



A novel procedure for selection of molecular descriptors: QSAR model for mutagenicity of nitroaromatic compounds

Branislav Stankovic¹ · Filip Marinkovic²

Received: 19 April 2024 / Accepted: 22 August 2024 / Published online: 29 August 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Nitroaromatic compounds (NACs) stand out as pervasive organic pollutants, prompting an imperative need to investigate their hazardous effects. Computational chemistry methods play a crucial role in this exploration, offering a safer and more time-efficient approach, mandated by various legislations. In this study, our focus lay on the development of transparent, interpretable, reproducible, and publicly available methodologies aimed at deriving quantitative structure–activity relationship models and testing them by modelling the mutagenicity of NACs against the *Salmonella typhimurium* TA100 strain. Descriptors were selected from Mordred and RDKit molecular descriptors, along with several quantum chemistry descriptors. For that purpose, the genetic algorithm (GA), as the most widely used method in the literature, and three alternative algorithms (Boruta, Featurewiz, and ForwardSelector) combined with the forward stepwise selection technique were used. The construction of models utilized the multiple linear regression method, with subsequent scrutiny of fitting and predictive performance, reliability, and robustness through various statistical validation criteria. The models were ranked by the Multi-Criteria Decision Making procedure. Findings have revealed that the proposed methodology for descriptor selection outperforms GA, with Featurewiz showing a slight advantage over Boruta and ForwardSelector. These constructed models can serve as valuable tools for the quick and reliable prediction of NACs mutagenicity.

Keywords QSAR · Mutagenicity · Nitroaromatic compounds · Feature selection · Machine learning · Quantum chemical calculations

Introduction

Nitroaromatic compounds (NACs) are a class of aromatic molecules that contain at least one nitro group attached to the benzene ring. Due to their unique structure, NACs are utilized in a wide variety of applications. The most prominent applications are in the field of explosive materials, where molecules such as 2,4,6-trinitrotoluene, 1,3,5-trinitrobenzene, and hexanitrobenzene are used (Akhavan 2022). Besides that, for example, 2,4-dinitrotoluene and its derivatives are used in the synthesis of polyurethanes, dyes,

medicaments, and rubbers (Lent 2015), while 1-nitropyrene and 2-nitronaphthalene are found in fuels (Hayakawa 2016).

Due to electron-withdrawing nitro groups, NACs are persistent pollutants, i.e., resistant to oxidation and hydrolysis. Thus, their degradation is not efficacious enough, leading to accumulation in the environment (Zhang et al. 2018). The primary source of NACs in the environment is emission, mainly by industry and vehicles. However, a minor emission originates from nonanthropogenic sources, i.e., as a product of the metabolism of microorganisms (Tiwari et al. 2019). On the other side, the secondary source of NACs is the photochemical reactions of parent molecules with nitrogen oxide (Arey et al. 1986; Hayakawa 2016). Notably, NACs are not only omnipresent in the environment but are also detected in food sources (Deng et al. 2015; Tiwari et al. 2019).

Among other pollutants, several NACs have been singled out as priority pollutants by the United States Environmental Protection Agency, such as nitrobenzene, 2,4-dinitrotoluene, 2,6-dinitrotoluene, 2-nitrophenol, 4-nitrophenol, etc. (Tiwari et al. 2019). Additionally, the toxic effects of NACs and

Responsible Editor: Marcus Schulz

✉ Branislav Stankovic
branislav.stankovic@vin.bg.ac.rs

¹ Department for Nuclear and Plasma Physics, Vinča Institute of Nuclear Sciences –National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia

² Faculty of Physics, University of Belgrade, Belgrade, Serbia

their relation with PM_{2.5} particles, have raised great concern among scientists and environmental protection agencies (Loomis et al. 2013). NACs exhibit a myriad of toxic effects. They were reported to be mutagenic and carcinogenic, as well as, to cause gastrointestinal, neurological, and reproductive disorders, respiratory and skin irritation, allergic reactions, immunotoxicity, methemoglobinemia, endocrine system impairment, etc. (Hsu et al. 2007; Kovacic and Somanathan 2014). For instance, diesel exhaust, containing 6-nitrochrysene, 1-nitropyrene, and 1,6-dinitropyrene, has been identified as carcinogenic and mutagenic to mammals (Benbrahim-Tallaa et al. 2012). Additionally, 2,4,6-trinitrotoluene, 2,4-dinitrotoluene, and their metabolites have been linked to adverse reproductive effects in crickets and salamanders (Bazar et al. 2008; Karnjanapiboonwong et al. 2009).

Mutagenicity is often assessed by the Ames test with *Salmonella typhimurium* assays, which is renowned for its high sensitivity to base-pair substitution and frameshift mutagens (Hornberg et al. 2014). However, the Ames test has a few drawbacks: it requires significant time, costs, and manpower. On the other side, since millions of organic chemicals are used in industry, it is impossible to evaluate all of them through the Ames test experimentally. European Union Registration, Evaluation, Authorization, and Restriction of Chemicals legislation recommends the use of in silico methods to comply with Organization for Economic Co-Operation and Development (OECD) principles, ensuring the provision of reliable toxicity data (Gozalbes and de Julián-Ortiz 2018). Several in silico approaches have been developed to address challenges in safety evaluation and risk assessment, as well as drug and material discovery, with quantitative structure–activity relationship (QSAR) standing out (Hong 2023; Huang et al. 2021; Ambure et al. 2019; Halder and Cordeiro 2021; Halder et al. 2022). QSAR is often coupled with Monte Carlo (Toropova et al. 2023; Toropov et al. 2023) and quantum chemistry (Ostojić et al. 2014a, 2014b; Stanković et al. 2016a, 2016b) methods and is recommended for early detection of various toxic chemicals (Khan et al. 2019a, 2019b). Further, mutagenicity in *S. typhimurium* has demonstrated a close correlation with carcinogenicity in rodent and human bioassays (Mortelmans and Zeiger 2000), which consequently led to numerous QSAR studies on NACs toxicity in rats (Keshavarz and Akbarzadeh 2019; Daghighi et al. 2022).

Among various properties, the lipophilicity and saturation of aromatic compounds hold particular significance due to their role in enabling molecules to penetrate cellular membranes. Specifically, NACs promote the fluidization of the phospholipid bilayer of a cell membrane, facilitating their subsequent accumulation within the cell. This entry into the redox cycle allows NACs to induce oxidative stress, triggering the production of reactive radicals that bond to the

exocyclic amino group of the phosphodiester bonds, ultimately causing mutations (Yu et al. 2016). In addition to lipophilicity and saturation, properties related to electronegativity also play a crucial role in determining mutagenicity, as electronegative atoms have the potential to intercalate into the space between two adjacent base pairs of DNA, inducing structural changes.

Wang et al. (2005) modelled the *Salmonella typhimurium* TA98 strain mutagenicity of NACs. Besides other parameters, the model contains the energy of the lowest unoccupied molecular orbital (LUMO), highlighting the significance of nitroreduction in NACs mutagenicity. To examine NACs mutagenicity, Gramatica (2007) used topological molecular descriptors, which are known for their efficiency in predicting mutagenicity (Estrada 2002; Pérez-Garrido et al. 2010, 2014). Zhang et al. (2008) developed a QSAR model of *S. typhimurium* TA98 strain mutagenicity of nitronaphthalenes and methylnitronaphthalenes using common quantum chemical (QC) descriptors. They involved energies of both LUMO and HOMO (highest occupied molecular orbital) in the model. Ding et al. (2017) constructed QSAR models to predict the TA98 strain mutagenicity of NACs. Model consist of several descriptors: electrophilic index, hydrophobicity, the partial atomic charge on the carbon attached to the nitro group, the sum of molar refractivity of substituents at the *ortho* positions, indicators if a molecule has more than one nitro group and more than two fused rings, dipole moment, solvation free energy, edge adjacency index, as well as few descriptors related with charge distribution, size, and the methyl groups. Hao et al. (2019) investigate the mutagenicity of NACs towards *S. typhimurium* TA100 strain using a QC descriptor-based QSAR model. The best model includes energy of HOMO and four 2D descriptors related to lipophilicity and complexity of the structure. Jillella et al. (2020) constructed a QSAR model for the mutagenicity of nitro and amino aromatic compounds in a test with *Salmonella typhimurium* TA98. They categorized descriptors into subclasses related to unsaturation, size and properties of rings, hydrophobicity, and electronegativity. In the case of mutagenicity in the test without S9 activation, they showed that mutagenicity increases with the size of the molecules, unsaturation of the molecule, lipophilicity, and the number of N atoms, but decreases with branching and electron-richness of the molecule. Further, according to their findings, mutagenicity is lower for molecules with pentacyclo or hexacyclo rings, as well as in systems with more rings, whereas it is higher for molecules with higher ring complexity.

This study aims to develop transparent, interpretable, reproducible, and publicly available methodologies for deriving quantitative structure–activity relationship models. The focus is on enhancing feature selection techniques, which literature has not explored extensively. The developed methodologies were applied to estimate the mutagenicity of

NACs against the *Salmonella typhimurium* TA100 strain. Further, as aforementioned, many QC descriptors were used in modelling NACs mutagenicity. However, their calculation is significantly more time-consuming compared to the molecular descriptors. This is even more pronounced when they are calculated with greater precision. Consequently, this study aims to evaluate whether QC descriptors can enhance the prediction of NACs mutagenicity in the *Salmonella typhimurium* TA98 strain and if they can be computed efficiently while maintaining good precision.

Materials and methods

Dataset

The investigation was focused on a set of 48 nitroaromatic compounds (Fig. 1A in Appendix A of the Supplementary Information (SI)) sourced from Benigni's report for the OECD (2004). Mutagenicity assessments were conducted using the Ames test with *Salmonella typhimurium* TA100 strain, performed without S9 activation. Mutagenicity values were given as the logarithm of the number of revertant per nanomole (rev/nmol). It is noteworthy that the OECD validation principle No. 1, emphasizing "a defined endpoint", is satisfied by this dataset (OECD 2007).

Splitting of data in the training and the test set followed the stratified sampling approach outlined in the paper of Hao et al. (2019). The molecules were sorted based on their mutagenic activity, organized into groups of four, and one molecule was selected from each group for inclusion in the test set. The remaining NACs were assigned to the training set. Additionally, both the least and the most mutagenic compounds were intentionally included in the training set to ensure that the molecules selected for the test set fall within the applicability domain of the subsequently developed QSAR model. Details regarding the compound names, CAS numbers, corresponding log TA100 values, and the quantum chemical descriptors selected by the ML models developed in this study have been provided in Appendix A of the SI.

Molecular descriptors

In this study, both 2D molecular and QC descriptors were employed to comprehensively characterize the NACs. Molecular descriptors were calculated by Mordred (Moriwaki et al. 2018) and RDKit (Landrum et al. 2023) packages. Geometry optimization of selected molecules was performed by the B3LYP method with the 6–31 + G(d,p) basis set. Vibrational frequency analysis was conducted to ensure the identification of minima on potential surfaces. The values of descriptors were then refined by single-point

calculations performed on the optimized structures using the 6–311 + +G(2d,2p) basis set with the same functional.

From the output files, electronic energies, dipole moments, and polarizability tensor values were extracted. Additionally, electron energies of neutral molecules, as well as energies of cations and anions of NACs, were calculated on the same optimized geometries. These parameters were then utilized for calculating various commonly used energy-related descriptors, including ionization energy, electron affinity, hardness, softness, chemical potential, and electrophilic index, as defined in the literature (Ostojčić et al. 2014a, 2014b; Stanković et al. 2016a). In addition to these conventional descriptors, the values of modified chemical potential and electrophilic index proposed by Gazquez et al. (2007) were incorporated. This choice was influenced by recent evidence suggesting their utility in classifying the carcinogenic activity of activated metabolites derived from nitroaromatics (Halabi et al. 2022).

In addition to the above-mentioned descriptors, the aim was to consider QC descriptors calculated using different, less time-consuming yet sufficiently accurate methods. Therefore, polarizability obtained by a ML method based on symmetry-adapted Gaussian process regression and the smooth overlap of atomic positions (Wilkins et al. 2019) was included.

Selection of descriptors and modelling

The schematic representation of model development process and in detail scheme of feature selection process has been given in Fig. 1. Before the selection process, descriptors with missing values for at least one molecule or containing non-numerical values were discarded. Following this step, the dataset was split into training and test sets and normalized so that each descriptor was in the range of 0–1. As the initial step of the feature selection, both constant and nearly constant (i.e. those with a standard deviation in the training set less than 0.05) descriptors were omitted, as suggested in the literature (Tropsha et al. 2003). For the selection process, the genetic algorithm (GA) was employed along with three algorithms that, to our knowledge, have not been previously utilized in the QSAR literature, namely Boruta, Featurewiz, and the ForwardSelector algorithm. For more detailed information and references on these algorithms, please refer to section “Software”.

In the case of GA, a two-step procedure was implemented as in Hao et al. (2019). First, a set of descriptors was narrowed to ensure that no pair of descriptors had a pair-wise correlation higher than 0.95. More precisely, a descriptor with a higher sum of correlation coefficients with the remaining descriptors was omitted. In the second step, a GA was employed for the final selection of up to 7 descriptors. The algorithm parameters were set as follows: population

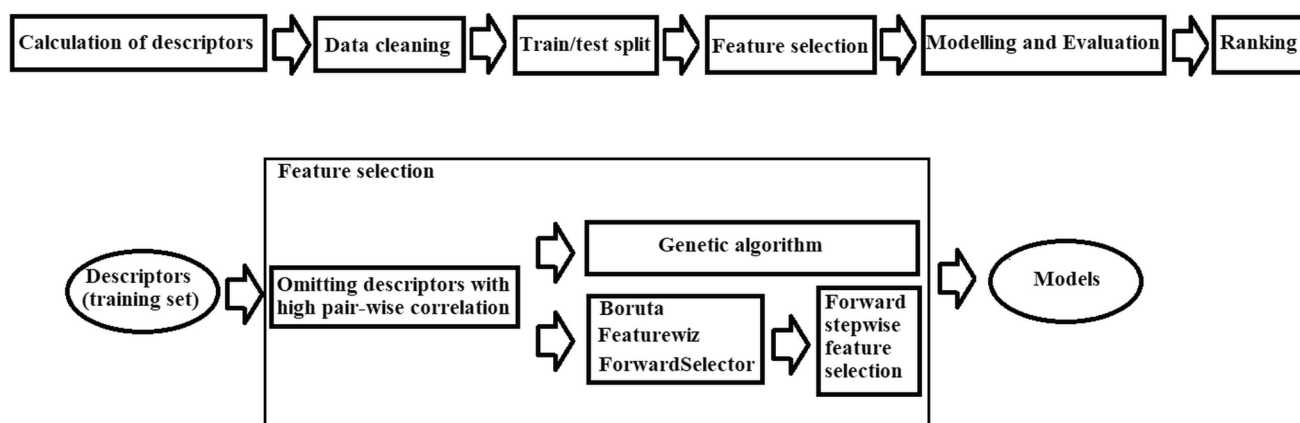


Fig. 1 Schematic representation of model development process and in detail scheme of feature selection process

size as 100, mutation rate as 20, and number of generations as 500. Coefficients of determination evaluated in leave-one-out cross-validation was used as the fitness function.

For Boruta, Featurewiz, and ForwardSelector algorithms, different threshold for acceptable pair-wise correlation between descriptors (i.e. 0.80, 0.85, 0.90, 0.95, and 0.99) were applied. As a subsequent step of descriptor selection, forward stepwise feature selection was applied to create models, commencing from each of the selected descriptors. During this process, new descriptors were gradually incorporated into the models to optimize the correlation coefficient in leave-one-out cross-validation.

Modelling was utilized using the multiple linear regression method incorporating up to 7 descriptors. The parameters of the model were estimated using the ordinary least square procedure, aligning with OECD validation principle No.2, "model building with an unambiguous algorithm".

Methods for inspection of chemical space

The OECD validation principle No.3 emphasizes that the model should have "a defined domain of applicability" (OECD 2007). This study utilized the leverage approach (please see Appendix B of the SI) to define the applicability domain (AD) (Gramatica 2007; Gadaleta et al. 2016). Alongside the leverage approach, for further insight into chemical space, a plot of the second vs. the first principal component and a plot of molecule weight (MW) against the Wildman-Crippen partition coefficient (SlogP) was employed.

Evaluation of models

Model evaluation with "appropriate measures of goodness-of-fit, robustness and predictivity" aligns with the OECD validation principle No.4. A range of commonly used

statistics was employed, as documented in the literature (Ding et al. 2017; Hao et al. 2019, 2020). These included measures for goodness-of-fit and robustness, such as the root mean square error in both the training set, $RMSE_{tr}$, and cross-validation, $RMSE_{cv}$, coefficients of determination in the training set, R^2 , as well as in leave-one-out, Q_{LOO}^2 , and leave-many-out, Q_{LMO}^2 , cross-validation. Additionally, the adjusted coefficients of determination, R_{adj}^2 , concordance correlation coefficient in the training set CCC_{tr} , and internal validation, CCC_{cv} , along with F -value were assessed. Predictive performance was estimated using root mean square error in the test set, $RMSE_{ext}$, mean absolute error in the test set, MAE_{ext} , alongside three metrics derived from coefficients of determination: Q_{F1}^2 , Q_{F2}^2 , and Q_{F3}^2 . Furthermore, correlation coefficient in test set, CCC_{ext} , slopes of regression lines of predicted vs. experimental values, k , and experimental vs. predicted values, k' , (both through the origin), along with parameters introduced by Roy et al. (2012): $\langle r_m^2 \rangle$ and Δr_m^2 , and Golbraikh and Tropsha (2002): R_{ext}^2 , $(R_{ext}^2 - R_0'^2)/R_{ext}^2$, and $(R_{ext}^2 - R_0'^2)/R_{ext}^2$ were consider. Additional details on all statistics can be found in Appendix B of the SI.

Following the approach outlined by Hao et al. (2019), an acceptable model should satisfy the following criteria: 1) $Q_{LOO}^2 > 0.70$, 2) $R_{ext}^2 > 0.5$, 3) $|R_0^2 - R_0'^2| < 0.30$, and 4) $(R_{ext}^2 - R_0'^2)/R_{ext}^2 < 0.10$ and $0.85 < k < 1.15$ or $(R_{ext}^2 - R_0'^2)/R_{ext}^2 < 0.10$ and $0.85 < k' < 1.15$. However, aim was to ensure that the model also meets more stringent criteria as suggested in the literature (Ding et al. 2017; Hao et al. 2020):

5) $R^2, Q_{Fn}^2 > 0.70$, 6) $|R^2 - Q_{LOO}^2| < 0.10$, 7) both criteria in 4) should be fulfilled, 8) $r_m^2 > 0.65$, 9) $\Delta r_m^2 < 0.20$ and 10) $CCC_{ext} \geq 0.85$.

To mitigate multicollinearity and chance correlation, a three-fold selection process was employed. Initially, the QUIK (Q Under Influence of K) rule proposed by Todeschini

et al. (2004) was applied. Following the recommendation of the literature (Hao et al. 2019, 2020), only models exhibiting a difference of more than 0.05 between two K indexes were retained for further assessment. Subsequently, the Variance Inflation Factor (VIF) was calculated, and models where each descriptor had a VIF value lower than 5 (Singh et al. 2008) advanced to the final test, in which Y-scrambling was applied to ensure the robustness of the chosen models. In this step, the target variable was reordered, and the coefficients of determination for the training set, R_{Yscr}^2 , and leave-one-out cross-validation, Q_{Yscr}^2 , were calculated. Models with significantly low values of R_{Yscr}^2 and Q_{Yscr}^2 were accepted.

Selection of the best model

The process of selecting the optimal model involved the application of the Multi-Criteria Decision Making (MCDM) procedure, more precisely the weighted sum model (Pesode et al. 2023). By this approach, all the measures of goodness-of-fit, robustness, and predictivity from the previous section were normalized and summed to rank the various models. In addition to the MCDM value, the number of descriptors was also considered a factor in determining the most suitable model. This dual consideration ensures a comprehensive assessment, incorporating both the model's overall performance according to MCDM and the simplicity reflected in the number of descriptors utilized.

Software

Modelling was performed in Python programming language (version 3.10.12). For that purpose, several standard libraries were used, such as NumPy (version 1.25.2), pandas (version 1.5.3), sklearn (version 1.2.2), statsmodels (version 0.14.1). Visualization of results was done using seaborn (version 0.13.1).

Quantum-chemical descriptors were calculated using the Gaussian 16 package (Frisch et al. 2016). Molecular descriptors were derived by Mordred (Moriwaki et al. 2018) and RDKit (Landrum et al. 2023) packages. Particularly, Mordred version 1.2.0 (available at <https://github.com/mordred-descriptor/mordred>) and RDKit version 2023.9.5 (available at <https://github.com/rdkit/rdkit>) was applied.

Feature selection using GA, was performed by genetic_selection module version 0.6.0 (available at <https://github.com/manuel-calzolari/sklearn-genetic>), as well as in the QSARINS software version 2.2.2 (Gramatica et al. 2013; Gramatica 2014), and the best model was selected. Besides GA, for a feature selection, three algorithms were used: Featurewiz (version 0.5.7, available at <https://github.com/AutoViML/featurewiz>), Boruta (version 0.3 available at https://github.com/scikit-learn-contrib/boruta_py; Kursu and Rudnicki 2010), ForwardSelector within step-select package

(version 0.1.1, available at <https://github.com/chris-santiago/steps>).

In the case of the Boruta algorithm, RandomForestRegressor was employed as an estimator. Parameter max_iter (the number of maximum iterations to perform) was set to 20, while for the rest of the parameters, the default values were taken. For the RandomForestRegressor, the parameter max_depth (the maximum depth of the tree) was set to 5. All other parameters retained default values. Given the stochastic nature of the Boruta algorithm, relatively low values for these parameters were chosen, and the algorithm was run multiple times to thoroughly explore the parameter space and ensure the reproducibility of the method. Specifically, all descriptors from 500 runs were considered, as an additional 100 runs did not yield any new descriptors. In the case of Featurewiz and ForwardSelector algorithms, default adjustments were applied.

Results and discussion

Chemical space distribution

The distribution of chemical space plays a pivotal role in assessing the predictive capability of the developed model. A commonly employed method to visualize chemical space is plot of dependence between MW and SlogP for the training set and test set (Fan et al. 2018; Hao et al. 2019). As illustrated in Fig. 2, the molecules within the datasets exhibit a chemically diverse profile. Additionally, it can be seen that the training set and the test set share similar chemical space, indicating consistency and generalizability in the model's applicability.

To better understand the potential applicability of the models derived in this study, the dataset and its division

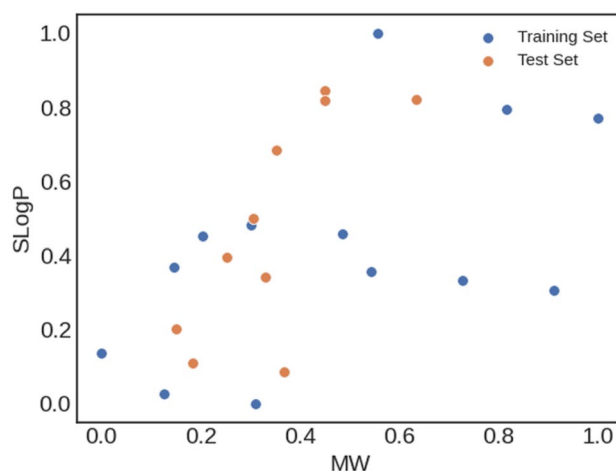


Fig. 2 The plot of SlogP vs MW for training and test set

into the training and test sets will be briefly discussed. Molecules within dataset contain a different number of nitro groups (22 with one, 15 with two, 9 with three, and 2 with four groups) and benzene rings (17 with one, 16 with two, 7 with three, and 8 with four rings). Notably, 3 NACs feature a nitrogen atom in the pyrrole ring, while 2 others have it in the pyridine ring. Furthermore, 15 NACs include one methyl group, and 10 have cyclopentane rings (3 of which contain a cyclopentanone moiety).

The log TA100 values span in a range from -2.10 to 4.74. The broad range of mutagenic activities makes the dataset suitable for QSAR analysis, while the diverse moieties in NACs provide a robust foundation for model development, potentially applicable to other NACs containing only C, H, N, and O atoms. However, it's crucial to note that both molecules with the pyridine ring and both nitroanthracenes are placed in the test set. Since the models developed here will not be trained on molecules with such geometries, predicting the mutagenic activities of these molecules might pose a challenge for certain models, a point that will be discussed later in this study.

Model selection and evaluation

To model NACs mutagenicity, 1821 2D molecular descriptors (1613 Mordred and 208 RDKit) and 10 QC descriptors were calculated. Calculating 3D descriptors is time-consuming and may require rigorous calculations to obtain reliable values. Therefore, this study included only those descriptors that previous studies indicated to be particularly significant in modelling NACs mutagenicity, as well as that can be calculated with satisfactory precision (Ostojić et al. 2014a, 2014b; Stanković et al. 2016a; Hao et al. 2019).

During the data cleaning step, which involved omitting descriptors with missing or non-numerical values for at least one molecule, the number of descriptors was reduced to 1135. Following the first step of feature selection, which involved removing constant and nearly constant descriptors, the count further decreased to 1124. As detailed in section “[Selection of descriptors and modelling](#)”, at this point descriptor selection branches in four different paths.

Following an established rule that the ratio between the size of the training set and the number of descriptors should be higher or equal to 5, modelling was performed using the multiple linear regression method incorporating up to 7 descriptors (Gramatica et al. 2013). To ensure fair comparisons with results from the literature (Hao et al. 2019), model development was also conducted in the QSARINS software (version 2.2.2) (Gramatica et al. 2013; Gramatica 2014). The modelling approach in QSARINS mirrored that of Hao et al., though with different descriptors. This comparative analysis allows for a comprehensive evaluation of our results in relation to prior findings.

In the feature selection step employing GA, 100 models with up to 7 descriptors were selected for further evaluation. When QSARINS software is used, less than 50 percent of models pass the check for low collinearity based on the QUIK rule. Notably, approximately 10 percent of the top-best models were rejected in this process. Further, after the step that filters out models not meeting the criteria for metrics defined in “2.5 Evaluation of models”, only around one-fourth of the initially selected models remained. Additionally, obtained models often suffer from reduced applicability domain. The situation is exacerbated when employing the GeneticSelectionCV algorithm, which produced around 10 applicable models, while issues with the applicability domain persisted. QSARINS software (on average) generates better models, probably due to the better choice of the initial population. Given the challenges encountered with GA, coupled with the stochastic nature of the algorithm, the idea of finding alternative feature selection methodologies that are more time-efficient in searching the descriptor space and yet yield well-designed models comes to mind.

Given the lack of consensus in the literature regarding the ideal threshold for intercorrelation between descriptors, this study aimed to investigate the effect of the threshold for acceptable pair-wise correlation between descriptors. Table 1 summarizes the number of descriptors selected by each algorithm at different intercorrelation thresholds. Featurewiz and ForwardSelector consistently yield the same set of descriptors. However, in the case of the Boruta algorithm, the number of descriptors can vary between runs. Thus, both the average number of descriptors and the overall number of unique descriptors (i.e., those selected in 500 runs) has been presented.

The Featurewiz algorithm stands out by selecting the fewest descriptors. A lower starting number of descriptors can substantially reduce the time needed for subsequent steps. This is particularly significant when model performances are similar. In this particular case, the following steps (i.e., forward stepwise feature selection and model evaluation) are the most time-consuming part. On the other hand, they take only a few seconds per selected descriptor, which hardly limits the applicability of the algorithm for such small datasets. Nevertheless, in future research, significant effort will be

Table 1 The number of descriptors which can be selected by different feature selection algorithms and intercorrelation thresholds between descriptors. For the Boruta algorithm numbers in parentheses shows the average number of descriptors in a single run

threshold	0.80	0.85	0.90	0.95	0.99
algorithm					
Featurewiz	11	15	19	19	19
ForwardSelector	17	30	35	35	35
Boruta	15 (11)	22 (16)	28(20)	43(27)	52 (27)

made to ensure the methodology has good scalability with larger dataset sizes.

The decrease in the intercorrelation threshold significantly affects the number of descriptors only in the case of the Boruta algorithm. Specifically, with the other two algorithms, changes in the number of descriptors are noticeable primarily around the threshold value of 0.85. Additionally, for threshold values of 0.90, 0.95, and 0.99, the differences in the performances of the selected models are negligible. On the contrary, further decreasing the threshold to 0.80 has a significant impact on the performances of the ForwardSelector and Boruta algorithms, particularly affecting parameters related to accuracy. Therefore, for simplicity, results for thresholds 0.99 and 0.85 have been analysed.

Nine models, labelled as model 1–9 (Table 2 and Appendix C of the SI), have been presented in this study. Models 1–5 were derived using a threshold of 0.99 and different algorithms. As one of the aims of this research is to explore whether QC descriptors can enhance the modelling of NACs mutagenicity, models 1, 2, and 3 were selected as the best models that do not contain QC descriptors obtained in the step with Featurewiz, Boruta, and ForwardSelector algorithm, respectively. Since the Boruta algorithm didn't select any QC descriptors, only two models with QC descriptors have been presented here: model 4 (Featurewiz algorithm) and model 5 (ForwardSelector algorithm). Additionally, to investigate the extent to which the intercorrelation threshold affects modelling, the three best models selected with an intercorrelation threshold of 0.85 were considered. Model 6 was obtained starting from the Featurewiz algorithm, model 7 from the ForwardSelector algorithm, and model 8 from the Boruta algorithm. Finally, model 9 represents the best model selected using GA, as the most common approach in the literature.

Each QSAR equation has been provided in two forms expressing the relation between mutagenicity and scaled (Eq. C.1a–C.9a in Appendix C of SI) and unscaled (Eq. C.1b–C.9b in Appendix C of SI) descriptors. In the former, the importance of each descriptor can be estimated based on the corresponding coefficient—the higher the coefficient, the greater the importance. In the latter, predicting the mutagenicity of compounds not present in the dataset becomes more straightforward. Coefficients in Eq. C.1a–C.9a vary within two orders of magnitude, indicating the significance of all descriptors.

Metrics for nine models are presented in Tables 2 and 3. Correlation coefficients R^2 and R^2_{ext} of the QSTR models fall within the ranges of 0.845–0.957 and 0.783–0.886, respectively. Additionally, Q^2_{LOO} and Q^2_{Fn} take values within intervals of 0.783–0.921, and 0.761–0.876, respectively. The comparability of parameters from internal and external validation to R^2 indicates the reliability and stability of the models. Apart from good precision, all models demonstrate more

Table 2 Internal fitting parameters of the models

No	Descriptors	R^2	adj R^2	RMSE _{tr}	CCC _{tr}	F	Q^2_{LOO}	RMSE _{ev}	CCC _{ev}	Q^2_{LMO}	R^2_{Yscr}	Q^2_{Yscr}
1	AAAT51v, VE2_Dzse, GATS1are, AT58dv, GATS4i, CIC4	0.926	0.910	0.473	0.961	60	0.879	0.604	0.938	0.864	0.155	-0.413
2	Vabc, FpDensityMorgan1, GATS2p, EState_VSA3, GATS1pe	0.929	0.918	0.461	0.963	79	0.896	0.558	0.947	0.889	0.141	-0.296
3	Xc-3d, FpDensityMorgan1, C3SP2, fr_bicyclic, CIC4, NumAliphaticCarbocycles	0.951	0.941	0.384	0.975	94	0.921	0.487	0.960	0.903	0.166	-0.344
4	EA, FpDensityMorgan1, C3SP2, AT58dv	0.933	0.924	0.450	0.965	108	0.910	0.522	0.954	0.910	0.090	-0.257
5	AvgPolar-QM, VE2_Dzse, GATS1se, EState_VSA3, MinEStateIndex, GATS4i,	0.925	0.909	0.475	0.961	60	0.885	0.588	0.941	0.867	0.180	-0.331
6	MolMR, ipc, GATS2v, GATS1pe, Kappa3, FpDensityMorgan1	0.943	0.931	0.415	0.971	80	0.907	0.528	0.953	0.878	0.173	-0.349
7	JG16, AAAT55v, GATS1se, GATS2v, GATS6s	0.880	0.860	0.601	0.936	44	0.838	0.697	0.915	0.790	0.151	-0.264
8	AAAT55i, SpAbs_Dzp, CIC2, GATS1c	0.845	0.826	0.682	0.916	42	0.783	0.809	0.887	0.780	0.105	-0.262
9	ATSC0m, ATSC2p, ZMIC3, FpDensityMorgan1, NumRotatableBonds	0.957	0.950	0.358	0.978	135	0.943	0.414	0.971	0.919	0.136	-0.290

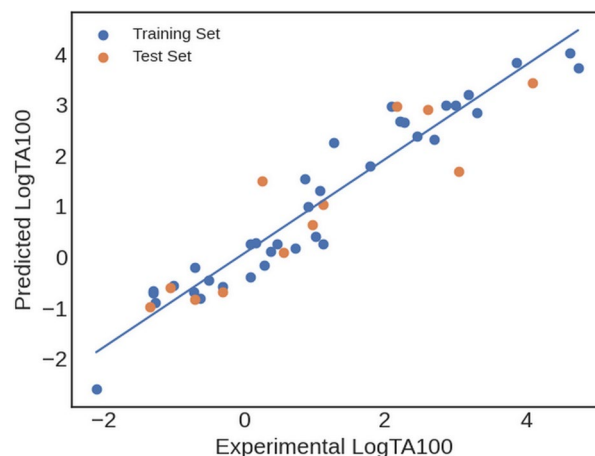
Table 3 External validation parameters of the models

No	R_{ext}^2	RMSE_{ext}	Q_{F1}^2	Q_{F2}^2	Q_{F3}^2	CCC_{ext}	MAE_{ext}	r_m^2	Δr_m^2	k	k'	$(R_{\text{ext}}^2 - R_0^2)/R_{\text{ext}}^2$	$(R_{\text{ext}}^2 - R_0^2)/R_{\text{ext}}^2$
1	0.835	0.670	0.836	0.835	0.851	0.911	0.544	0.798	0.058	0.999	0.877	0.000	0.008
2	0.819	0.757	0.790	0.790	0.810	0.902	0.592	0.780	0.001	0.867	0.995	0.003	0.003
3	0.785	0.787	0.774	0.773	0.794	0.884	0.608	0.737	0.095	0.907	0.925	0.000	0.019
4	0.886	0.611	0.863	0.863	0.876	0.935	0.444	0.838	0.056	0.881	1.037	0.001	0.008
5	0.849	0.666	0.837	0.837	0.852	0.921	0.528	0.816	0.058	0.931	0.948	0.006	0.000
6	0.824	0.746	0.796	0.795	0.815	0.904	0.604	0.785	0.014	0.868	0.998	0.002	0.004
7	0.813	0.732	0.804	0.803	0.822	0.900	0.613	0.774	0.078	0.968	0.882	0.011	0.000
8	0.838	0.667	0.837	0.837	0.852	0.913	0.525	0.803	0.002	1.004	0.874	0.002	0.002
9	0.783	0.807	0.762	0.761	0.784	0.883	0.675	0.739	0.086	0.916	0.904	0.016	0.000

than satisfactory accuracy. CCC_{ext} surpasses 0.883 (goes up to 0.935) and is close to the CCC_{tr} (0.916–0.978) and CCC_{cv} (0.887–0.971), indicating a consistent distribution of the target variable between the training and test sets. Values of r_m^2 fall within the intervals of 0.737–0.838, while $\Delta r_m^2 < 0.1$, which is a particularly good result. Furthermore, values of k and k' are within ranges of 0.867–1.04 and 0.874–1.037, while $\frac{(R_{\text{ext}}^2 - R_0^2)}{R_{\text{ext}}^2}$ and $\frac{(R_{\text{ext}}^2 - R_0^2)}{R_{\text{ext}}^2}$ are of the order of magnitude 10^{-7} – 10^{-3} . Thus, even stricter criteria from the literature are satisfied.

All models listed in Table 2 adhere to the QUIK rule and none of the descriptors within them exhibit a VIF exceeding 5, indicating low multicollinearity. Low values of R_{Yscr}^2 and Q_{Yscr}^2 signify the absence of chance correlation. Thus, it can be said that the presented models are suitable for predicting *Salmonella typhimurium* TA100 mutagenicity of NACs. The model 1 was selected as the best model among models 1–9. A scatter plot depicting the experimental vs. predicted mutagenicity values for the training and test sets (Fig. 3) confirms the excellent fitness and predictability. Notably, despite the presence of both NACs with a pyridine ring and nitroanthracenes in the test set, NACs with a pyridine exhibit deviations comparable to other classes of molecules, while nitroanthracenes represent the only NACs for which the absolute difference between experimental and predicted values exceeds one log unit.

Among all models, model 9 exhibits the best performance in internal fitting but turns out to be the worst in external validation, indicating a high level of overfitting. Consequently, the MCDM method rates this model as the least favourable. During the comparison of various models by the MCDM method, an interesting observation emerged — the designation of the best model often hinges on the specific set of models being compared. For example, model 4 stands out as the best when considering all models selected by the Featurewiz algorithm, with model 1 following closely as the second best. However, when comparing the most well-performing models selected by all algorithms, model 1 emerges

**Fig. 3** Scatter plot of predicted versus experimental mutagenicity for the molecules in the training and test set for the model 1

as the top choice. Moreover, the selection of the best models is contingent upon the metrics considered, making it challenging to unequivocally determine the best model.

Of all the models, model 1 has a value of k closest to 1, while model 4 has a relatively poor value for this statistics, although the majority of other parameters are comparable and in favor of model 4. Since model 9 shows better performance in internal fitting than all models selected by the Featurewiz algorithm, comparing models 1 and 4 in the same set with model 9 reduces the difference in internal fitting scores between models 1 and 4, making model 1 better ranked. It should also be noted that another MCDM model might rank model 4 as the best. Additionally, model 5 is well-ranked, as its measures of goodness-of-fit, robustness, and predictivity are relatively high compared to other models. Thus, it is evident that QC descriptors can be a significant component of models for predicting NACs mutagenicity.

Among the three alternative algorithms, Featurewiz appears to be the most promising. Not only that models built by Featurewiz are slightly better rated than other models, but the entire procedure following its use is less

time-consuming, offering excellent performances even at a lower intercorrelation threshold. Formally, while model 6 is ranked as the fourth-best model, its notable feature is the ability to be obtained with a threshold as low as 0.8. This has the potential to significantly accelerate the model development process, especially for larger datasets.

In the existing literature, only a limited number of models have been developed on the dataset investigated here. Gramatica (2007) utilizing two topological molecular descriptors, achieving correlation coefficient ranges of 0.768–0.885 in the training set, 0.705–0.861 in LOO cross-validation, and 0.504–0.841 in external validation. The models presented in this study demonstrate better performance, which is expected given the higher limit set for the maximal number of descriptors. A more recent study by Hao et al. (2019) derived models showing great fitting performances. However, it was found that the model developed by Hao et al. faced multicollinearity issues, as it couldn't pass the QUIK rule and exhibited high VIF values. This prevented a direct comparison with the models developed in this study.

Calculating QC descriptors by ML methods

As said, the cost associated with the calculation of QC descriptors by precise ab initio and DFT methods limits their widespread use in QSAR studies. Consequently, most studies rely on semi-empirical or HF calculations, predominantly utilizing HOMO and LUMO energies. Although Koopmans' theorem suggests that HOMO/LUMO energy should reflect the negative values of IP/EA, the unreliability in predicting EA arises from significant orbital relaxation effects on LUMO eigenvalues (Zhang and Musgrave 2007). Therefore, there is a pressing need to develop an efficient yet accurate methodology for calculating quantum chemical descriptors. In recent years, various attempts have been made to develop ML procedures for the rapid and accurate computation of these descriptors.

As far as our knowledge extends, no model for predicting EA has achieved both high accuracy and a large AD. In contrast, Wilkins et al. (2019) introduced a method based on symmetry-adapted Gaussian process regression and the smooth overlap of atomic position descriptors for predicting polarizability. Their model demonstrated superior accuracy compared to moderate-precision QC methods and exhibited scalability with an increase in molecule size. As can be seen from Fig. 4, there is an excellent correlation between values of α obtained by the DFT and the ML approach. Moreover, substituting the ML-predicted values into model 5 results in negligible changes in metrics. Therefore, utilizing ML methods for calculating QC descriptors, given its comparable computational efficiency with 2D molecular descriptors, emerges as a valuable approach to enhance QSAR studies such as the one presented here.

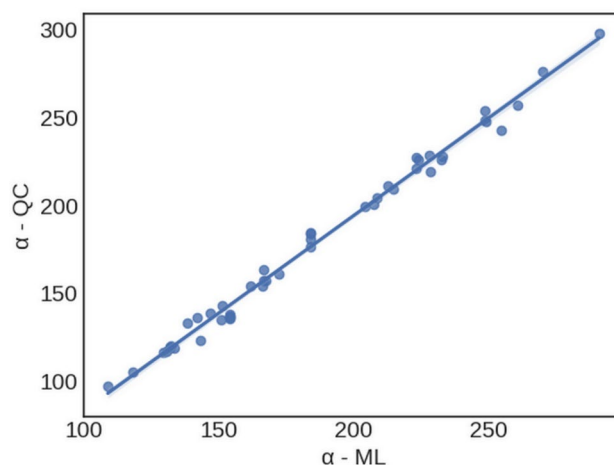


Fig. 4 Scatter plot of average polarizability calculated with QC method (α -QC) and ML (α -ML) method of Wilkins et al. (Wilkins et al. 2019)

Mechanistic interpretation

OECD validation principle No. 5 requires “a mechanistic interpretation, if possible” (OECD 2007). With this in mind, the discussion will first address each of the descriptors selected by the best model according to MCDM.

The best model identified in this study comprises six descriptors (AATS1v, GATS1are, VE2_{Dzse}, ATS8dv, CIC4, and GATS4i). AATS1v represents the averaged Moreau-Broto autocorrelation of lag 1 weighted by van der Waals volume. This descriptor shows the diversity of atoms within bonds by volume. It increases with the size and number of heteroatoms in the structure (and thus, consequently, with the number of nitro groups). As a lag 1 autocorrelation descriptor, AATS1v considers chain lengthening and branching, but also charge transfer and Coulomb interactions, as it is related to the distribution of heavier atoms. Given the positive correlation between the size, the number of NO₂ groups, and the number of heteroatoms in rings with NACs mutagenicity (Yu et al. 2016; Hao et al. 2019; Jillessa et al. 2020), the substantial role of a descriptor such as AATS1v in the mutagenicity model is expected. However, AATS1v yields the same value for different isomers, necessitating additional descriptors to differentiate between them. In other models, two autocorrelation descriptors weighted by van der Waals volume were chosen: GATS2v (the Geary coefficient of lag 2) and AATS5v (the averaged Moreau-Broto coefficient of lag 5). Additionally, five descriptors related to the size and shape of molecules were selected: FpDensityMorgan1 (Morgan fingerprint density of radius 1), Vabc (van der Waals volume of the molecule), ATSC0m (centered Moreau-Broto autocorrelation of lag 0 weighted by mass), Xc-3d (3-ordered Chi cluster weighted by sigma electrons), and Kappa3 (3rd Kier and Hall kappa molecular

shape index). AATS5v, as Moreau-Broto coefficient of lag 5, exhibits lower values when two nitro groups are in the *meta* position (topological distance of 5 between O and N from different groups) compared to *ortho* or *para* positions. Moreover, all ATSC descriptors depend on molecule size, accounting for dispersion interaction strength. The importance of considering the size and shape of the molecule in mutagenicity assessment is underscored by the fact that several descriptors, which will be elaborated on below, are dependent on these two properties.

Two descriptors related to the electronegativity were selected GATS1are (Geary coefficient of lag 1 weighted by Allred-Rochow electronegativity) and VE2_Dzse (average coefficient of the last eigenvector from the Barysz matrix weighted by Sanderson electronegativity). Differences in electronegativity between two adjacent atoms can indicate higher hydrophilicity, potentially leading to lower mutagenicity (Yu et al. 2016; Jillella et al. 2020). These descriptors are also influenced by the size of the molecule, but decrease with it. Therefore, a negative correlation is expected, as indicated by Eq. C.1. Additionally, VE2_Dzse is lower for the less substituted molecules and molecules with fewer hetero atoms in the rings and it can differentiate isomers, making it a valuable descriptor ($R^2=0.76$). Apparently, the distribution of electronegative atoms within a molecule might be particularly important for determining mutagenicity, as similar descriptors are selected in other models. Notably, GATS1pe (Geary coefficient of lag 1 weighted by Pauling electronegativity) was selected in models 2 and 6, while GATS1se (Geary coefficient of lag 1 weighted by Sanderson electronegativity) is part of model 5.

ATS8dv, the Moreau-Broto autocorrelation of lag 8 weighted by valence electrons, is sensitive to steric hindrance and long-range rearrangements, thus accounting for patterns in branching and dispersion forces. In addition to being part of model 1, it is also selected in model 4. This descriptor exhibits a negative coefficient in corresponding equations, aligning with the established knowledge that branching and steric hindrance decrease the toxicity of NACs (Kuz'min et al. 2008; Mondal et al. 2020; Jillella et al. 2020). Moreover, ATS8dv, as a lag 8 descriptor, has a value equal to 0 for small molecules where there are not 2 atoms at a topological distance of 8, thus considering molecule size to some extent.

The 4-ordered complementary information content (CIC4) is a descriptor that describes connectivity in the molecule, considering both its size and atomic composition. For compounds with complex structures and diverse atom types, CIC4 has a small value, while for simple molecules with uniform atom types, the value is large. Accordingly, coefficient in Eq. C.1 is negative. This descriptor is related to intermolecular interactions and as such it has been selected in the research of rodent carcinogenicity (Li et al. 2022).

Additionally, other information content descriptors such as 2-ordered complementary information content (CIC2), 3-ordered Z-modified information content (ZMIC3), and information content of the coefficients of the characteristic polynomial of the adjacency matrix (I_{pc}) were chosen in different models.

GATS4i, Geary coefficient of lag 4 weighted by ionization potential, tends to have higher values when the same substituents are in the *meta* position. Moreover, this descriptor is capable of estimating interactions like Coulombic, dipolar, and hydrogen bonding interactions. Similarly, the averaged Moreau-Broto autocorrelation of lag 5 weighted by ionization potential (AATS5i) was chosen in model 8. As indicated in the literature (Kuz'min et al. 2008; Gooch et al. 2017; Mondal et al. 2020; Hao et al. 2020) the positioning of substituents significantly influences the NACs toxicity, a fact further highlighted in studies involving small datasets consisting solely of different isomers of molecules (Ostojić et al. 2014a; Stanković et al. 2016a). While GATS4i alone may not exhibit a significant correlation with mutagenicity in the current dataset, by examining derivatives of pyrene, it becomes apparent that molecules with nitro groups in the *meta* position tend to have lower mutagenicity. To further investigate the significance of GATS4i as a descriptor of mutagenicity, a model consisting of all descriptors from model 1 except GATS4i was examined. This model performed worse, particularly in terms of accuracy-related parameters. For instance, the value of Δr_m^2 increased from 0.058 to 0.098. Moreover, the overall MCDM value decreases significantly, resulting in this model being one of the lower-ranking.

The similarity in MCDM values among all models suggests that the determination of the best model is contingent on the chosen set of metrics used for evaluation and the MCDM model. Consequently, it implies that several of these models or descriptors might offer improved descriptions of mutagenicity when applied to a different, potentially larger, dataset. With this consideration, the mechanistic interpretation of these models will be explored.

Descriptors related to the intrinsic state consider factors like chain length, branching, heteroatoms, and unsaturation. GATS6s (Geary coefficient of lag 6 weighted by intrinsic state) is a part of model 7 and accounts for steric hindrance and long-range rearrangements, leading to a negative coefficient in Eq. C.7. MinEStateIndex, the minimal value of electrotopological state (EState), is included in model 5. The EState value, affected by intrinsic state and topology, estimates the deficiency of pi and lone pair electrons. Thus, it can serve as a measure of affinity toward nucleophilic attack, justifying the positive coefficient in Eq. C.5. EState_VSA3 descriptor (sum of van der Waals surface areas with EState in the range of $4.69 \leq x < 9.17$) was selected in models 2 and 5. EState_VSA descriptors can be related to specific

electrostatic interactions of backbone atoms and the presence of reactive sites, explaining positive coefficients in the respective equations.

Charge distribution within the molecule is captured by GATS1c (Geary coefficient of lag 1 weighted by Gasteiger charge) and JGI6 (6-ordered mean topological charge). Similar to GATS1are, GATS1c exhibits a negative coefficient (Eq. C.8), as an increase in its value can be associated with higher hydrophilicity of the molecule. JGI6 is often referred to as a measure of dipole momentum and depends on the size and shape of the molecule. The descriptor's value is higher than zero for compounds where charge transfer can occur over a distance of 6, with monosubstituted benzenes and *para*-substituted benzenes having the highest values. Descriptors related to charge distribution account for Coulomb, dipolar interactions, and H-bonding. The positive coefficient in Eq. C.7 is straightforward to understand, as mutagenicity increases with both size and dipole momentum, leading to higher interactions between the molecule and the enzyme.

Another class of descriptors is one that counts certain moieties in a molecule's structure. For instance, in model 3, *fr_bicyclic* represents the number of bicyclic fragments and *NumAliphaticCarbocycles* counts aliphatic carbocycles. These descriptors help model 3 estimate both the size and aromaticity of NACs. However, due to their partial similarity, discussing their signs in the model equation is not straightforward. Nevertheless, both of these descriptors significantly improved the model. Omitting *fr_bicyclic* notably decreases model performances. Thus, for instance, it reduces the value of Q^2_{LOO} from 0.92 to 0.87 and Q^2_{LMO} from 0.89 to 0.80. The exclusion of *NumAliphaticCarbocycles* has lower effect, primarily impacting the values of $RMSE_{tr}$ and $RMSE_{cv}$, which increases from 0.38 to 0.44 and from 0.49 to 0.55, respectively. C3SP2, the number of sp² carbons bound to 3 other carbons, is also part of model 3. In this particular case, C3SP2 counts from the number of methyl groups in the structure. The negative coefficient in Eq. C.3 aligns with literature findings that a higher number of methyl groups decreases mutagenicity (Kuz'min et al. 2008; Jillella et al. 2020). Similarly, *NumRotatableBonds* in model 9, counts the number of substituents (i.e. nitro and methyl groups).

Electron affinity, included in model 4, is a descriptor reflecting the reduction potential of the nitro group—an essential step in mutagenic activation (Wang et al. 2005; Zhang et al. 2008). Accordingly, molecules with higher electron affinity are expected to be more easily activated, contributing to their increased mutagenicity. This aligns with the positive coefficient observed in Eq. C.4. Averaged polarizability (α) is one of the descriptors with the highest correlation with mutagenicity, $R^2=0.68$. In support of this, other descriptors related to polarizability are parts of the models. *SpAbs_Dzp* is the graph energy from the Barysz matrix

weighted by polarizability, while GATS2p and ATSC2p are the Geary and centered Moreau-Broto autocorrelation coefficient of lag 2 weighted by polarizability. Molar refractivity (MolMR) is incorporated in model 6, and it is directly proportional to polarizability. However, MolMR calculations, which sum the contributions of individual atoms, yield the same values for all isomers. In contrast, although α has demonstrated effectiveness as a mutagenicity descriptor for various isomer classes with $R^2 > 0.99$ (Ostojić et al. 2014a; Stanković et al. 2016a) and took a part in the QSAR model for toxicity of nitrobenzene derivatives against *Tetrahymena pyriformis* (Niazi et al. 2008), QC calculations are time-consuming. Given that many enzymes feature hydrophobic active sites, polarizability, directly related to dipole-induced dipole and dispersive interactions, emerges as a valuable descriptor in QSAR models. This aligns with the notion that stronger interactions between NACs and enzymes are facilitated by higher polarizability and is further substantiated by the positive coefficients in the model equation.

Applicability domain analysis

The applicability domain of model 1 was estimated through an analysis of both the plot of the second vs. the first principal component (Fig. 5) and the Williams plot (Fig. 6). Figure 5 underscores the high degree of similarity between the training and test sets, implying a well-executed split into two subsets (training and test) and a well-designed model selection process.

A more robust estimation of the applicability domain can be achieved by the leverage method. As the leverage of each compound in both training and test sets falls below the threshold ($h^*=0.58$), it can be concluded that no response outlier is present. Therefore, the predictions of mutagenicity

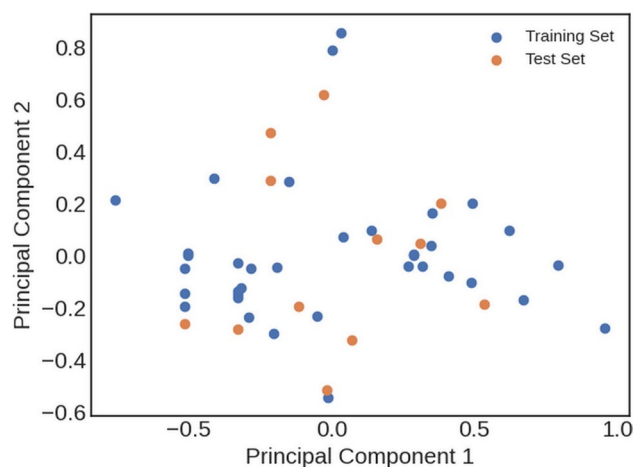


Fig. 5 The plot of the second vs the first principal component of descriptors from model 1

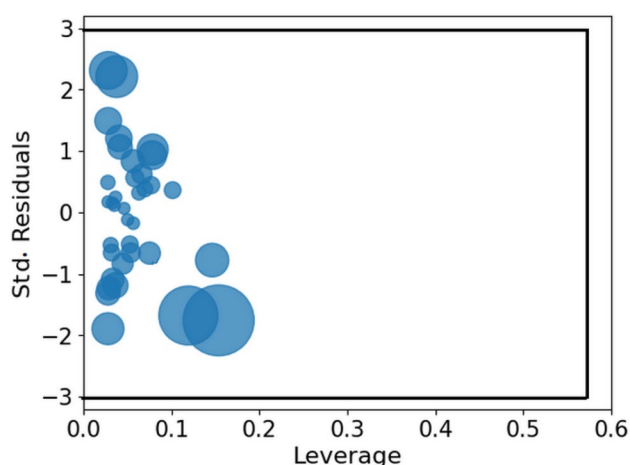


Fig. 6 The Williams plot for the model 1

of NACs could be extrapolated by model 1. Moreover, no structural outliers were identified. In essence, even the two nitroanthracenes are within AD. Thus, there is a stable basis to assert that the proposed model is suitable for predicting mutagenicity across a broad spectrum of NACs.

It is worth noting that many models selected from GA exhibited Y-outliers. This aligns with the previously noted trend that overfitting is mitigated in models derived through the Featurewiz, ForwardSelect, and Boruta algorithm workflow.

Conclusions

The QSAR models developed in this study aimed to predict the mutagenicity of NACs against *Salmonella typhimurium* TA100. The approach involved the utilization of Mordred, RDKit, and quantum-chemical descriptors and used innovative, transparent, interpretable, reproducible, and publicly available feature selection methodologies.

Three feature selection algorithms were used (Boruta, Featurewiz, and ForwardSelector), each followed by forward stepwise feature selection. Models were ranked using a Multi-Criteria Decision Making procedure, i.e. the weighted sum model. According to this procedure, the Featurewiz algorithm performs slightly better than the other two approaches. Nevertheless, all three approaches yield better results than the genetic algorithm, the most commonly used state-of-the-art algorithm in the field. Furthermore, it was found that many models obtained by the genetic algorithm suffer from multicollinearity and overfitting, while also having a narrower applicability domain compared to the three approaches proposed here.

Evaluating the impact of the pair-wise correlation threshold on model performance, the study revealed that

lowering the threshold from 0.99 to 0.85 does not significantly compromise performance, opening possibilities for more time-efficient model development. The research discussed the importance of quantum-chemical descriptors in QSAR models. Although this aspect seems promising, final conclusions need to be reached after more detailed analysis, possibly involving larger datasets. In the end, using machine learning methods existing in the literature, it was shown that average polarizability, as one of the quantum-chemical descriptors, can be efficiently calculated without sacrificing accuracy.

The constructed models showcase excellent fitting ability and external predictive performance. Notably, for their simplicity, robustness, alignment with OECD criteria, and surpassing existing literature models, these QSAR models prove their reliability for practical use, thereby potentially reducing the necessity for experimental procedures for estimating mutagenicity of NACs against *Salmonella typhimurium* TA100. However, the derived models can be used only for molecules containing C, H, N, and O atoms. Moving forward, the study plans to test the derived methodologies on larger sets of NACs and aromatic compounds with a wider variety of substituents, as well as other datasets related to chemical toxicity and other chemical properties.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11356-024-34800-x>.

Acknowledgements B.S. is grateful to Dr Dusan Ramljak from The Pennsylvania State University, Malvern, USA for fruitful discussions during working on conceptualization of a research and proof reading the article.

Author's contributions Branislav Stankovic: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Visualization, Writing—original draft;

Filip Marinkovic: Investigation, Project administration, Visualization, Software, Writing—review & editing

Funding This investigation was supported by the Vinča Institute of Nuclear Sciences and Faculty of Physics through the agreement with the Ministry of Science, Technological Development and Innovation of the Republic of Serbia under Grant. No. 451–03–66/2024–03/200017 and 451–03–66/2024–03/ 200162.

Data availability Any additional data will be made available on request.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

Competing interest The authors have no relevant financial or non-financial interests to disclose.

References

- Akhavan J (2022) The chemistry of explosives. The Royal Society of Chemistry, London. <https://doi.org/10.1039/9781839168802>
- Ambure P, Halder AK, Gonzalez Diaz H, Cordeiro MNDS (2019) QSAR-Co: an open source software for developing robust multitasking or multitarget classification-based QSAR models. *J Chem Inf Model* 59:2538–2544. <https://doi.org/10.1021/acs.jcim.9b00295>
- Arey J, Zielinska B, Atkinson R, Winer AM, Ramdahl T, Pitts JN (1986) The formation of nitro-PAH from the gas-phase reactions of fluoranthene and pyrene with the OH radical in the presence of NOX. *Atmos Environ* 20:2339–2345. [https://doi.org/10.1016/0004-6981\(86\)90064-8](https://doi.org/10.1016/0004-6981(86)90064-8)
- Bazar MA, Quinn MJ, Mozzachio K, Johnson M (2008) Toxicological responses of red-backed salamanders (*Plethodon cinereus*) to subchronic soil exposures of 2,4,6-trinitrotoluene. *Environ Toxicol Chem* 27:1393–1398. <https://doi.org/10.1016/j.envpol.2006.10.007>
- Benbrahim-Tallaa L, Baan RA, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Guha N, Loomis D, Straif K (2012) Carcinogenicity of diesel-engine and gasoline-engine exhausts and some nitroarenes. *Lancet Oncol* 13:663–664. [https://doi.org/10.1016/S1470-2045\(12\)70280-2](https://doi.org/10.1016/S1470-2045(12)70280-2)
- Daghighi A, Casanola-Martin GM, Timmerman T, Milenković D, Lučić B, Rasulev B (2022) In silico prediction of the toxicity of nitroaromatic compounds: Application of ensemble learning QSAR approach. *Toxics* 10:746. <https://doi.org/10.3390/toxics10120746>
- Deng KL, Wong TY, Wang YN, Leung EMK, Chan W (2015) Combination of precolumn nitro-reduction and ultraperformance liquid chromatography with fluorescence detection for the sensitive quantification of 1-nitronaphthalene, 2-nitrofluorene, and 1-nitropyrene in meat products. *J Agric Food Chem* 63:3161–3167. <https://doi.org/10.1021/acs.jafc.5b00523>
- Ding YL, Lyu YC, Leong MK (2017) In silico prediction of the mutagenicity of nitroaromatic compounds using a novel two-QSAR approach. *Toxicol in Vitro* 40:102–114. <https://doi.org/10.1016/j.tiv.2016.12.013>
- Estrada E (2002) Physicochemical interpretation of molecular connectivity indices. *J Phys Chem A* 106:9085–9091. <https://doi.org/10.1021/jp026238m>
- Fan TJ, Sun GH, Zhao LJ, Cui X, Zhong RG (2018) QSAR and classification study on prediction of acute oral toxicity of N-nitroso compounds. *Int J Mol Sci* 19:3015. <https://doi.org/10.3390/ijms19103015>
- Frisch MJ, Trucks GW, Schlegel HB et al (2016) Gaussian 16. Gaussian Inc, Wallingford
- Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. *Int J Quant Struct Prop Relatsh* 1:45–63. <https://doi.org/10.4018/IJQSPR.2016010102>
- Gazquez JL, Cedillo A, Vela A (2007) Electrodonating and electroaccepting powers. *J Phys Chem A* 111:1966–1970. <https://doi.org/10.1021/jp065459f>
- Golbraikh A, Tropsha A (2002) Beware of q²! *J Mol Graph Model* 20:269–276. [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
- Gooch A, Sizochenko N, Rasulev B, Gorb L, Leszczynski J (2017) In vivo toxicity of nitroaromatics: a comprehensive quantitative structure-activity relationship study. *Environ Toxicol Chem* 36:2227–2233. <https://doi.org/10.1002/etc.3761>
- Gozalbes R, de Julian-Ortiz JV (2018) Applications of chemoinformatics in predictive toxicology for regulatory purposes, especially in the context of the EU REACH legislation. *Int J Quant Struct Prop Relatsh* 3:1–24. <https://doi.org/10.4018/IJQSPR.2018010101>
- Gramatica P (2007) Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 26:694–701. <https://doi.org/10.1002/qsar.200610151>
- Gramatica P (2014) External evaluation of QSAR Models, in addition to crossvalidation: verification of predictive capability on totally new chemicals. *Mol Inf* 33:311–314. <https://doi.org/10.1002/minf.201400030>
- Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem* 34:2121–2132. <https://doi.org/10.1002/jcc.23361>
- Halabi A, Ricón A, Chamorro E (2022) Machine learning predictive classification models for the carcinogenic activity of activated metabolites derived from aromatic amines and nitroaromatics. *Toxicol in Vitro* 81:105347. <https://doi.org/10.1016/j.tiv.2022.105347>
- Halder AK, Cordeiro MNDS (2021) QSAR-Co-X: an open source toolkit for multitarget QSAR modelling. *J Cheminform* 13:29. <https://doi.org/10.1186/s13321-021-00508-0>
- Halder AK, Moura AS, Cordeiro MNDS (2022) Moving average-based multitasking in silico classification modeling: where do we stand and what is next? *Int J Mol Sci* 23:4937. <https://doi.org/10.3390/ijms23094937>
- Hao YX, Sun GH, Fan TJ, Sun XD, Liu YD, Zhang N, Zhao LJ, Zhong RG, Peng YZ (2019) Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods. *Ecotoxicol Environ Saf* 186:109822. <https://doi.org/10.1016/j.ecoenv.2019.109822>
- Hao YX, Sun GH, Fan TJ, Tang XY, Zhang J, Liu YD, Zhang N, Zhao LJ, Zhong RG, Peng YZ (2020) In vivo toxicity of nitroaromatic compounds to rats: QSTR modelling and interspecies toxicity relationship with mouse. *J Hazard Mater* 399:122981. <https://doi.org/10.1016/j.jhazmat.2020.122981>
- Hayakawa K (2016) Environmental behaviors and toxicities of polycyclic aromatic hydrocarbons and nitropolycyclic aromatic hydrocarbons. *Chem Pharm Bull* 64:83–94. <https://doi.org/10.1248/cpb.c15-00801>
- Hong H (2023) QSAR in Safety Evaluation and Risk Assessment. Elsevier, Amsterdam. <https://doi.org/10.1016/C2022-0-00536-5>
- Hornberg JJ, Laursen M, Brenden N, Persson M, Thougard AV, Toft DB, Mow T (2014) Exploratory toxicology as an integrated part of drug discovery. Part II: screening strategies. *Drug Discov Today* 19:1137–1144. <https://doi.org/10.1016/j.drudis.2013.12.008>
- Hsu CH, Stedeford T, Okochi-Takada E, Ushijima T, Noguchi H, Muro-Cacho C, Holder JW, Banasik M (2007) Framework analysis for the carcinogenic mode of action of nitrobenzene. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 25:155–184. <https://doi.org/10.1080/10590500701399234>
- Huang T, Sun G, Zhao L, Zhang N, Zhong R, Peng Y (2021) Quantitative Structure-Activity Relationship (QSAR) Studies on the Toxic Effects of Nitroaromatic Compounds (NACs): A Systematic Review. *Int J Mol Sci* 22:8557. <https://doi.org/10.3390/ijms22168557>
- Jillella GK, Khan K, Roy K (2020) Application of QSARs in identification of mutagenicity mechanisms of nitro and amino aromatic compounds against *Salmonella typhimurium* species. *Toxicol in Vitro* 65:104768. <https://doi.org/10.1016/j.tiv.2020.104768>
- Karnjanapiboonwong A, Zhang B, Freitag CM, Dobrovolsky M, Salice CJ, Smith PN, Kendall RK, Anderson TA (2009) Reproductive toxicity of nitroaromatics to the cricket, *Acheta domestica*. *Sci Total Environ* 407:5046–5049. <https://doi.org/10.1016/j.scitotenv.2009.05.048>
- Keshavarz MH, Akbarzadeh AR (2019) A simple approach for assessment of toxicity of nitroaromatic compounds without using

- complex descriptors and computer codes. SAR QSAR Environ Res 30:347–361. <https://doi.org/10.1080/1062936X.2019.1595135>
- Khan K, Benfenati E, Roy K (2019a) 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. <https://doi.org/10.1016/j.scitotenv.2019.05.048>
- Khan K, Roy K, Benfenati E (2019b) Ecotoxicological QSAR modeling of endocrine disruptor chemicals. *J Hazard Mater* 369:707–718. <https://doi.org/10.1016/j.jhazmat.2019.02.019>
- Kovacic P, Somanathan R (2014) Nitroaromatic compounds: Environmental toxicity, carcinogenicity, mutagenicity, therapy and mechanism. *J Appl Toxicol* 34:810–824. <https://doi.org/10.1002/jat.2980>
- Kursa M, Rudnicki W (2010) Feature Selection with the Boruta Package. *J Stat Softw* 36:1–13. <https://doi.org/10.1002/jat.2980>
- Kuz'min VE, Muratov EN, Artemenko AG, Gorb L, Qasim M, Leszczynski J (2008) The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study. *J Comput-Aided Mol Des* 22:747–759. <https://doi.org/10.1007/s10822-008-9211-x>
- Landrum G, Tosco P, Kelley B et al. (2023) rdkit/rdkit: Release_2023.09.5. Zenodo. <https://doi.org/10.5281/zenodo.591637>
- Lent EM (2015) Wildlife toxicity assessment for 2,4-dinitrotoluene and 2,6-dinitrotoluene. In: Williams MA, Reddy G, Quinn MJ Jr, Johnson MS (eds) *Wildlife toxicity assessments for chemicals of military concern*. Elsevier, Amsterdam, pp 415–419. <https://doi.org/10.1016/C2013-0-13473-3>
- Li F, Fan T, Sun G, Zhao L, Zhong R, Peng Y (2022) Systematic QSAR and iQCCR modelling of fused/non-fused aromatic hydrocarbons (FNFAHs) carcinogenicity to rodents: reducing unnecessary chemical synthesis and animal testing. *Green Chem* 24:5304–5319. <https://doi.org/10.1039/d2gc00986b>
- Loomis D, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H, Straif K (2013) The carcinogenicity of outdoor air pollution. *Lancet Oncol* 14:1262–1263. [https://doi.org/10.1016/S1470-2045\(13\)70487-X](https://doi.org/10.1016/S1470-2045(13)70487-X)
- Mondal D, Ghosh K, Baidya AT, Gantait AM, Gayen S (2020) Identification of structural fingerprints for in vivo toxicity by using Monte Carlo based QSTR modeling of nitroaromatics. *Toxicol Mech Methods* 30:257–265. <https://doi.org/10.1080/15376516.2019.1709238>
- Moriwaki H, Tian YS, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. *J Cheminformatics* 10:1–14. <https://doi.org/10.1186/s13321-018-0258-y>
- Mortelmans K, Zeiger E (2000) The Ames Salmonella/microsome mutagenicity assay. *Mutat Res* 455:29–60. [https://doi.org/10.1016/S0027-5107\(00\)00064-6](https://doi.org/10.1016/S0027-5107(00)00064-6)
- Niazi A, Jameh-Bozorgi S, Nori-Shargh D (2008) Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines. *J Hazard Mater* 151:603–609. <https://doi.org/10.1016/j.jhazmat.2007.06.030>
- OECD (Organization for Economic Co-Operation and Development) (2007) *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q) SAR] Models*. OECD Environment Health and Safety Publications Series on Testing and Assessment No. 69, Paris
- Ostojić B, Stanković B, Đorđević D (2014a) The molecular properties of nitrobenzantrone isomers and their mutagenic activities. *Chemosphere* 104:228–236. <https://doi.org/10.1016/j.chemosphere.2013.11.057>
- Ostojić B, Stanković B, Đorđević D (2014b) Theoretical study of the molecular properties of dimethylanthracenes as properties for the prediction of their biodegradation and mutagenicity. *Chemosphere* 111:144–150. <https://doi.org/10.1016/j.chemosphere.2014.03.067>
- Pérez-Garrido A, Morales Helguera A, Caravaca López G, Cordeiro MNDS, Garrido Escudero A (2010) A topological substructural molecular design approach for predicting mutagenesis end-points of α , β -unsaturated carbonyl compounds. *Toxicology* 268:64–77. <https://doi.org/10.1016/j.tox.2009.11.023>
- Pérez-Garrido A, Girón-Rodríguez F, Morales Helguera A, Borges F, Combes RD (2014) Topological structural alerts modulations of mammalian cell mutagenicity for halogenated derivatives. *SAR QSAR Environ Res* 25:17–33. <https://doi.org/10.1080/1062936X.2013.820791>
- Pesode P, Barve S, Wankhede SV, Jadhav DR, Rawar SK (2023) Titanium alloy selection for biomedical application using weighted sum model methodology. *Mater Today: Proc* 7:724–728. <https://doi.org/10.1016/j.matpr.2022.08.494>
- Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. *J Chem Inf Model* 52:396–408. <https://doi.org/10.1021/ci200520g>
- OECD (2004) QSAR Expert Group. The report from the expert group on (Quantitative) Structure-Activity Relationships [(Q) SARs on the principles for the validation of (Q) SARs. Organisation For Economic co-Operation and Development 49:206
- Singh J, Singh S, Shaik B, Deeb O, Sohani N, Agrawal VK, Khadikar PV (2008) Mutagenicity of nitrated polycyclic aromatic hydrocarbons: a QSAR investigation. *Chem Biol Drug Des* 7:230–243. <https://doi.org/10.1111/j.1747-0285.2008.00629.x>
- Stanković B, Ostojić B, Popović A, Gruden M, Đorđević D (2016a) Theoretical study of nitrobenzofurans: A possible relationship between molecular properties and mutagenic activity. *J Hazard Mater* 318:623–630. <https://doi.org/10.1016/j.jhazmat.2016.07.035>
- Stanković B, Ostojić B, Popović A, Gruden M, Đorđević D (2016b) Substituted naphthalenes: Stability, conformational flexibility and description of bonding based on ETS-NOCV method. *Chem Phys Lett* 661:136–142. <https://doi.org/10.1016/j.cplett.2016.08.056>
- Tiwari J, Tarale P, Sivanesan S, Bafana A (2019) Environmental persistence, hazard, and mitigation challenges of nitroaromatic compounds. *Environ Sci Pollut Res* 26:28650–28667. <https://doi.org/10.1007/s11356-019-06043-8>
- Todeschini R, Consonni V, Mauri A, Pavan M (2004) Detecting “bad” regression models: multicriteria fitness functions in regression analysis. *Anal Chim Acta* 515:199–208. <https://doi.org/10.1016/j.aca.2003.12.010>
- Toropov AA, Toropova AP, Roncaglioni A, Benfenati E (2023) In silico prediction of the mutagenicity of nitroaromatic compounds using correlation weights of fragments of local symmetry. *Mutat Res Genet Toxicol Environ Mutagen* 891:503684. <https://doi.org/10.1016/j.mrgentox.2023.503684>
- Toropova AP, Toropov AA, Roncaglioni A, Benfenati E (2023) The enhancement scheme for the predictive ability of QSAR: A case of mutagenicity. *Toxicol in Vitro* 91:105629. <https://doi.org/10.1016/j.tiv.2023.105629>
- Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* 22:69–77. <https://doi.org/10.1002/qsar.200390007>
- Wang X, Lin Z, Yin D, Liu S, Wang L (2005) 2D/3D-QSAR comparative study on mutagenicity of nitroaromatics. *Sci China Ser B Chem* 48:246–252. <https://doi.org/10.1360/04yb0077>
- Wilkins DM, Grisafi A, Yang Y, Lao KU, DiStasio RA, Ceriotti M (2019) Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc Natl Acad Sci USA* 116:3401–3406. <https://doi.org/10.1073/pnas.1816132116>
- Yu M-H, Tsunoda H, Tsunoda M (2016) *Environmental Toxicology: Biological and Health Effects of Pollutants*, 3rd edn. CRC Press, Boca Raton

- Zhang G, Musgrave CB (2007) Comparison of DFT methods for molecular orbital eigenvalue calculations. *J Phys Chem A* 111:1554–1561. <https://doi.org/10.1021/jp061633o>
- Zhang Z, Niu J, Zhi X (2008) A QSAR Model for Predicting mutagenicity of nitronaphthalenes and methylnitronaphthalenes. *Bull Environ Contam Toxicol* 81:498–502. <https://doi.org/10.1007/s00128-008-9540-4>
- Zhang CL, Yu YY, Fang Z, Naraginti S, Zhang YH, Yong YC (2018) Recent advances in nitroaromatic pollutants bioreduction by electroactive bacteria. *Process Biochem* 70:129–135. <https://doi.org/10.1016/j.procbio.2018.04.019>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.