated by the following Eq. (4):

$$\Delta\beta = \sum \beta_{ns} - \sum \beta_{s} \quad (4)$$

Here, Σβs is the summed β values for all the sigma bonds (VEM (Valence Electron Mobile environment) sigma contribution) and Σβns is the summed β values for all the non-sigma bonds including lone electron pairs capable of resonance if any (VEM non-sigma contribution) (Roy and Das, 2017). This descriptor contributes positively towards the mutagenicity as indicated by its positive regression coefficient. This indicates that the mutagenicity of the nitro-aromatic compounds increases with an increase in the numerical value of unsaturation in the form of double bonds ("=") as can be seen in compounds **99, 111** and **271,** while a low level of unsaturation in the form of double bonds will lead to less mutagenic chemicals mainly seen in case of **154, 182** and **228**. Presence of more unsaturation prevailed in most of the highly mutagenic nitro-aromatic compounds. Again, ETA_Beta_ns, which is a measure of electron richness in a molecule, is an index of non-sigma electrons including lone pair of electrons. The index is calculated from all the π bonds and electron lone pairs present on heteroatoms, carbonyls and the atoms which are capable of resonance like N, O, S and halogens present in aromatic compounds. In contrast to ETA_dBeta, this index exerted a negative effect towards mutagenicity possibly due to an

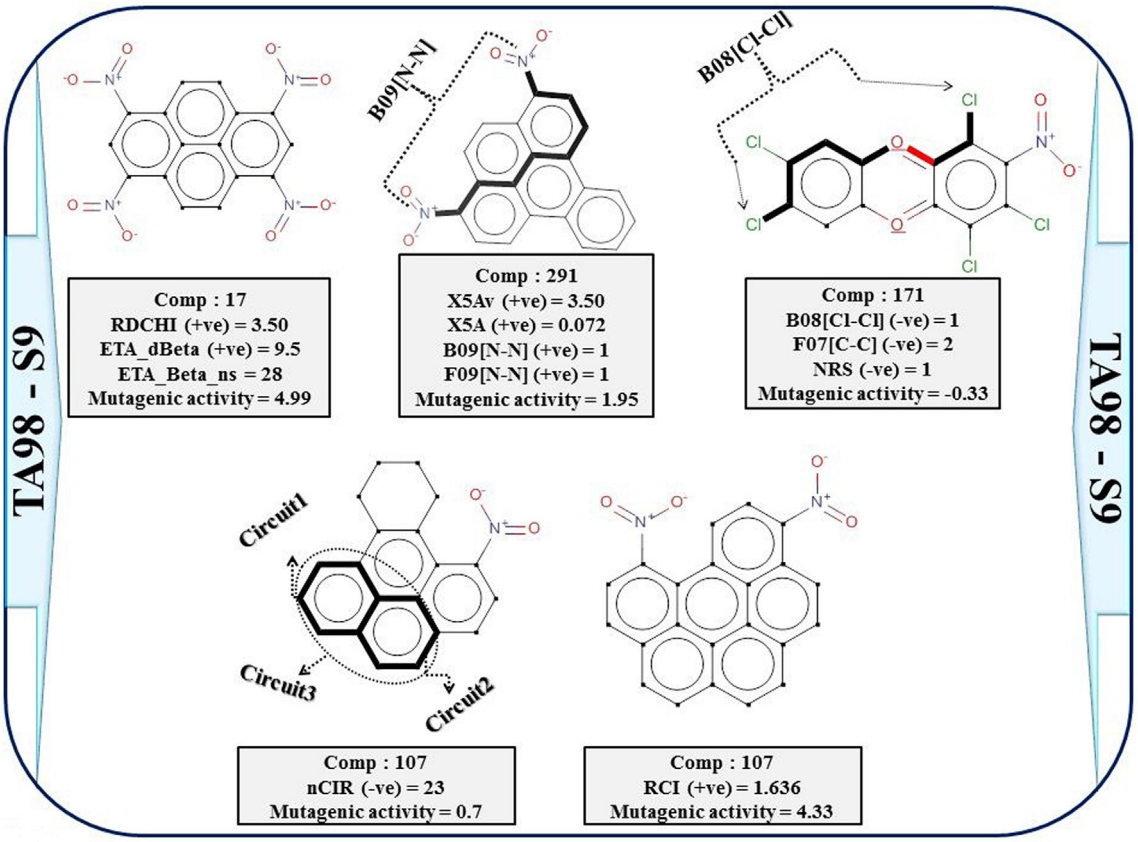

Fig. 3. Molecular features enhancing or reducing mutagenicity of TA98-S9.

increase in polar bulk in the molecule mainly due to sum effects of polar atoms such as N, O and S. Some of the low mutagenic chemicals with higher electron richness include compounds **4**, **141** and **171**.

The second major sub-class having a major influence on mutagenicity is the presence of various ring systems. The descriptor nCIR represents the count of the circuits (larger loop around two or more rings) in a molecule, the circuit being a self-returning path, i.e., a walk with no repeated vertices other than its first and last ones (Mauri et al., 2006). The negative regression coefficient of this descriptor indicates that an increase in the number of circuits (which are mostly due to presence of more fused aromatic rings such as pentacyclo or hexacyclo aromatics) reduces the potential of the mutagenicity within nitro-aromatics as evidenced from the compounds **88** and **284**. On the other hand, a lower value of this descriptor (with a decrease in the count of fused rings) enhances mutagenicity of nitro-aromatics as evident from compounds **68** and **264**. Another variable NRS representing the number of ring system, which gives the proportion of cyclic content when compared with the whole molecule, had a negative correlation with mutagenicity of bacterial species as seen in Eq. (5)

$$NRS = (nBO - B_R) - (nSK - A_R) + 1 \quad (5)$$

Here, nBO and nSK are the total numbers of bonds and atoms in the H-depleted molecular graph, respectively; B_R and A_R are the number of atoms and bonds belonging to rings, respectively. Thus, we can infer that with an increase in the ring proportion against the entire molecule, there will be a considerable decrease in the mutagenicity value as seen in the compounds **220, 224** and **233** and vice versa in compounds **54, 183** and **238**. The third variable RCI denoting ring complexity index gives the ratio of summed ring size of all the single cycles, over the total number of atoms in the ring systems, and it is calculated from Eq. (6).

$$C_R = \frac{R}{A_R} \quad (6)$$

Here, R and A_R are the total ring size and total number of atoms belonging to any ring system, respectively. Due to the positive regression coefficient, the high ring complexity causes higher mutagenicity in nitro-aromatics. Some examples with higher ring complexity values showing higher mutagenic characteristics include compounds **111, 176** and **179** (presence of many monocyclic rings), whereas opposite was observed in case of compounds **216**, **223** and **226.**

The presence of higher lipophilic bulk mainly due to carbon skeleton as represented by X5A and X5AV (average valence connectivity index of order 5) exhibited a positive influence in enhancing mutagenicity. The higher values of these descriptors correspond to an increase in size and non-polar surface area of the molecule. Thus, the mutagenicity of the compounds may increase with an increase in the surface area and size of the molecules as shown by the compounds **48**, **280** and **290** (presence of bromine and chlorine atoms as well as polycyclic hydrocarbon in the structure tends to increase the lipophilicity). On the other hand, the opposite may happen with the reduction of size, surface area and lipophilicity of the molecules as can be seen from compounds **16, 74** and **181** (with simple structures like nitrobenzene, flourobenzene having a small size and low molecular bulk). Another variable F07[C−C] contributing negatively towards mutagenicity was found to be less significant in controlling toxicity of nitro-aromatics owing to its lower VIP score.

The last group of variables denotes the presence of more electronegative elements in a molecule such as nitrogen and chlorine where the former contributes positively towards mutagenicity while the latter has a negative correlation coefficient. The descriptor F09[N−N] stands for frequency of two nitrogen atoms at the topological distance 9 contributing to the electronegativity in the nitro-aromatics, capable of undergoing alkylation (Rosenkranz and Klopman, 1995). As per Eq. (1), the descriptor F09[N−N] is positively correlated with the mutagenicity of nitro-aromatics. The presence of more electronegative atoms in the

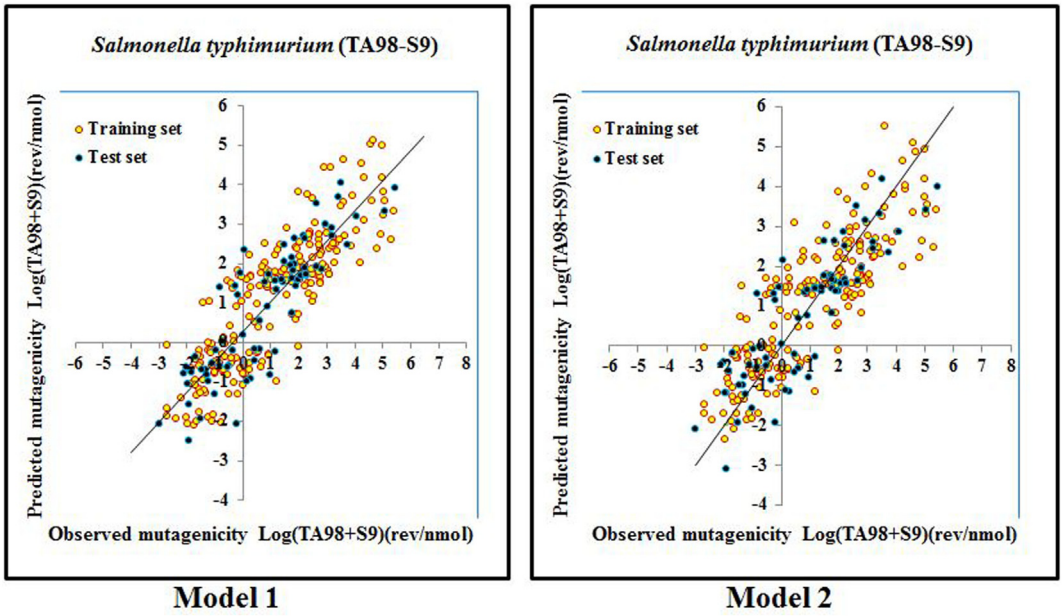

Fig. 4. Scatter plots of observed v/s predicted responses against model 1 and 2 of TA98-S9.

form of nitrogen increases the mutagenicity of nitro-aromatics as evident from several examples like **133**, **160** and **244** being more mutagenic than other molecules such as **138**, **142** and **143** which are devoid of this fragment. Another closely related variable B09[N—N] stands for the presence or absence of two nitrogen atoms at topological distance 9 also has a positive influence on mutagenicity of nitro-aromatics. The last variable appearing in the TA98-S9 model is B08[Cl—Cl] representing the presence or absence of two chlorine atoms located at the topological distance 8. Although this descriptor has a negative regression coefficient in the equation, the descriptor was found to be relatively insignificant as suggested by very low VIP score of 0.2 (found in three molecules i.e. **132**, **141** and **171**). The coefficient plot as shown in Fig. S1–S2 depicts (SI-2) the contribution of individual variables with reference to the algebraic sign. The chemical features enhancing or reducing mutagenicity of TA98-S9 endpoints are summarized in Fig. 3. The goodness of fit for the training set and corresponding predictive ability (test set) in the form of scatter plots are represented in Fig. 4.

3.2. Modeling of Salmonella typhimurium strain (TA98) with microsomal activation (S9) using 309 aromatic and hetero aromatic compounds

The dataset of 309 compounds was further split into a training set containing 232 compounds for model building and a test set of 77 molecules in order to validate the models. The final two models for TA98 + S9 also exhibited moderate robustness as both the models could explain nearly 70% of the training set variance (68% in terms of LOO) while for the test set 68–70% variance was predicted. For details of the metric values, one can see Eqs. (7)–(8) as given below. Both the final models were derived of two highly significant variables namely ETA_Epsilon_3 (a measure of electronegative atom count), a sub-type of extended topochemical indices and an *E*-state index namely SaaaC (sum of atom-type E-State of::C:) having VIP scores of 1.55–1.77 and 1.25–1.45, respectively. Apart from the descriptor B06[C—C], the remaining variables were considered relatively less significant as suggested by their respective low VIP scores. The loading plot of TA98 + S9 models highlights more towards impact of mutagenicity enhancing factor which are placed in close proximity to each other on co-ordinate 1. The Y-randomization plots gave intercepts for determinant coefficients of 0.0002 to −0.0009 and those for cross-validated determinant coefficients of −0.25 and −0.21 proving robustness and non-random nature of the models. Like the TA98-S9 models, intercorrelation among the descriptors were kept below 0.9 in order to avoid problems of over fitting. The variables appearing in the TA98 + S9 models were grouped into three subcategories based on their features. The first group consists of descriptors showing presence of more electronegative elements include ETA_Epsilon_3, B02[N—N], F02[N—N], nPyridines, nImidazoles and SaaNH, while the second group represented hydrophobic moieties such as C-034, C-027, B06[C—C], SaaaC and D/Dtr09. The remaining variable sssCH was grouped in the third group influencing branching in a molecule. The various qualitative plots generated by SIMCA tool are given in Fig. S3–S4 in Supplementary Materials (SI-2).

3.2.1. Mechanistic interpretation of TA98 + S9 models

$$Log(TA98 + S9)rev/nmol = -54.056 + 120.748 \times \text{ETA\_Epsilon\_3} + 0.224 \times \text{SaaaC} + 0.614 \times \text{B06[C - C]} + 0.010 \times \text{D/Dtr09} - 1.473 \times \text{C - 034} + 1.585 \times \text{F02[N - N]} - 0.743 \times \text{SsssCH} - 1.658 \times \text{nPyridines} \quad (7)$$

$n_{training} = 232$, LV = 4, $R^2 = 0.701$, $Q^2 = 0.680$, $r^2_{m(LOO)} = 0.566$, $\Delta r^2_{m(LOO)} = 0.200$, MAE = 0.767

$n_{test} = 77$, $R^2_{pred} = 0.696$, $Q^2_{F2} = 0.695$, $r^2_{m(test)} = 0.588$, $\Delta r^2_{m(test)} = 0.162$, MAE = 0.747

$$Log(TA98 + S9)rev/nmol = -98.228 + 222.822 \times \text{ETA\_Epsilon\_3} + 0.09347 \times \text{SaaaC} + 3.535 \times \text{B02[N - N]} + 1.43394 \times \text{nImidazoles} - 0.753 \times \text{SaaNH} - 1.364 \times \text{C - 027} - 1.270 \times \text{nPyridines} - 0.775 \times \text{SsssCH} \quad (8)$$

$n_{training} = 232$, LV = 5, $R^2 = 0.700$, $Q^2 = 0.683$, $r^2_{m(LOO)} = 0.567$, $\Delta r^2_{m(LOO)} = 0.220$, MAE = 0.766

$n_{test} = 77$, $R^2_{pred} = 0.677$, $Q^2_{F2} = 0.676$, $r^2_{m(test)} = 0.568$, $\Delta r^2_{m(test)} = 0.144$, MAE = 0.768

The presence of electronegative atom count due to ETA_Epsilon_3 was found to be the most crucial descriptor enhancing mutagenicity.

The descriptor ETA_Epsilon_3 is a second generation extended topochemical variable (see Eq. (9)); it gives the summation of epsilon (ε) values relative to the total number of atoms including hydrogen in the connected molecular graph of the reference alkane.

$$\varepsilon_3 = \frac{[\sum \varepsilon]_R}{N_R} \tag{9}$$

Here, $\Sigma\varepsilon$ and N are summation of electronegative atoms, heteroatoms and multiple (double or triple) bonds replaced by carbon and single bonds and total number of atoms including hydrogen respectively in the molecular graph of the original structure. Here, R denotes the parent reference alkane having no functional moiety present within the molecule. The positive regression coefficient of ETA_Epsilon_3 suggests that with an increase in the electronegative element content (mainly nitrogen), the tendency of molecules to behave as potent mutagenic entity is enhanced in aromatics and hetero-aromatics. Other variables denoting the presence of high electronegative element content in the aromatic or hetero-aromatic amine datasets include a functional group count descriptor nImidazoles (number of imidazoles), 2D atom pair descriptors F02[N−N] (frequency of two nitrogens at topological distance 2) and B02[N−N] (presence or absence of the 2 nitrogen atoms at topological distance 2); all three of them were positively correlated with the mutagenicity of the studied chemicals. Some examples with higher electronegative element content (mainly nitrogen) include **115S, 116S** and **118S** whereas an inverse effect was seen in compounds having lower nitrogen content, for example **105S** and **288S**. The last variable denoting the electronegative count in a molecule was SaaNH (−-NH—moiety, where (−-) denotes aromatic bonds). The last fragment (aromatic bond-NH-aromatic bond) exerted a very little contribution to enhance the mutagenicity of aromatic and hetero aromatic compounds (low VIP score). A lot of such fragments were the parts of more toxic imidazole ring, and thus it can be inferred that the imidazole moiety plays a crucial role in regulating the mutagenicity of aromatic and hetero aromatic compounds. Some examples of compounds with higher electronegative content and enhanced mutagenicity include **115S**, **118S** and **119S,** whereas the reverse was seen in compounds **81S, 150S** and **265S**. With imidazole enhancing mutagenic potency of aromatic and hetero-aromatic compounds, a reverse was observed with the presence of pyridines. Despite it showing a negative regression coefficient, we cannot confidently comment on the mutagenic potency of pyridines as the level of significance of this variable is very low with VIP score of < 0.5.

The second subgroup represents the lipophilic bulk of the organic chemicals. Lipophilicity being an important criterion having a positive influence on mutagenicity, it is represented by several attributes in the models, the most important being an atom centered fragment SaaaC which represents sum of aromatic carbons ((−-C(−-)–), where "–"represents an aromatic bond). The positive coefficient of the SaaaC denotes that with an increase in the number of aromatic rings surrounded by carbons enhances the hydrophobicity thereby enhancing the mutagenicity values, as observed in the compounds **88S**, **177S** and **205S** having higher mutagenicity values of **3.8, 3.23** and **3.5**, respectively. Conversely, with a decrease in the number of aromatic rings surrounding by carbons decreases the mutagenic potential of the molecule as seen in compounds **15S**, **21S** and **298S** with mutagenicity values of −3, −3 and −2.7 respectively. The other lipophilicity enhancing variables appearing in the model was B06[C−C] (presence or absence of two carbons at topological distance 6) and D/Dtr09 (distance/detour ring index of order 9), both of these variables represent a larger chain length in the carbon skeleton. Higher values of these descriptors tend to increase the non-polar surface area and bulkiness of the molecule. It was observed that the mutagenicity of the compounds increases with an increase in non-polar surface area as evident from the molecules **174S**, **177S** and **220S** where most of them contained a flouranthane moiety having a larger non-polar surface. Similarly with a reduced non-polar surface area as seen in compounds **28S, 48S** and **304S** mutagenicity decreases. The remaining two variables enhancing lipophilicity include C-034 (R–CR..X) and C-027 (R–CH—X) (where R is any group linked through carbon; X is any electronegative atom (O, N, S, P, Se, halogens); − is an aromatic bond as in benzene or delocalized bonds such as the N,O bond in a nitro group; .. denotes aromatic single bonds). These two variables represent very less number of molecules in spite of having a close relation with ETA_epsilon_3 descriptor where the latter has a larger influence in controlling mutagenicity of aromatic and hetero-aromatics chemicals for the same compounds as seen in **115S** and **118S**.

The last variable SsssCH represents the presence of tertiary carbon atoms and denotes branching in the molecule. The descriptor is an atom type *E*-state index calculated from the sum of E-states of > CH– fragment. This descriptor contributed negatively towards the mutagenicity of the aromatic and hetero aromatic compounds. Thus, we can infer that the highly branched organic chemicals tend to have lower potential to cause mutagenicity in *Salmonella typhimurium* when compared to the less branched molecules. Some of the more branched and less mutagenic compounds include **10S**, **84S** and **290S**, while the reverse was seen with compounds **88S**, **90S** and **102S**. The chemical features enhancing or reducing mutagenicity of TA98-S9 endpoints are summarized in Fig. 5**.** The scatter plots showed that the points were close to the line of fit for both TA98 + S9 models (see Fig. 6).

3.3. Modeling mutagens with molecular fragments derived from SiRMS software

In many cases, the conventional molecular descriptors fail to provide the definite features actually responsible for the desired response. To obviate such deficiency, the authors have additionally incorporated simplex molecular variables in order to identify the actual fragments present within the molecules with contribution to bacterial mutagenicity. The SiRMS variables constitute a group of 1D-4D tetratomic fragments; however, to avoid the complications of energy minimizations needed for 3D and 4D fragment computation, only 2D descriptors were used in the present study. The final selected models with fragments were slightly better in predictivity for the test sets when compared to the previous models for the respective endpoints (See Eqs. (10)–(13)).

3.3.1. Fragmental QSAR models against TA98-S9

$$\begin{aligned} Log(TA98-S9)rev/nmol = {} & -8.50 + 3.28 \times RDCHI \\ & + 0.0543 \times Fr5(d\_a)/I\_I\_I\_I\_I \\ & /1\_2s,\ 2\_4a,\ 3\_5a,\ 4\_5a/(Box2) \\ & + 0.0337 \times S\_A(lip)/B\_C\_C\_C/1\_2a,\ 3\_4a \\ & /3(Box1) + 0.264 \times Fr5(att)/E\_E\_E\_E\_E \\ & /1\_2s,\ 1\_3s,\ 2\_4a,\ 3\_5a/(Box3) \\ & + 1.15 \times F09[N-N] - 0.841 \times NRS \\ & - 0.127 \times F07[C-C] \\ & - 0.0126 \times S\_A(rep)/A\_B\_B\_B/1\_2s,\ 3\_4a \\ & /3(Box4) \end{aligned}$$

$n_{training} = 219$, LV = 5, $R^2 = 0.738$, $Q^2 = 0.710$, $r^2_{m(LOO)} = 0.604$, $\Delta r^2_{m(LOO)} = 0.187$, MAE = 0.847

$n_{test} = 72$, $R^2_{pred} = 0.747$, $Q^2_{F2} = 0.739$, $r^2_{m(test)} = 0.640$, $\Delta r^2_{m(test)} = 0.176$, MAE = 0.753 (10)

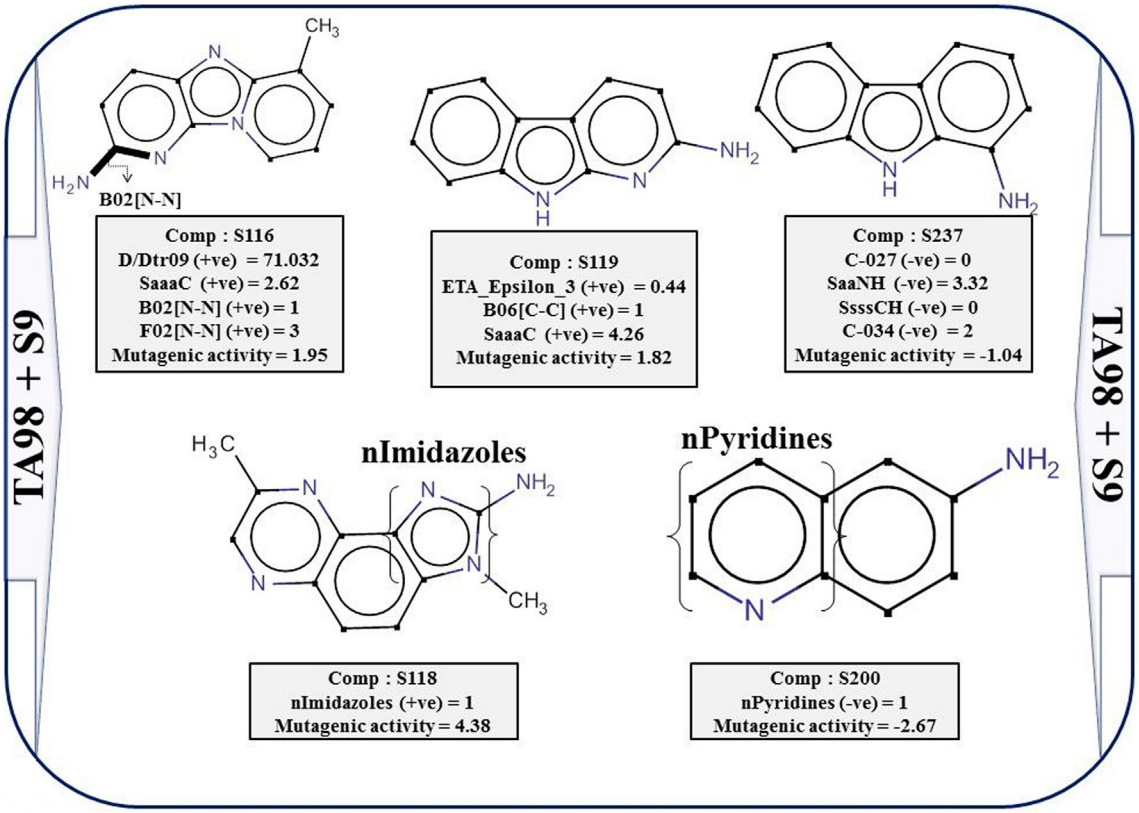

Fig. 5. Molecular features enhancing or reducing mutagenicity of TA98 + S9.

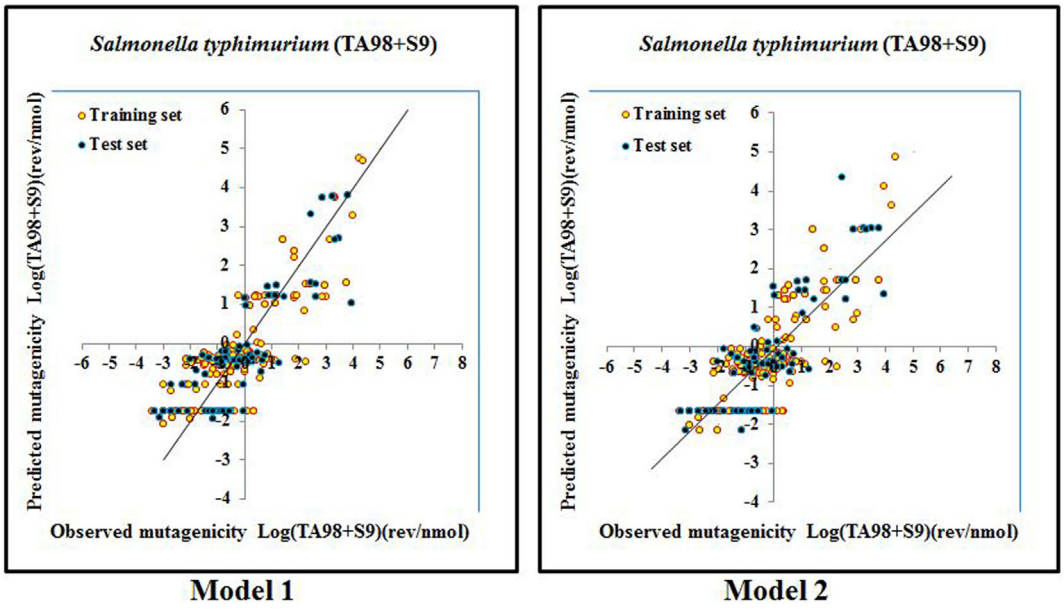

Fig. 6. Scatter plots of observed v/s predicted responses against model 1 and 2 of TA98 + S9.

$$Log(TA98 - S9)rev/nmol = -7.66 + 2.97 \times RDCHI + 1.33 \times F09[N - N] + 0.51 \times S\_A(chg)/B\_B\_B\_C /1\_2s, 2\_3a, 3\_4a/6(Box5) + 0.0821 \times Fr5(d\_a)/I\_I\_I\_I\_I /1\_2s, 2\_4a, 3\_5a, 4\_5a/(Box2) + 0.0319 \times S\_A(lip)/B\_C\_C\_C/1\_2a, 3\_4a /3(Box1) - 0.0204 \times S\_A(rep)/A\_B\_B\_B /1\_2s, 3\_4a/3 - 0.787 \times Fr5(d\_a)/A\_I\_I\_I\_I /1\_3d, 2\_4s, 3\_5s, 4\_5d/(Box6) - 1.061 \times NRS$$

$n_{training} = 219$, LV = 6, $R^2 = 0.739$, $Q^2 = 0.709$, $r^2_{m(LOO)} = 0.603$, $\Delta r^2_{m(LOO)} = 0.187$, MAE = 0.825

$n_{test} = 72$, $R^2_{pred} = 0.757$, $Q^2_{F2} = 0.749$, $r^2_{m(test)} = 0.650$, $\Delta r^2_{m(test)} = 0.195$, MAE = 0.750 (11)

The TA98-S9 models with fragmental variables showed a slightly better robustness having six additional variables in addition to the previous descriptors. The fragmental QSAR models identified four features correlated positively with the bacterial mutagenicity (see Box 1–3 and 5 of Fig. 7) while the remaining two fragments (see Box 4, 6 of Fig. 7) exerted a negative influence in controlling mutagenicity of nitro aromatics. Another major notable point here is that the positively correlated features were rich in aromatic bonds along with nitrogen of the nitro group. The fragments of TA98-S9 models mainly hinted towards lipophilic and electronegative group dependent mutagenicity of nitro aromatic chemicals. The scatter plots for the TA98-S9 SiRMS models are given in Fig. 8.

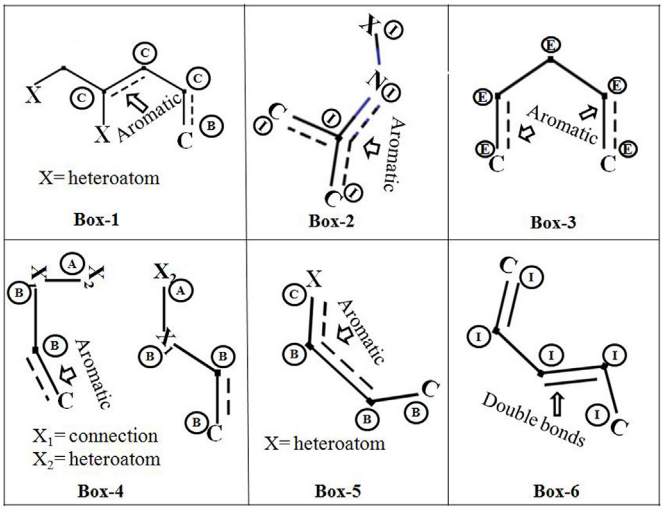

Fig. 7. Simplex representation of molecular structures (SiRMS) fragments appearing in both of the TA98-S9 model.

3.3.2. Fragmental QSAR models against TA98 + S9

$$Log(TA98 + S9)rev/nmol = =-3.035 + 0.229 \times Fr3(lip)/C\_C\_C /1\_3a, 2\_3a/(Box5) + 0.32516 \times Fr3(lip)/B\_C\_C/1\_3a, 2\_3s /(Box1) + 0.908 \times Fr5(elm)/C\_C\_C\_C\_C /1\_2s, 2\_3a, 3\_4s, 4\_5a/(Box7) + 0.46272 \times S\_A(chg)/A\_B\_D\_D /1\_3a, 2\_4a/3(Box3) - 1.659 \times Fr5(chg)/B\_B\_C\_C\_D /1\_4a, 2\_3a, 2\_4a, 4\_5a/(Box8) + 1.24174 \times Fr3(elm)/C\_N\_N/1\_2s, 1\_3a /(Box2) + 0.816 \times Fr5(lip)/B\_B\_B\_B\_C /1\_2s, 2\_5a, 3\_4a, 4\_5a/(Box6) - 0.052 \times Fr3(chg)/B\_C\_C/1\_2s, 1\_3s /(Box4)$$

$n_{training} = 232$, LV = 4, $R^2 = 0.726$, $Q^2 = 0.704$, $r^2_{m(LOO)} = 0.596$, $\Delta r^2_{m(LOO)} = 0.197$, MAE = 0.733

$n_{test} = 77$, $R^2_{pred} = 0.760$, $Q^2_{F2} = 0.754$, $r^2_{m(test)} = 0.663$, $\Delta r^2_{m(test)} = 0.147$, MAE = 0.485 (12)

$$Log(TA98 + S9)rev/nmol = -3.170 + 0.240 \times Fr3(lip)/C\_C\_C /1\_3a, 2\_3a/(Box5) + 0.328 \times Fr3(lip)/B\_C\_C/1\_3a, 2\_3s /(Box1) + 0.458 \times S\_A(chg)/A\_B\_D\_D /1\_3a, 2\_4a/3(Box3) + 0.822 \times Fr5(elm)/C\_C\_C\_C\_C /1\_2s, 2\_3a, 3\_4s, 4\_5a/(Box7) - 1.824 \times Fr5(chg)/B\_B\_C\_C\_D /1\_4a, 2\_3a, 2\_4a, 4\_5a/(Box8) + 1.379 \times Fr3(elm)/C\_N\_N/1\_2s, 1\_3a /(Box2) + 0.853 \times Fr5(lip)/B\_B\_B\_B\_C /1\_2s, 2\_5a, 3\_4a, 4\_5a/(Box6) - 0.509 \times Fr5(d\_a)/D\_I\_I\_I\_I /1\_2s, 2\_5a, 3\_5a, 4\_5a/(Box9)$$

$n_{training} = 232$, LV = 6, $R^2 = 0.725$, $Q^2 = 0.702$, $r^2_{m(LOO)} = 0.594$, $\Delta r^2_{m(LOO)} = 0.194$, MAE = 0.731

$n_{test} = 77$, $R^2_{pred} = 0.760$, $Q^2_{F2} = 0.760$, $r^2_{m(test)} = 0.663$, $\Delta r^2_{m(test)} = 0.157$, MAE = 0.496 (13)

The performance of TA98 + S9 fragmental QSAR models were superior when compared to the previous two models with conventional descriptors providing better robustness as well as predictivity. Both the models solely consist of fragmental variables with nine fragments in total. Six (see Box 2–7 of Fig. 9) out of nine variables exerted positive contributions towards mutagenicity while the remaining three variables (see Box 1, 8 and 9 of Fig. 9) exerted a negative influence on bacterial mutagenicity. Like the TA98-S9 models, lipophilicity proved to be a major contributing feature for mutagenicity as evident from its repetition in several positively correlated fragments in the TA98 + S9 models. Among the negatively correlated features was branching, as seen in Box 9 of Fig. 9. Additionally, plots of observed against predicted response for theTA98 + S9 fragment models are shown in Fig. 10.

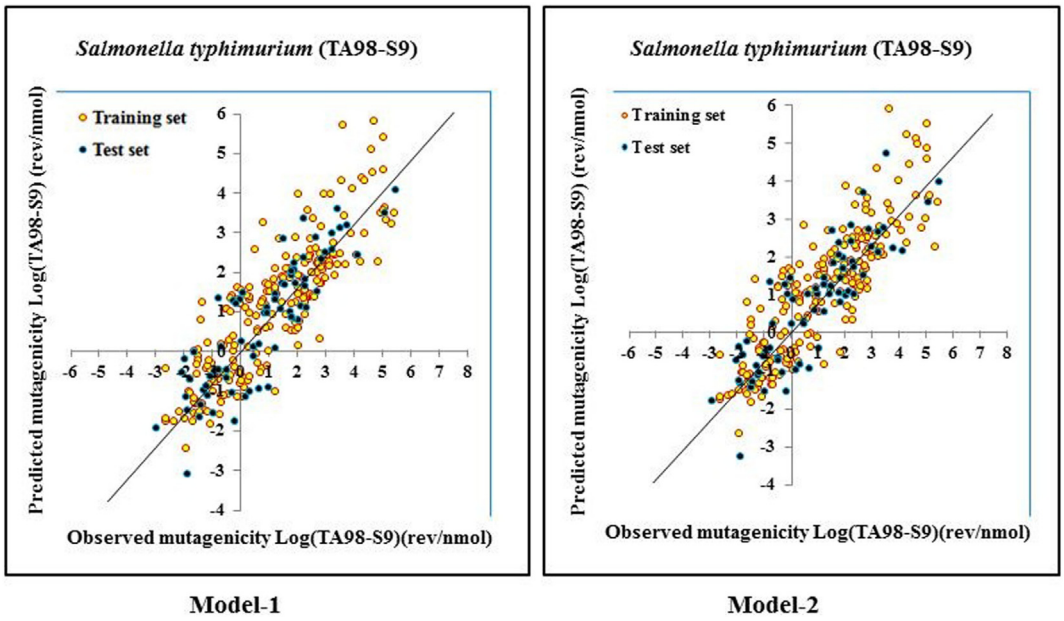

Fig. 8. Scatter plots of observed v/s predicted responses against models 1 and 2 of TA98-S9 fragment models.

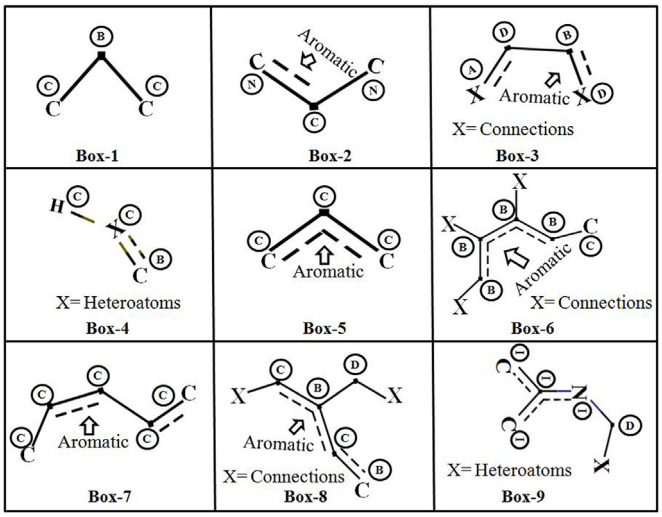

Fig. 9. Simplex representation of molecular structures (SiRMS) fragments appearing in both of the TA98 + S9 model.

3.4. Summary of the overall finding of individual QSAR models: A mechanistic view

To summarize, the authors have assembled various model features into four different groups based on their physicochemical attributes responsible for enhancing or reducing mutagenicity. All the chemical features obtained in the process can simply be grouped into four classes: (1) effect of lipophilicity on mutagenicity, (2) mutagenicity due to electronegative atoms such as nitrogen, (3) mutagenicity due to unsaturation, and finally (4) effect of branching on mutagenicity.

3.4.1. Effect of non-polar moieties on mutagenicity

The lipophilicity of polycyclic, aromatic and planar chemical structures enables them to readily penetrate cellular membranes (Yu et al., 2016). Furthermore, these moieties undergo metabolism (phase-I and II), hence, these molecules are converted into more water-soluble entities in order to be removed from the body easily. However, these chemical structures can also be converted to more mutagenic or carcinogenic metabolites because of their heavy lipophilic nature. Some of the highly contributing lipophilicity variables such as higher connectivity indices (X5Av, X5A), atom pair indices (B08[Cl—Cl] and F07[C—C]) and fragmental variables (Box 1 and 4 of Fig. 7 and Box 4–9 in Fig. 9) contributed to the mutagenicity of the compounds of the studied dataset. Since these molecules are highly lipophilic in nature, they tend to promote fluidization of phospholipid bilayer of cell membrane in order to facilitate more and more accumulation within the cell. This, in turn, it can lead to the formation of several intermediates. Finally, there is a high possibility of these intermediates entering into redox cycle thereby causing oxidative stress and inducing the production of reactive radical cations inside the cell (Yu et al., 2016). These radical cations have potential to form covalent bonds with the exocyclic amino group of the phosphodiester bonds, a leading cause for mutations leading to genotoxicity. Fig. 11 schematically demonstrates the probable lipophilicity induced mutagenicity.

3.4.2. Effect of electronegativity on mutagenicity

Electronegative features present in the planar polycyclic aromatic molecules intercalate into DNA in the space between two adjacent base pairs via nucleophilic aromatic substitution. This might induce the changes in DNA structure like double helix unwinding and elongation of DNA strands. Several such features were predominant in all the developed QSAR models against both the endpoints. Some of the most predominant descriptors correlating electronegativity against mutagenicity include ETA indices like ETA_epsilon_3 and ETA_Beta_ns, atom pair variables (F02[N—N, F09[N—N] and B02[N—N]), presence of imidazole ring and several simplex fragments (see Box 2, 3 and 5 in Fig. 7 and Box 2 and 3 of Fig. 9**).** These structural modifications may lead to mutagenicity because of the sequential alterations in the DNA strands; furthermore, the DNA gyrase may not distinguish the actual DNA and mutated DNA leading to more lethal conditions. Besides mutations, these features can also lead to retardation or inhibition of transcription and replication. The process is summarized in Fig. 12**.**

3.4.3. Effect of unsaturation on mutagenicity

There is also a sufficient number of features present within the QSAR models possibly hinting towards unsaturation leading to mutations. Some of the important descriptors include ETA indices like ETA_Beta_ns and ETA_dBeta, connectivity indices like RDCHI and several molecular fragments such as Box 2 and 6 of Fig. 7 and Box 3 of Fig. 9**.** Like lipophilicity, unsaturation is also capable of augmenting

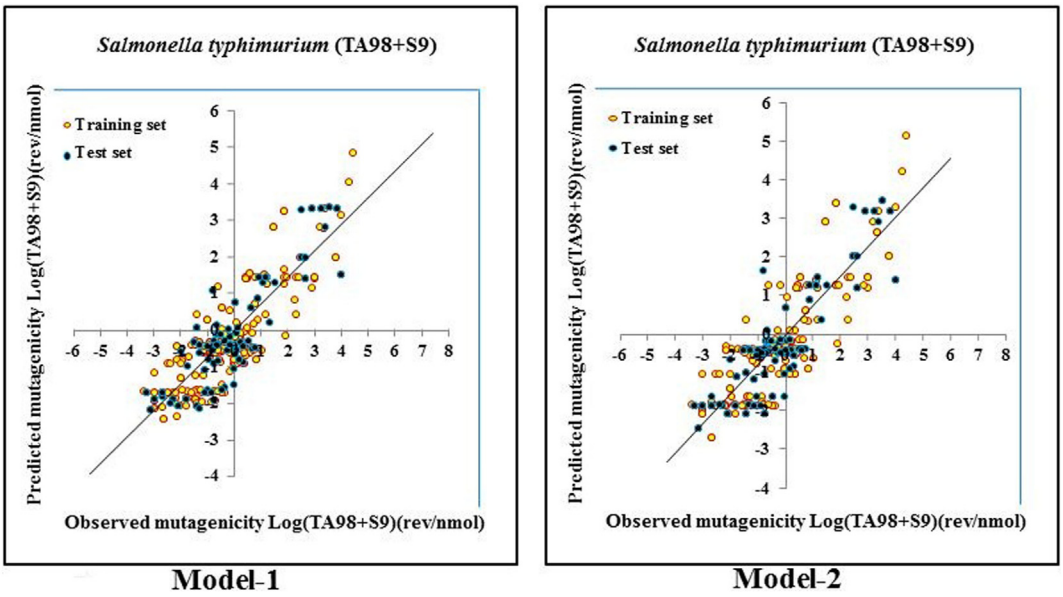

Fig. 10. Scatter plots of observed v/s predicted responses against models 1 and 2 of TA98 + S9 fragment models.

membrane fluidization which results in more invasions of the organic chemicals within the cell. The process is followed by metabolism giving reactive intermediates leading to adduct formation as shown in Fig. 13. Two types of adducts are reported to form with these reactive intermediates named "Bay region adducts" and "Fjord region adducts" where the former is less reactive than the other (Munoz and Albores, 2011; Yu et al., 2016). These adducts go on to form the covalent bonds with the exocyclic amino group of the adenine and guanine as shown in Fig. 13. To be specific, "Bay region adducts" form covalent bonds with guanine, whereas "Fjord region adducts" form covalent bonds with adenine (Munoz and Albores, 2011; Yu et al., 2016).

3.4.4. *Effect of branching on mutagenicity*

The last effective attribute having a negative influence on mutagenicity was found to be enhanced branching within the molecules. Some of variables influencing more branching in the molecules include SsssCH and several fragments as shown in Box 1, 2, 8 and 9 of Fig. 9**.** In general, highly branched organic chemicals show lower toxicity. The reason of their less toxic behaviour could lie in their physical properties which alters the lipophilic bulk and increase of the hydrophilic nature. This is followed by the cascade of membrane fluidization which is sufficiently hampered due to more hydrophilic nature of the entity thus ending the mutagenic pathway as explained in lipophilic section. For ease of understanding, please see Fig. 14.

3.5. *Applicability domain analysis*

The applicability domains (ADs) of individual models were checked using DModX approach embedded in Simca-P (version10.0), a recommended method for PLS models. The AD was checked at 99% confidence level with D-critical limit of 0.01. All the developed QSAR models could cover a minimum of 95% hypothetical AD space with

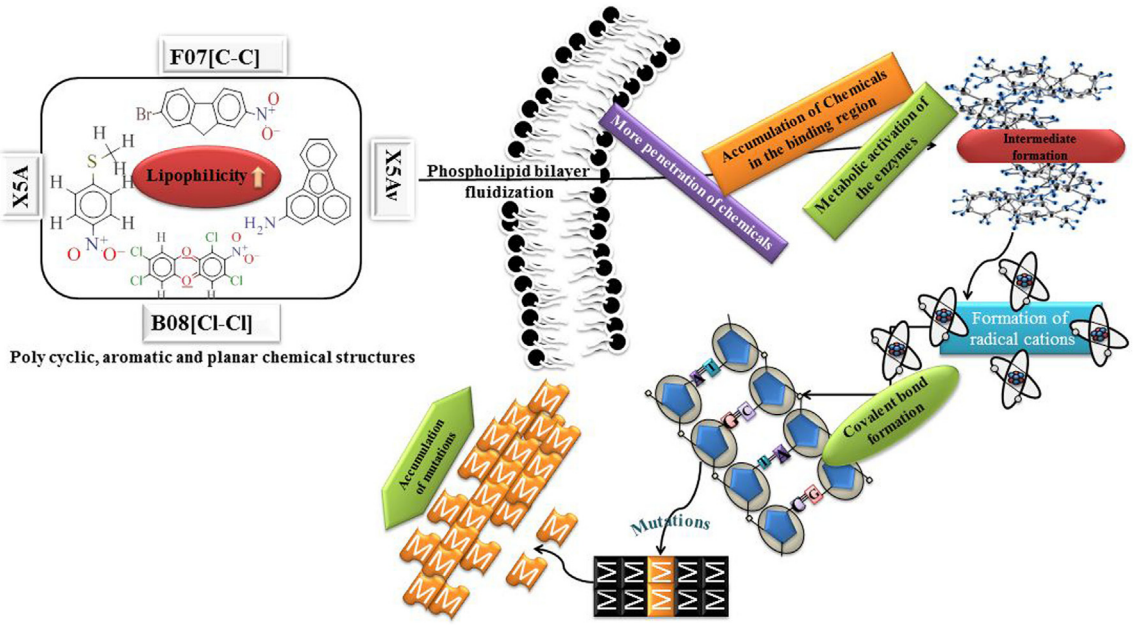

Fig. 11. Effect of lipophilic moieties on mutagenicity present in the organic chemicals.

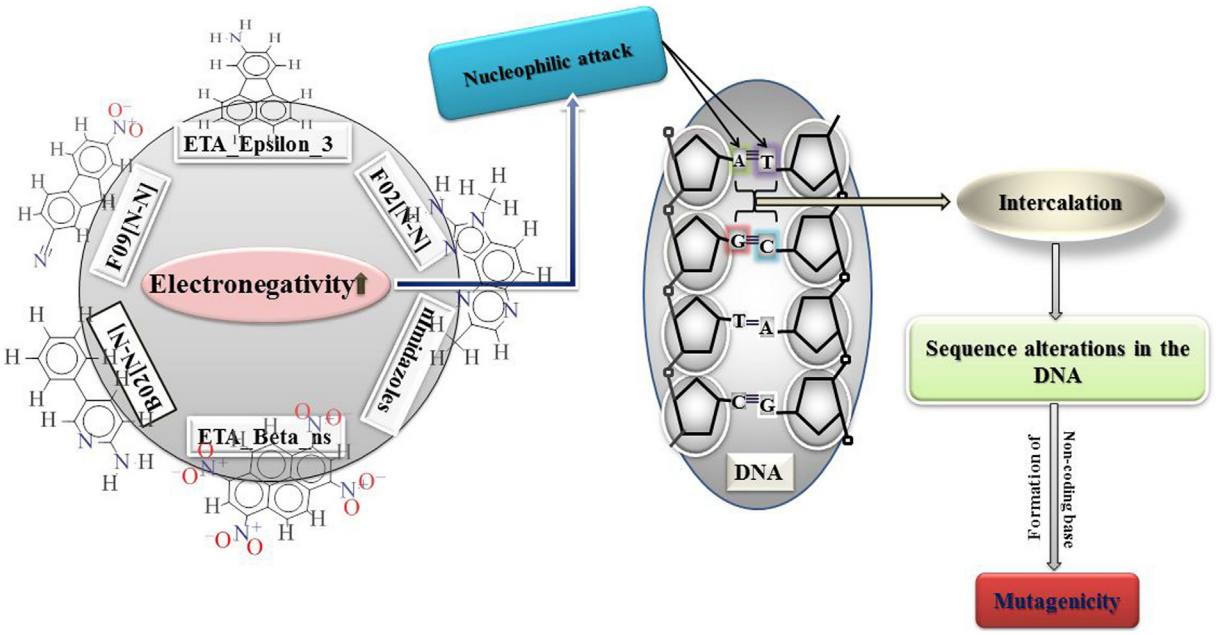

Fig. 12. Effect of electronegative moieties on mutagenicity present in the organic chemicals.

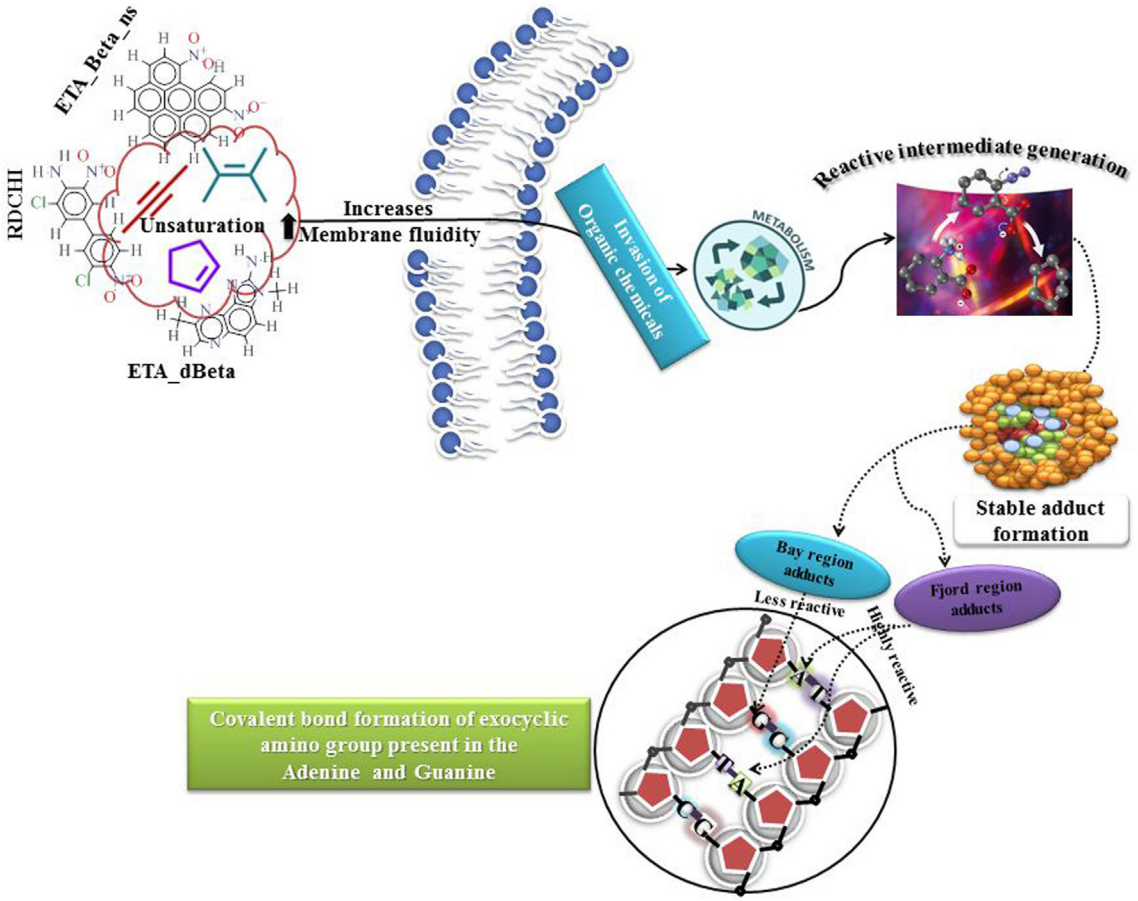

Fig. 13. Effect of unsaturation on mutagenicity present in the organic chemicals.

TA98-S9 models covering maximum of 96–99% of chemical space. Although a number of outliers were identified in the training set and a number of chemicals were outside the domain in the test set, the outliers were retained in the final models since the developed models could predict their mutagenicity with absolute error of less than two (< 2) log units. For the details of AD analysis, please see Fig. S5–8 in SI-2.

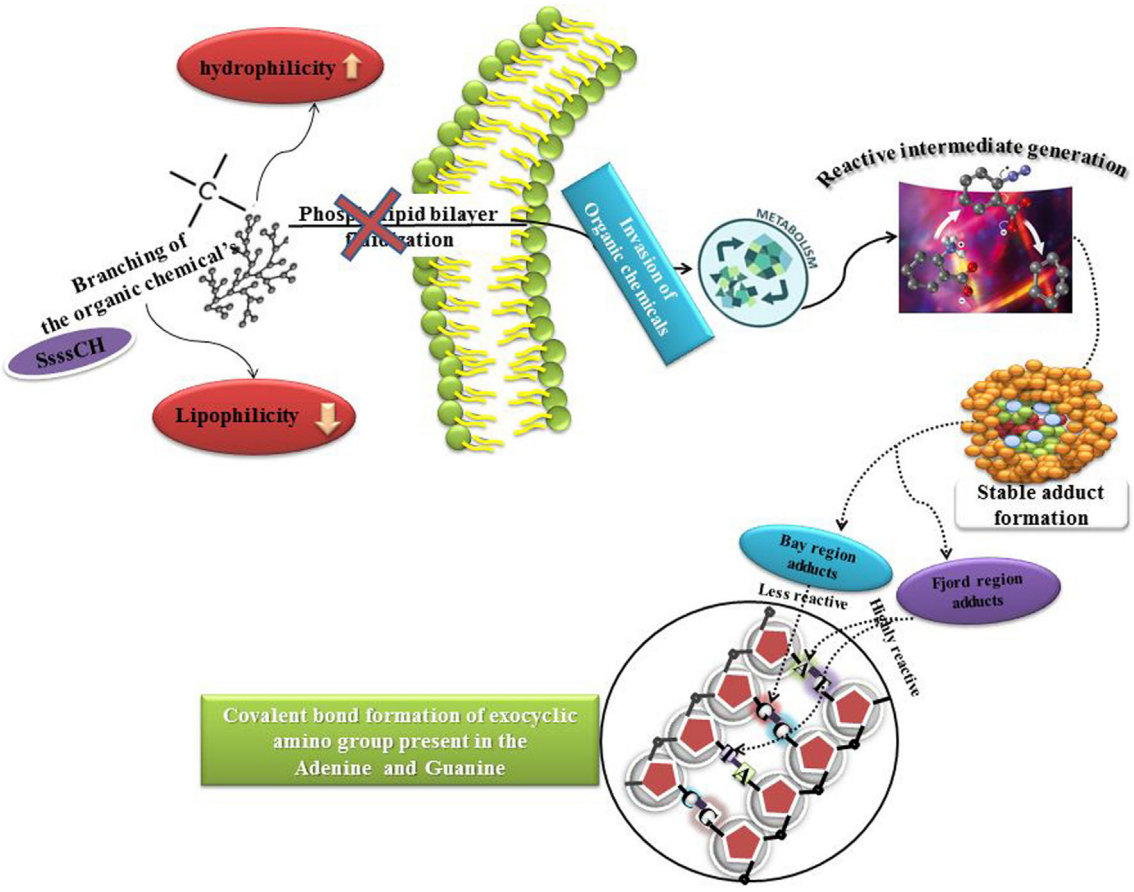

Fig. 14. Effect of unsaturation on mutagenicity present in the organic chemicals.

3.6. Comparison of present models with existing QSAR models on mutagenicity

The application of novel QSAR models cannot be justified unless compared with practicing standards. Thus, we have tried to give a brief comparison with some of the known models already developed in order to study the probable mode of mutagenicity of selective organic chemicals against *Salmonella typhimurium*. The authors make a point here that the present manuscript outperforms all the previous linear QSAR models in terms of both statistical quality as well as domain of applicability. For a detailed comparison, please see Table 2.

4. Conclusion

The present paper proposes sufficiently robust QSAR models employing simple 2D variables generated from Dragon and PaDEL-descriptor softwares against the TA98-S9 and TA98 + S9 endpoints. Additionally, more predictive QSAR models consisting of simplex fragmental variables were also proposed. The data division, model development and AD estimation were performed following the strict guidelines for QSAR model validation. The study focuses mainly on the mechanistic approach of QSAR application by providing a detailed analysis of probable cause of mutagenicity by simply taking into

Table 2
The brief comparison of already existing linear 2D QSAR models with our models on mutagenicity of selected organic chemicals against *S. typhimurium*.

Sr no.	Working group	Type of chemicals	Endpoint(S)	Dataset size	R^2	Q^2	Q^2F_1	Remarks[a]
1	Current manuscript	Nitro aromatics	TA98-S9	291	0.73	0.71	0.76	Model without fragments
		Nitro aromatics	**TA98-S9**	**291**	**0.74**	**0.71**	**0.76**	**Model with fragments**
		Amino aromatics	TA98 + S9	309	0.70	0.68	0.70	Model without fragments
		Amino aromatics	**TA98 + S9**	**309**	**0.73**	**0.70**	**0.76**	**Model with fragments**
2	Ding et al. (2017)	Nitro aromatics	TA98-S9	282	0.72	0.69	0.70	Complicated descriptors used
3	Bhat et al. (2005)	Amino aromatics	TA98 + S9	181	0.67	0.64	0.65	Multiple linear regression model
4	Leong et al. (2010)	Amino aromatics	TA98 + S9	122	0.35	–	0.77	No valid PLS models
5	Garg et al. (2002)	Amino aromatics	TA98 + S9	43	0.85	–	–	Very small dataset, no validation set
6	Gadaleta et al. (2017)	Azoaromatics	Non specific	354	–	–	–	Classification model
7	Pasha et al. (2008)	Azoaromatics	TA98 + S9	43	0.95	0.51	0.65	3D model, non-robust
8	Ren (2003)	Phenols	*Tetrahymena pyriformis*	200	0.65	–	–	No validation set, use of more complex descriptors
9	Gramatica et al. (2007)	nitro-PAHs[b]	TA100-S9	48	0.88	0.86	0.75	Very small dataset and used of more complex descriptors
10	Abbasitabar and Zare-Shahabadi (2017)	Phenols	*Tetrahymena pyriformis*	206	0.72	0.69	0.69	Use of more complex descriptors

[a] The best models are highlighted with **bold.**
[b] Poly aromatic hydrocarbons.

account QSAR equation (OECD principle 5 for model interpretations). From the statistical point of view, the models were validated using some stringent metrics such as r_m^2 and MAE. Finally, the established models were compared with many already existing QSAR models on mutagenicity against *S. typhimurium* and related species. The obtained QSAR model outperforms almost all the existing linear QSAR models for at least some of the validation parameters. The authors strongly believe that the performed methodology will greatly help various groups of researchers working in the field of mutagenicity of synthetic organic chemicals. Lastly, the developed models can also be used to screen untested or unknown or not yet synthesized chemicals based on their acute mutagenic potential.

Author contributions

All the coauthors have equally contributed to this work. The authors have read and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

GKJ thanks the Department of Pharmaceuticals, Ministry of Chemicals and Fertilizers, Govt. of India for a fellowship. KK thanks Indian Council of Medical Research, New Delhi for financial support in the form of a senior research fellowship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tiv.2020.104768.

References

Abbasitabar, F., Zare-Shahabadi, V., 2017. In silico prediction of toxicity of phenols to Tetrahymena pyriformis by using genetic algorithm and decision tree-based modeling approach. Chemosphere 172, 249–259.
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.
Bhat, K.L., Hayik, S., Sztandera, L., Bock, C.W., 2005. Mutagenicity of aromatic and heteroaromatic amines and related compounds: A QSAR investigation. QSAR Comb. Sci. 24, 831–843.
Cadet, J., Voituriez, L., Hruska, F.E., Kan, L.-S., Leeuw, F.A.A.M., Altona, C., 1985. Characterization of thymidine ultraviolet photoproducts. Cyclobutane dimers and 5, 6-dihydrothymidines. Can. J. Chem. 63, 2861–2868.
Cheng, T.-F., Hu, X., Gnatt, A., Brooks, P.J., 2008. Differential blocking effects of the acetaldehyde-derived DNA lesion N2-ethyl-2`-deoxyguanosine on transcription by multisubunit and single subunit RNA polymerases. J. Biol. Chem. 283, 27820–27828.
Chung, K.-T., Cerniglia, C.E., 1992. Mutagenicity of azo dyes: Structure-activity relationships. Mutat. Res. Rev. Gen. Tox. 277, 201–220.
Chung, K.-T., Kirkovsky, L., Kirkovsky, A., Purcell, W.P., 1997. Review of mutagenicity of monocyclic aromatic amines: Quantitative structure-activity relationships. Mutat. Res. Rev. Mutat. 387, 1–16.
Dearden, J.C., 2016. The history and development of quantitative structure-activity relationships (QSARs). IJQSPR. 1, 44.
Degan, P., Shigenaga, M.K., Park, E.-M., Alperin, P.E., Ames, B.N., 1991. Immunoaffinity isolation of urinary 8-hydroxy-2′ deoxygunosine and 8-hydroxyguanine and quantitation of 8-hydroxy-2′ deoxyguanosine in DNA by polyclonal antibodies. Carcinogenesis 12, 865–871.
Ding, Y.-L., Lyu, Y.-C., Leong, M.K., 2017. In silico prediction of the mutagenicity of nitroaromatic compounds using a novel two-QSAR approach. Toxicol. in Vitro 40, 102–114.
Fraga, C.G., Shigenaga, M.K., Park, J.-W., Degan, P., Ames, B.N., 1990. Oxidative damage to DNA during aging: 8-hydroxy-2′-deoxyguanosine in rat organ DNA and urine. Proc. Natl. Acad. Sci. U. S. A. 87, 4533–4537.
Fu, P.P., 1990. Metabolism of nitro-polycyclic aromatic hydrocarbons. Drug Metab. Rev. 22, 209–268.
Fu, P.P., Herreno-Saenz, D., 1999. Nitro-polycyclic aromatic hydrocarbons: A class of genotoxic environmental pollutants. J. Environ. Sci. Health C 17, 1–43.
Gadaleta, D., Porta, N., Vrontaki, E., Manganelli, S., Manganaro, A., Sello, G., Honma, M., Benfenati, E., 2017. Integrating computational methods to predict mutagenicity of aromatic azo compounds. J. Environ. Sci. Health C 35, 239–257.
Garg, A., Bhat, K.L., Bock, C.W., 2002. Mutagenicity of aminoazobenzene dyes and related structures: A QSAR/QPAR investigation. Dyes Pigments 55, 35–52.
Golmohammadi, H., Dashtbozorgi, Z., Acree Jr., W.E., 2012. Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. Eur. J. Pharm. Sci. 47, 421–429.
Gramatica, P., Pilutti, P., Papa, E., 2007. Approaches for externally validated QSAR modelling of nitrated polycyclic aromatic hydrocarbon mutagenicity. SAR QSAR Environ. Res. 18, 169–178.
Hossain, K.A., Roy, K., 2018. Chemometric modeling of aquatic toxicity of contaminants of emerging concern (CECs) in *Dugesia japonica* and its interspecies correlation with daphnia and fish: QSTR and QSTTR approaches. Ecotoxicol. Environ. Saf. 166, 92–101.
Khan, K., Roy, K., 2019. Ecotoxicological QSAR modelling of organic chemicals against *Pseudokirchneriella subcapitata* using consensus predictions approach. SAR QSAR Environ. Res. 30, 665–681.
Khan, K., Kar, S., Sanderson, H., Roy, K., Leszczynski, J., 2019. Ecotoxicological modeling, ranking and prioritization of pharmaceuticals using QSTR and i-QSTTR approaches: Application of 2D and fragment based descriptors. Mol. Inform. 38, 1800078.
Khan, K., Baderna, D., Cappelli, C., Toma, C., Lombardo, A., Roy, K., Benfenati, E., 2019a. Ecotoxicological QSAR modeling of organic compounds against fish: application of fragment based descriptors in feature analysis. Aquat. Toxicol. 212, 162–174.
Khan, K., Benfenati, E., Roy, K., 2019b. Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: Ranking and prioritization of the DrugBank database compounds. Ecotoxicol. Environ. Saf. 168, 287–297.
Khan, K., Roy, K., Benfenati, E., 2019c. Ecotoxicological QSAR modeling of endocrine disruptor chemicals. J. Hazard. Mater. 369, 707–718.
Kuz'min, V.E., Artemenko, A.G., Polischuk, P.G., Muratov, E.N., Hromov, A.I., Liahovskiy, A.V., Andronati, S.A., Makan, S.Y., 2005. Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. J. Mol. Model. 11, 457–467.
Leong, M.K., Lin, S.-W., Chen, H.-B., Tsai, F.-Y., 2010. Predicting mutagenicity of aromatic amines by various machine learning approaches. Toxicol. Sci. 116, 498–513.
Maisanaba, S., Prieto, A.I., Pichardo, S., Jorda-Beneyto, M., Aucejo, S., Jos, A., 2015. Cytotoxicity and mutagenicity assessment of organomodified clays potentially used in food packaging. Toxicol. in Vitro 29, 1222–1230.
Mauri, A., Consonni, V., Pavan, M., Todeschini, R., 2006. Dragon software: An easy approach to molecular descriptor calculations. Match 56, 237–248.
Moretti, M., Marcarelli, M., Villarini, M., Fatigoni, C., Scassellati-Sforzolini, G., Pasquini, R., 2002. In vitro testing for genotoxicity of the herbicide terbutryn: Cytogenetic and primary DNA damage. Toxicol. in Vitro 16, 81–88.
Munoz, B., Albores, A., 2011. DNA damage caused by polycyclic aromatic hydrocarbons: mechanisms and markers. Select. Topics DNA Rep. 201, 125–143.
O'Brien, T., Mandel, H.G., Pritchard, D.E., Patierno, S.R., 2002. Critical role of chromium (Cr)- DNA interactions in the formation of Cr-induced polymerase arresting lesions. Biochemistry 41, 12529–12537.
Pasha, F.A., Muddassar, M., Chung, H.W., Cho, S.J., Cho, H., 2008. Hologram and 3D-quantitative structure toxicity relationship studies of azo dyes. J. Mol. Model. 14, 293–302.
Rashid, K.A., Arjmand, M., Sandertnann, H., Mumma, R.O., 1987. Mutagenicity of chloroaniline/lignin metabolites in the *Salmonella*/microsome assay. J. Environ. Sci. Heal. B 22, 721–729.
Ravanat, J.-L., Douki, T., Cadet, J., 2001. Direct and indirect effects of UV radiation on DNA and its components. J. Photochem. Photobiol. B 63, 88–102.
Ren, S., 2003. Ecotoxicity prediction using mechanism-and non-mechanism-based QSARs: A preliminary study. Chemosphere 53, 1053–1065.
Rosenkranz, H.S., Klopman, G., 1995. Relationships between electronegativity and genotoxicity. Mutat. Res. Fund. Mol. M 328, 215–227.
Roy, K., Das, R.N., 2017. The "ETA" Indices in QSAR/QSPR/QSTR Research, Pharmaceutical Sciences: Breakthroughs in Research and Practice. IGI Global, pp. 978–1011.
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.
Roy, K., Mitra, I., Ojha, P.K., Kar, S., Das, R.N., Kabir, H., 2012. Introduction of rm2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. Chemometr. Intell. Lab. 118, 200–210.
Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometr. Intell. Lab. 152, 18–33.
Sabbioni, G., 1994. Hemoglobin binding of nitroarenes and quantitative structure-activity relationships. Chem. Res. Toxicol. 7, 267–274.
Stead, A.G., Hasselblad, V., Creason, J.P., Claxton, L., 1981. Modeling the Ames test. Mutat. Res. Environ. Mutat. 85, 13–27.
Tomita, I., Nakamura, Y., Aoki, N., Inui, N., 1982. Mutagenic/carcinogenic potential of DEHP and MEHP. Environ. Health Perspect. 45, 119–125.
Tsuboy, M.S., Angeli, J.P.F., Mantovani, M.S., Knasmuller, S., Umbuzeiro, G.A., Ribeiro, L.R., 2007. Genotoxic, mutagenic and cytotoxic effects of the commercial dye CI Disperse Blue 291 in the human hepatic cell line HepG2. Toxicol. in Vitro 21, 1650–1655.
Umetrics, M., 2013. User Guide to SIMCA. MKS Umetrics AB, Malmo (Sweden).
Vrtis, K.B., Markiewicz, R.P., Romano, L.J., Rueda, D., 2013. Carcinogenic adducts induce distinct DNA polymerase binding orientations. Nucleic Acids Res. 41, 7843–7853.
Wang, X., Lin, Z., Yin, D., Liu, S., Wang, L., 2005. 2D/3D-QSAR comparative study on mutagenicity of nitroaromatics. SCI China Ser. B 48, 246–252.

Wilson Iii, D.M., Barsky, D., 2001. The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA. Mutat. Res./DNA Repair 485, 283–307.
Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. 58, 109–130.
Yap, C.W., 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474.
Yasui, M., Matsui, S., Ihara, M., Laxmi, Y.R.S., Shibutani, S., Matsuda, T., 2001. Translesional synthesis on a DNA template containing N 2-methyl-2â€^{2}-deoxyguanosine catalyzed by the Klenow fragment of *Escherichia coli* DNA polymerase I. Nucleic Acids Res. 29, 1994–2001.
Yu, M.-H., Tsunoda, H., Tsunoda, M., 2016. Environmental Toxicology: Biological and Health Effects of Pollutants. CRC Press.