International Journal of
Molecular Sciences

MDPI

Article

QSAR and Classification Study on Prediction of Acute Oral Toxicity of *N*-Nitroso Compounds

Tengjiao Fan †, Guohui Sun †, Lijiao Zhao *, Xin Cui and Rugang Zhong

Beijing Key Laboratory of Environmental & Viral Oncology, College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100124, China; fantengjiao2014@emails.bjut.edu.cn (T.F.); sunguohui@bjut.edu.cn (G.S.); cuixin1201@bjut.edu.cn (X.C.); lifesci@bjut.edu.cn (R.Z.)

* Correspondence: zhaolijiao@bjut.edu.cn; Tel.: +86-10-6739-1667

† These authors contributed equally to this work.

Received: 10 September 2018; Accepted: 30 September 2018; Published: 3 October 2018

Abstract: To better understand the mechanism of in vivo toxicity of *N*-nitroso compounds (NNCs), the toxicity data of 80 NNCs related to their rat acute oral toxicity data (50% lethal dose concentration, LD_{50}) were used to establish quantitative structure-activity relationship (QSAR) and classification models. Quantum chemistry methods calculated descriptors and Dragon descriptors were combined to describe the molecular information of all compounds. Genetic algorithm (GA) and multiple linear regression (MLR) analyses were combined to develop QSAR models. Fingerprints and machine learning methods were used to establish classification models. The quality and predictive performance of all established models were evaluated by internal and external validation techniques. The best GA-MLR-based QSAR model containing eight molecular descriptors was obtained with $Q^2_{loo} = 0.7533$, $R^2 = 0.8071$, $Q^2_{ext} = 0.7041$ and $R^2_{ext} = 0.7195$. The results derived from QSAR studies showed that the acute oral toxicity of NNCs mainly depends on three factors, namely, the polarizability, the ionization potential (IP) and the presence/absence and frequency of C–O bond. For classification studies, the best model was obtained using the MACCS keys fingerprint combined with artificial neural network (ANN) algorithm. The classification models suggested that several representative substructures, including nitrile, hetero N nonbasic, alkylchloride and amine-containing fragments are main contributors for the high toxicity of NNCs. Overall, the developed QSAR and classification models of the rat acute oral toxicity of NNCs showed satisfying predictive abilities. The results provide an insight into the understanding of the toxicity mechanism of NNCs in vivo, which might be used for a preliminary assessment of NNCs toxicity to mammals.

Keywords: *N*-nitroso compounds; acute oral toxicity; QSAR; classification; toxicity mechanism

1. Introduction

N-nitroso compounds (NNCs) are an important class of potent toxicants that widely exist in the environment and diet [1]. The carcinogenicity, mutagenicity and toxicity of NNCs and their metabolites have been evaluated in various experiments [2–4]. Among the 300 NNCs that have been tested for their carcinogenic potential, more than 90% were proven to be carcinogenic in a wide variety of animal species [5,6]. Human exposure to NNCs occurs mainly through food, tobacco products, drugs, car interiors, and cosmetics [7]. However, NNCs may also be synthesized endogenously from precursors and nitrosating agents, mainly in the stomach, leading to the formation of potentially carcinogenic compounds [8–11]. Due to the potentially harmful effects of these compounds, it is necessary to study the mechanism of action of their biological effects, particularly the structure-activity relationship (SAR).

Quantitative structure-activity relationship (QSAR) and classification methods are ideal alternatives to biological experiments. Not only because of their higher efficiency and lower cost, but they can also provide rapid assessment of the potential impacts of chemicals on human health and the environment, including lethality or non-lethal adverse effects, as well as being able to predict biological or physicochemical properties [12,13]. Thus, the European Union (EU) published REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) regulation for promoting their applications in various fields in 2006. The QSAR and classification models have been developed as feedback to different legislation around the world (e.g., EU REACH) as well as to assist in reducing animal testing and designing greener chemicals [14–16].

Generally, the acute toxicity of most chemicals is mainly induced by a narcotic mechanism of action, which has been long termed as "membrane perturbation". Narcotic compounds certainly accumulate within biological membranes, thus, a number of effects at the membrane occur. If a compound can be identified as being unreactive or narcotic, its acute toxicity to a variety of species can be predicted accurately from the structure alone [17]. However, a number of compounds have specific toxic mechanisms of action (e.g., inhibition of specific enzymes or electrophilic/nucleophilic reaction) [17]. For example, NNCs can produce alkyldiazonium ions through metabolic activation by specific enzymes or spontaneous decomposition, followed by attacking bio-macromolecules (e.g., DNA and proteins) to exert their toxicity [4].

Previous studies have reported some predictive models of the carcinogenic potential of NNCs. Based on "di-region theory", a quantitative pattern recognition method performed for structure-carcinogenic activity relationship of NNCs gave rise to 97% correct classification using 10 descriptors [18]. In addition, the results suggested that the bifunctional alkylation between α and β sites or α and γ sites of NNCs provided important roles in their carcinogenesis. The support vector machine (SVM) and linear discriminant analysis (LDA) were used to develop a classification model of carcinogenic properties of 148 NNCs with seven descriptors [19]. The obtained results confirmed the discriminative capacity of the calculated descriptors and the total accuracy of SVM (95.2%) is better than that of LDA (89.8%). Using a topological substructure molecular descriptors (TOPS-MODE) approach, Helguera et al. constructed several QSAR models for predicting the carcinogenic effects of NNCs through different routes of administration for male and female rats [20,21]. Yuan et al. developed an LDA method to predict the carcinogenicity and further understand the carcinogenic mechanism of NNCs in rats using a TOPS-MODE approach. The results indicated that a good classification (carcinogenic and noncarcinogenic) value of 90.1% was obtained with a dataset of 111 NNCs [7].

Although several SAR studies in the perspective of carcinogenicity of NNCs have been reported, to our knowledge, there are still no related studies on the relationship between molecular structure or properties and acute oral toxicity of NNCs. In the present work, a dataset consisting of acute oral toxicity (LD_{50}) of 80 NNCs to rats was used to establish the QSAR and classification prediction models. The developed models were assessed using various statistical parameters and an external validation set. Based on the analysis of these developed models, some important information in connection with toxicity can be obtained, which may help us better understand the bio-transformation and toxic mechanism of NNCs in vivo. Moreover, these QSAR and classification models may provide a way to evaluate and predict the toxicity of many other untested NNCs, before they have adverse effects on both humans and the environment.

2. Results and Discussion

2.1. QSAR Models

2.1.1. Model Validation

The initial number of descriptors of MLR (multiple linear regression) model developed based on Dragon and DFT (density functional theory) were 457 after removal of constant value and high inter-correlated descriptors. Then, the further screening was executed by GA (genetic algorithm) [22] coupled with the MLR procedure, followed by the generation of 100 models. All 80 NNCs were ranked according to the toxicity value ($-\log LD_{50}$), then one was selected as the test set every five compounds and the remaining 64 compounds were used as the training set. According to the rule-of-thumb [23,24], the ratio of the number of compounds in the training set over the number of variables (descriptors) should have a value of at least 5, which allows the flexibility to build models using up to 13 descriptors in the present study. After utilizing QUIK (Q Under Influence of K) module, 44 models remained without multicollinearity. For acceptable QSAR predictive models, they should satisfy the following conditions [24,25]: (i) $Q^2_{loo} > 0.5$; (ii) $R^2_{ext} > 0.6$; (iii) $(R^2_{ext} - R_0^2)/R^2_{ext} < 0.1$ and $0.85 \leq k \leq 1.15$ or $(R^2_{ext} - R'^2_0)/R^2_{ext} < 0.1$ and $0.85 \leq k' \leq 1.15$; (iv) $|R_0^2 - R'^2_0| < 0.3$. R_0^2 and R'^2_0 are the mean coefficients of determination of experimental versus predicted values and predicted versus experimental values for regressions through the origin, respectively. k and k' are the corresponding slopes of regression lines through the origin. Finally, eight models were selected by MCDM (Multi-Criteria Decision Making) (Figure S1 in the Supplementary Materials), which Q^2_{loo} values ranged from 0.7214 to 0.7533 (R^2 = 0.7786 to 0.8071), as listed in Table 1. Among these models, six descriptors were observed with higher frequency than other descriptors, namely, MATS6p, MATS4i, SpMin7_Bh(i), JGI4, B01[C-O] and F04[C-O]. The best QSAR model with eight descriptors for the prediction of acute oral toxicity of NNCs was shown in Equation (1). The actual values of selected descriptors in the best QSAR model were presented in Table S1 in the Supplementary Materials.

$$\begin{gathered} -\log LD_{50} = 2.86 + 0.28\mathrm{nR06} - 0.55\mathrm{MATS6p} - 1.29\mathrm{MATS4i} - 12.07\mathrm{JGI4} - \\ 2.06\mathrm{SpMin7\_Bh(i)} - 0.50\mathrm{B01[C\text{-}O]} - 0.35\mathrm{F04[C\text{-}O]} - 7.36E_{\mathrm{HOMO}} \end{gathered} \tag{1}$$

$$\begin{gathered} N_{tr} = 65,\ Q^2_{loo} = 0.7533,\ R^2 = 0.8071,\ R^2_{adj} = 0.7796,\ F = 29.2961,\ RMSE_{tr} = 0.2661,\ CCC_{tr} = 0.8933 \\ N_{test} = 14,\ Q^2_{ext} = 0.7041,\ R^2_{ext} = 0.7195,\ RMSE_{test} = 0.2847,\ Q^2_{F1} = 0.7041,\ Q^2_{F2} = 0.7032, \\ Q^2_{F3} = 0.7794,\ CCC_{test} = 0.8062,\ (R^2_{ext} - R_0^2)/R^2_{ext} = 0.0215,\ |R_0^2 - R'^2_0| = 0.2642. \end{gathered}$$

N_{tr} and N_{test} represent the number of compounds in the training and test sets, respectively. One compound in the original test set was removed because it was a predictive outlier (grey open circle in Figure 1). The relatively high quality of fitting parameters (R^2, R^2_{adj} and RMSE) and internal cross-validation correlation coefficient (Q^2_{loo}) indicate that the model has good internal fitting ability and robustness. A test set containing 14 compounds independent from the training set was used for an external validation to confirm the predictive ability of the MLR model. As shown in Table 2, the predictive ability of this model is high, which is reflected by Q^2_{ext}, R^2_{ext} and $RMSE_{test}$ as 0.7041, 0.7195 and 0.2847, respectively. The good external prediction was also observed with high CCC_{ext} (Concordance Correlation Coefficient) value (0.8062). Furthermore, a Y-scrambling procedure gave significantly lower statistical parameters (R^2_{Yscr} = 0.1247, Q^2_{Yscr} = −0.1890) when compared to the original model, thus we considered that the proposed QSAR model was not obtained casually.

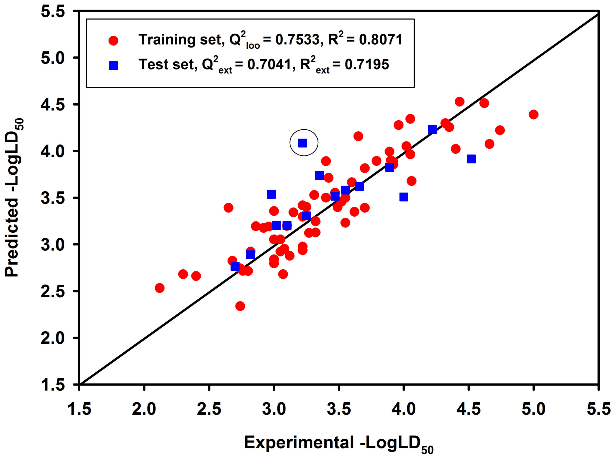


Figure 1. Experimental versus predicted toxicity values for compounds in the training set (red circle) and test set (blue square) of the best GA-MLR (genetic algorithm- multiple linear regression)-based quantitative structure-activity relationship (QSAR) model.

Table 1. Fitting and internal validation parameters of GA-MLR-based QSAR models selected by Multi-Criteria Decision Making (MCDM).

No.	Model No.	Number of Descriptors	Descriptors	R^2	R^2_{adj}	$RMSE_{tr}$	CCC_{tr}	F	Q^2_{loo}	$RMSE_{cv}$	CCC_{cv}	Q^2_{lmo}	R^2_{Yscr}	Q^2_{Yscr}
1	21	8	nR06 MATS6p MATS4i JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O] HOMO	0.8071	0.7796	0.2661	0.8933	29.2961	0.7533	0.3010	0.8651	0.7379	0.1247	−0.1890
2	27	8	nR06 MATS6p MATS4i JGI4 SpMin7_Bh(i) P_VSA_MR_1 B01[C-O] F04[C-O]	0.8033	0.7752	0.2688	0.8909	28.5870	0.7432	0.3071	0.8596	0.7267	0.1247	−0.1856
3	29	8	D/Dtr06 MATS6p MATS4i JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O] HOMO	0.8023	0.7740	0.2695	0.8903	28.3984	0.7504	0.3028	0.8632	0.7335	0.1268	−0.1880
4	33	7	MATS6p MATS4i GATS1m JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O]	0.7872	0.7611	0.2796	0.8809	30.1220	0.7322	0.3136	0.8520	0.7169	0.1097	−0.1644
5	34	7	Mp MATS6p MATS4i JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O]	0.7848	0.7584	0.2811	0.8794	29.6947	0.7262	0.3171	0.8484	0.7076	0.1100	−0.1636
6	36	7	MATS6p MATS4i JGI4 SpMin7_Bh(i) H-046 B01[C-O] F04[C-O]	0.7807	0.7538	0.2838	0.8768	28.9864	0.7276	0.3163	0.8491	0.7140	0.1077	−0.1700
7	37	7	ZM1Mad MATS6p MATS4i GGI4 SpMin7_Bh(i) B01[C-O] F04[C-O]	0.7797	0.7527	0.2844	0.8762	28.8222	0.7214	0.3199	0.8446	0.7045	0.1094	−0.1669
8	38	7	MATS6p MATS4i JGI4 SpMin5_Bh(s) P_VSA_MR_1 B01[C-O] F04[C-O]	0.7786	0.7514	0.2851	0.8755	28.6378	0.7223	0.3194	0.8463	0.7040	0.1104	−0.1621

Table 2. External validation parameters of GA-MLR-based QSAR models selected by MCDM.

No.	Model No.	Number of Descriptors	Descriptors	R^2_{ext}	$RMSE_{ext}$	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC_{ext}	*k*	*k′*
1	21	8	nR06 MATS6p MATS4i JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O] HOMO	0.5401 (0.7195)	0.3544 (0.2847)	0.5147 (0.7041)	0.5144 (0.7032)	0.6581 (0.7794)	0.7023 (0.8062)	0.9774 (0.9957)	1.0132 (0.9977)
2	27	8	nR06 MATS6p MATS4i JGI4 SpMin7_Bh(i) P_VSA_MR_1 B01[C-O] F04[C-O]	0.5100 (0.6534)	0.3709 (0.3080)	0.4685 (0.6538)	0.4681 (0.6527)	0.6255 (0.7418)	0.7003 (0.7934)	0.9784 (0.9944)	1.0110 (0.9994)
3	29	8	D/Dtr06 MATS6p MATS4i JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O] HOMO	0.5153 (0.7175)	0.3659 (0.2908)	0.4862 (0.6912)	0.4823 (0.6902)	0.6355 (0.7697)	0.6767 (0.7914)	0.9742 (0.9933)	1.0159 (0.9999)
4	33	7	MATS6p MATS4i GATS1m JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O]	0.4632 (0.5963)	0.3806 (0.3354)	0.4381 (0.5894)	0.4399 (0.5881)	0.6056 (0.6938)	0.6398 (0.7221)	0.9787 (0.9946)	1.0101 (0.9962)
5	34	7	Mp MATS6p MATS4i JGI4 SpMin7_Bh(i) B01[C-O] F04[C-O]	0.4712 (0.6587)	0.3813 (0.3116)	0.4381 (0.6455)	0.4377 (0.6443)	0.6041 (0.7356)	0.6508 (0.7622)	0.9752 (0.9944)	1.0138 (0.9977)
6	36	7	MATS6p MATS4i JGI4 SpMin7_Bh(i) H-046 B01[C-O] F04[C-O]	0.4726	0.3780	0.4478	0.4474	0.6109	0.6609	0.9804	1.0084
7	37	7	ZM1Mad MATS6p MATS4i GGI4 SpMin7_Bh(i) B01[C-O] F04[C-O]	0.6322	0.3295	0.5805	0.5802	0.7044	0.7851	0.9751	1.0171
8	38	7	MATS6p MATS4i JGI4 SpMin5_Bh(s) P_VSA_MR_1 B01[C-O] F04[C-O]	0.4710 (0.5055)	0.3786 (0.3721)	0.4461 (0.4944)	0.4457 (0.4927)	0.6097 (0.6229)	0.6702 (0.6971)	0.9855 (0.9941)	1.0030 (0.9946)

The linear correlation between the experimental and predicted values from the best GA-MLR-based QSAR model (No. 21) was shown in Figure 1, in which red circles and blue squares represent compounds in training set and test set, respectively. All the studied NNCs are distributed evenly on both sides of the optimal line, indicating the good predictive power of this model. In addition, we applied the best prediction model on several NNCs from the ZINC database and found potentially toxic compounds without tested toxicity on rats, the results are listed in Table S2 in the Supplementary Materials.

2.1.2. Outlier Analysis of MLR Model

In developing the QSAR model, outliers strongly influence the regression parameters of the model. As a result, models should be re-established after outliers are removed. Williams plot, which represents the AD of the MLR model, is shown in Figure 2. It is very important to note that hat values of all compounds are lower than the critical hat value (h^* = 0.415). Only one compound (**53**) in this study was identified as a predictive outlier because its standardized residual was slightly bigger than 3. In other words, the acute oral toxicity values of NNCs are generally well predicted by model 21 and they are reliable.

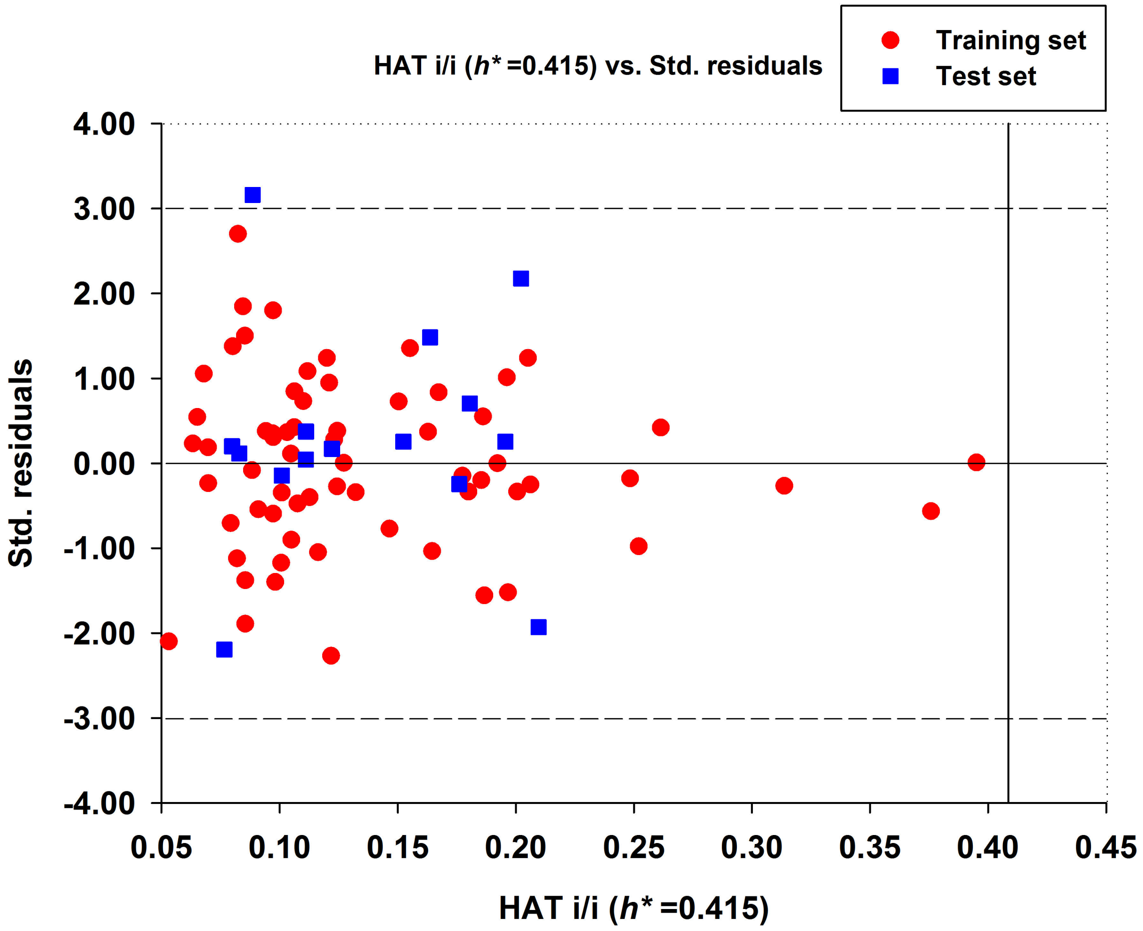


Figure 2. Williams plot for the best GA-MLR-based QSAR model. The transverse dash lines represent ±3 standard residual, vertical black line represents warning leverage h^* = 0.415.

2.1.3. Interpretation of Descriptors in MLR Model

Equation (1) indicates the best GA-MLR-based QSAR model consists of the following eight molecular descriptors: nR06, MATS6p, MATS4i, JGI4, SpMin7_Bh(i), B01[C-O], F04[C-O] and E_{HOMO}. The corresponding types and chemical meanings of molecular descriptors are listed in Table 3, and the detailed explanation can be found in Handbook of Molecular Descriptors [26]. nR06 means the number of 6-membered rings, which is the only variable positively correlated with the high in vivo toxicity of NNCs. There were 7 compounds that contain 6-membered rings in the 22 high toxic NNCs, their molecular structures and LD_{50} values were shown in Figure 3. MATS6p is the Moran autocorrelation of lag 6 weighted by atomic polarizability, indicating a relationship between molecular polarizability and toxicity. According to the handbook of molecular descriptor [26], the Moran coefficient of the autocorrelation descriptors usually takes a value ranging from −1 to +1. Positive autocorrelation produces positive values of the coefficient whereas negative autocorrelation corresponds to negative values. MATS6p tends to have low values when the polarizabilities of bonded atoms are large. Matteo Cassotti et al. also found a relationship between molecular polarizability and acute aquatic toxicity of 546 organic molecules [27]. Polarizable molecules are usually considered as 'soft' species, which tend to react with other soft species, it thus appears that more-polarizable molecules tend to have higher toxicities, and this might be due to the formation of covalent bonds involving the HOMO and LUMO of soft acids and bases [27]. The MATS4i is also the Moran coefficient of the autocorrelation descriptors, while SpMin7_Bh(i) belongs to Burden eigenvalues. The MATS4i and SpMin7_Bh(i) are both related to ionization potential (IP), which is defined as the energy needed to extract one electron from a chemical system. The equation is shown as below:

$$\mathrm{IP} = \mathrm{E}(\mathrm{N}_{\mathrm{el}}) - \mathrm{E}\ (\mathrm{N}_{\mathrm{el}} - 1) \quad (2)$$

where N_{el} is the number of electrons in the system. IP can be used to measure the capability of a molecule to give the corresponding positive ion. The low values of MATS4i correspond to the compounds that have C=C bonds. Zhang et al. [28] demonstrated a good relationship between epoxidation activation energies and IP, which means that the activation energy of epoxidation by P450s strongly depends on the conversion of the double bond in the olefin to a single bond in the product. There are also some NNCs containing olefins which can be activated by P450s to exert their toxicity to a certain degree. JGI4 is a kind of topological charge indices (Mean topological charge index of order 4) which can evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [29,30]. B01[C-O] and F04[C-O] are both 2D atom pairs descriptors that describe pairs of atoms and bond types connecting them in 2D space. They represent the presence/absence of C–O bond and frequency of C–O bond at corresponding topological distance, respectively. There was a negative correlation between these two descriptors and in vivo toxicity of NNCs. The last descriptor $\boldsymbol{E}_{\mathbf{HOMO}}$ is a quantum chemistry descriptor. Molecules with high HOMO (highest occupied molecular orbital) energy values can donate their electrons more easily compared to molecules with low HOMO energy values, and hence are more reactive. Therefore, within the validity of the Koopman's theorem, the $\boldsymbol{E}_{\mathbf{HOMO}}$ descriptor is also related to the IP, is a measure of the nucleophilicity of a molecule, and is important in modeling molecular properties and reactivity [31].

Table 3. Type and chemical meaning of molecular descriptors in the best QSAR model.

Descriptor	**Type**	**Chemical Meaning**
nR06	Ring descriptors	Number of 6-membered rings
MATS6p	2D autocorrelations	Moran autocorrelation of lag 6 weighted by polarizability
MATS4i	2D autocorrelations	Moran autocorrelation of lag 4 weighted by ionization potential
JGI4	2D autocorrelations	Mean topological charge index of order 4
SpMin7_Bh(i)	Burden eigenvalues	Smallest eigenvalue n. 7 of Burden matrix weighted by ionization potential
B01[C-O]	2D Atom Pairs	Presence/absence of C–O at topological distance 1
F04[C-O]	2D Atom Pairs	Frequency of C–O at topological distance 4
E_{HOMO}	QM descriptors	Highest occupied molecular orbital energy

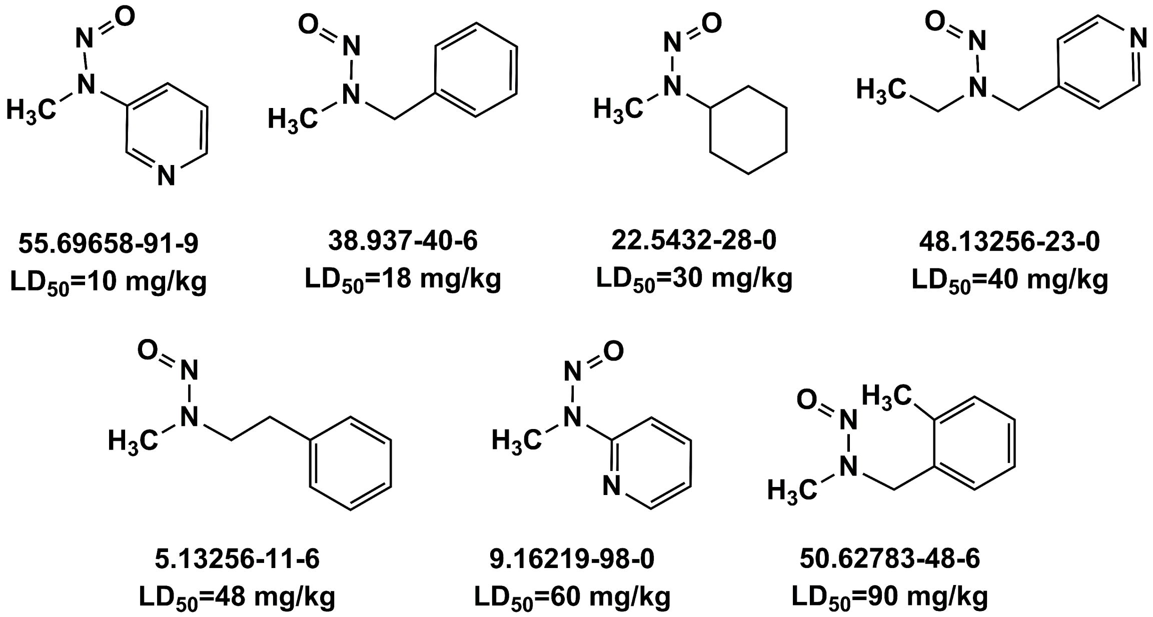


Figure 3. Several typical compounds that contain 6-membered rings in 22 high toxic *N*-nitroso compounds (NNCs).

Considering all the molecular descriptors, acute oral toxicity of NNCs is mainly associated with three properties: Polarizability, IP, and the presence/absence and frequency of C–O bond. In addition, types of structural fragments (i.e., nR06) and charge/electrons transfer in molecules also affect molecular toxicity of NNCs.

2.2. Classification Models

2.2.1. Data Set Analysis

In this study, a total of 80 NNCs collected from the US National Library of Medicine TOXNET ChemIDplus database were used for model building and validation. The 80 NNCs compounds were divided by the classification criterion of 200 mg/kg, then a dataset with 22 high toxic compounds and 58 low toxic compounds was obtained. The training set consisted of 15 high toxic and 41 low toxic compounds while the external test set contained 7 high toxic and 17 low toxic compounds. As an added precaution, it was verified that each set contained roughly the same percentage of high toxic compounds (training set = 26.8%, test set = 29.2%).

Chemical diversity is important to build a robust and reliable prediction model. We have investigated the chemical space distribution by calculating the molecule weight (MW) and Ghose−Crippen LogKow (ALogP) of the training set and the external test set [32]. The distribution scatter diagram was presented in Figure 4A. The scatter diagram showed that the chemical space of compounds in the external test set was within the scope of the training set. To further explore the chemical diversity of the data set, the Euclidian distance metrics of the data set was calculated. The training and external test sets were compared with each other, and the heat map of Euclidian distance metrics was shown in Figure 4B. It is clear that the similarity between the training and external test sets was low.

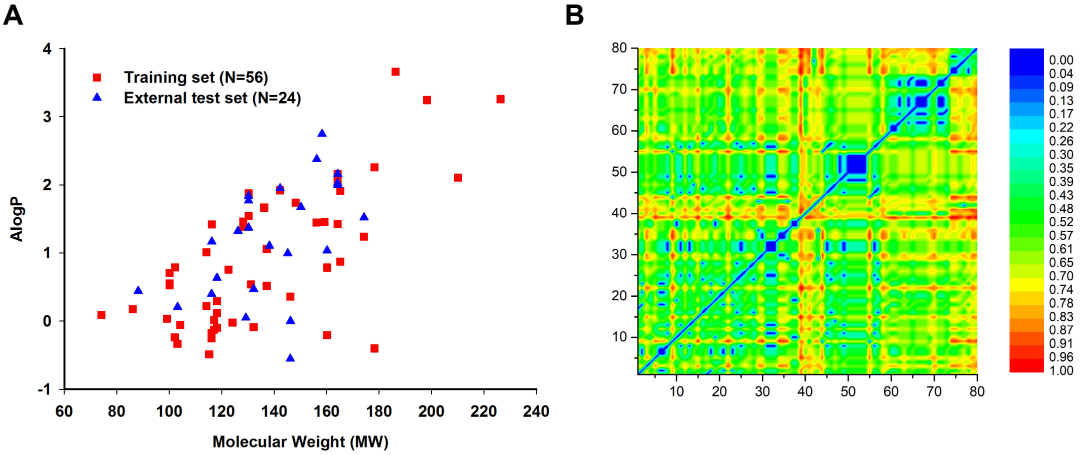


Figure 4. Chemical diversity analysis of the training and external test sets. (**A**) Chemical space was defined by molecular weight (MW) and Ghose−Crippen LogKow (ALogP). N represents the chemical number of different data sets. (**B**) Similarity heat map of Euclidian distance metrics calculated using MACCS keys fingerprint for the training and external test sets.

2.2.2. Performances of 10-Fold Cross-Validation

In our classification study, we built the combinatorial predictive models by using four different fingerprints along with seven statistical algorithms. As a result, a total of 28 binary classification models were generated. The detailed evaluation results of these models are shown in Figure 5. The performance of these models was evaluated by 10-fold cross-validation, and the best models were selected based on the values of CA (classification accuracy) and AUC (the area under the ROC curve). As shown in Figure 5, most models had CA values more than 0.6, except for MACCS-NB and PubChem-NB models. Similarly, most models were obtained with AUC values higher than 0.6, except for the SubFP-NB and SubFP-SVM models. According to the results, the top eight ranking models were MACCS-ANN, PubChem-ANN, SubFP-ANN, PubChem-LR, PubChem-RF, Est-ANN, MACCS-LR and MACCS-SVM. Their CA values were 0.732–0.839 and AUC values were 0.770–0.905. The values of specificity (SP) were higher than that of sensitivity (SE) in all models, which means that all models have higher prediction accuracy for low toxic compounds rather than high toxic compounds. The underlying reason might be that more low toxic compounds existed in the data set. The detailed performances of the top eight models are shown in Table 4. By comparing the performance of four fingerprints, we could draw a conclusion that the MACCS and PubChem fingerprints are appropriate for the classification study of NNCs regarding in vivo toxicity. Based on the well-defined structural fragments dictionary, MACCS molecular fingerprint is full of structural information [33]. In previous studies, MACCS and PubChem fingerprints had also been proven to outperform other fingerprints in classifier models [33,34]. By contrast, the Est fingerprint performed worst when the same machine learning methods were used. This might due to the nature of the Est fingerprint, where only 79 bits signified substructure patterns are involved. It seems that the 79 bits are too short to represent diverse fragments of all compounds. When using the same molecular fingerprint, ANN and LR algorithms were better than other methods (*k*NN, NB, SVM, RF and Tree) in this study. For example, the 10-fold cross-validation results using MACCS showed that the AUC values of MACCS-ANN and MACCS-LR models were 0.905 and 0.832 respectively, whereas the values were 0.738, 0.655, 0.770, 0.767 and 0.619 in MACCS-*k*NN, MACCS-NB, MACCS-SVM, MACCS-RF and MACCS-Tree models, respectively.

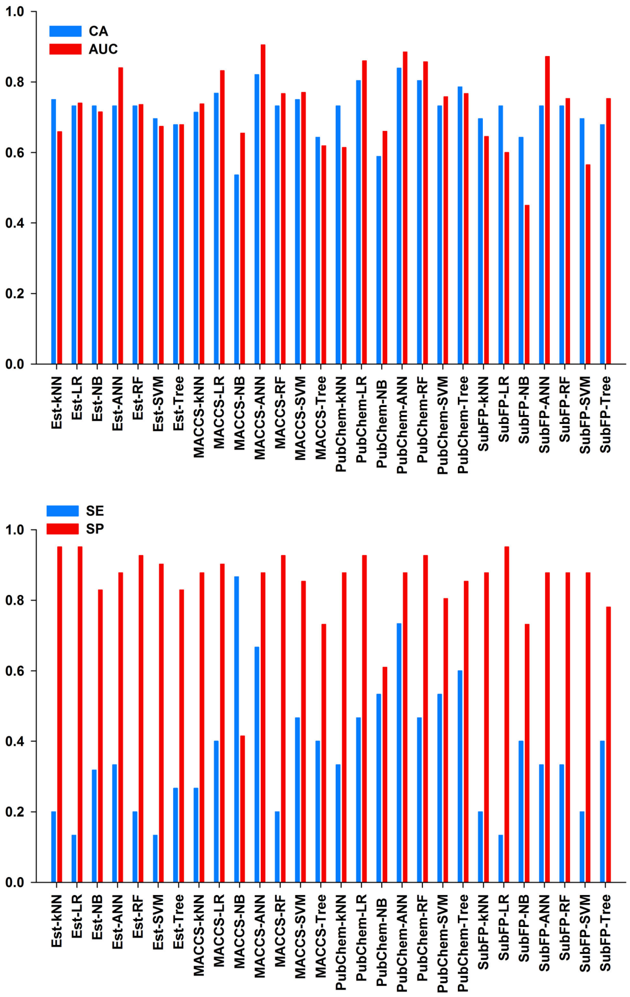


Figure 5. Performance of 10-fold cross-validation for the training set in 28 classification models. CA, classification accuracy; AUC, the area under the ROC curve; SE, sensitivity; SP, specificity.

Table 4. Performance of the top eight models for training set and external test set in classification study [1].

Data Set	Model	CA	SE	SP	AUC	TP	TN	FP	FN
Training set	MACCS-ANN	0.821	0.67	0.88	0.905	10	36	5	5
	PubChem-ANN	0.839	0.73	0.88	0.885	11	36	5	4
	SubFP-ANN	0.732	0.33	0.88	0.872	5	36	5	10
	PubChem-LR	0.804	0.47	0.93	0.860	7	38	3	8
	PubChem-RF	0.804	0.47	0.93	0.857	7	38	3	8
	Est-ANN	0.732	0.33	0.88	0.840	5	36	5	10
	MACCS-LR	0.768	0.40	0.90	0.832	6	37	4	9
	MACCS-SVM	0.750	0.47	0.85	0.770	7	35	6	8
Test set	MACCS-ANN	0.792	0.29	1.00	0.992	2	17	0	5
	PubChem-ANN	0.708	0.29	0.88	0.765	2	15	2	5
	SubFP-ANN	0.667	0.29	0.82	0.626	2	14	3	5
	PubChem-LR	0.792	0.43	0.94	0.889	3	16	1	4
	PubChem-RF	0.708	0.14	0.94	0.693	1	16	1	6
	Est-ANN	0.750	0.14	1.00	0.790	1	17	0	6
	MACCS-LR	0.875	0.57	1.00	0.899	4	17	0	3
	MACCS-SVM	0.875	0.71	0.94	0.958	5	16	1	2

[1] Notes: CA, classification accuracy; SE, sensitivity; SP, specificity; AUC, the area under the ROC curve; TP, the number of true positive compounds; TN, the number of true negative compounds; FP, the number of false positive compounds; FN, the number of true negative compounds.

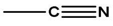
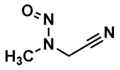
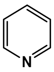
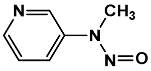
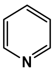
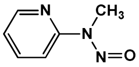
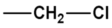
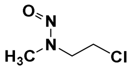
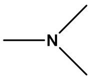
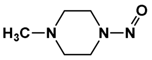
2.2.3. Performance of External Test Set

The external test set was utilized for testing the top eight models. The performance of the eight best models for test set is also shown in Table 4. The CA and AUC values ranged from 0.667 to 0.875 and 0.626 to 0.992 for external test set, respectively. Except for PubChem-RF model and the model using SubFP fingerprint, all models exhibited good predictive performance for external test set with both CA and AUC values higher than 0.7. Similar to the training set, the values of SP in these models were significantly higher than that of SE, which reflected almost perfect predictive ability for low toxic compounds in these models. Especially, the highest accuracy of 100% for low toxic compounds (SP) was obtained in MACCS-ANN, Est-ANN and MACCS-LR models. However, in all generated 28 models, only the MACCS-SVM model had good accuracy for high toxic compounds with SE value of 0.71. We supposed that the higher predictive accuracy for low toxic compounds in external test set was caused by the imbalance of high toxic compounds and low toxic compounds with a ratio of 0.292. Among these models, the MACCS-ANN model (CA = 0.792, AUC = 0.992) yielded the best performance, followed by MACCS-LR (CA = 0.875, AUC = 0.958) and MACCS-SVM (CA = 0.875, AUC = 0.899) models for the external test set. It is worth noting that the longer bits of fingerprint did not always get better results. For example, the PubChem fingerprint that contains 881 bits substructure patterns did not produce the best classification performance in this study. On the basis of our results in training and test sets, the MACCS fingerprint might be the best choice for the classification study of NNCs in terms of in vivo toxicity. Seven machine learning methods were used in this study. From the overall prediction performance, we can conclude that two algorithms, namely ANN and LR, produced the best results, in which models using ANN algorithm were slightly superior to those models using LR algorithm. As we know, LR is a widely used technique of choice for statistical modeling in which the outcome of interest is binary [35]. ANN is a type of algorithm that has great potential to execute nonlinear statistical modeling and provide a new alternative to LR, the most commonly used method for establishing predictive models for binary outcomes in medicine [35]. ANN offers a set of advantages, such as detecting complex nonlinear relationships between dependent and independent variables, detecting all possible interactions between predictor variables, requiring less formal statistical training and the availability of multiple training algorithms. We recommend that the outstanding performance of ANN in 10-fold cross-validation and external validation is because of its special algorithm [35]. In general, the prediction results showed the stable robustness and good prediction accuracy of the models.

2.2.4. Identification of Privileged Substructures as Structural Alerts

To investigate the structural features between high toxic and low toxic NNCs, the IG method and substructure frequency analysis were performed to recognize privileged substructures (fragments) in the training and external test sets based on SubFP fingerprints [33,36,37]. The higher the information gain value, the more important the substructure. These chemical features contribute to investigate the relationship between structure and the acute oral toxicity of NNCs. Details of IG values and frequencies of each fragment occurred in the high and low toxic classes are shown in Table S3 in the Supplementary Materials. From the results of the IG analysis and frequency values of privileged substructures, we found 30 substructures responsible for in vivo toxicity of NNCs. Some representative privileged substructures and known compounds containing these substructures are listed in Table 5. Among these 30 substructures, the following five substructures, namely nitrile, Hetero N nonbasic, Heteroaromatic, Alkylchloride, and Tertiary aliph amine appeared more frequently in high toxic class rather than low toxic class of NNCs (Table 5). This implies that these six substructures can be considered as structural alerts for high toxic NNCs in vivo, and then can be used as the screening alert fragments to predict potential toxicity of new potential NNCs. For example, the compounds *N*-Nitrosomethylaminoacetonitrile (**14**) containing a nitrile fragment and 2-Chloro-*N*-methyl-*N*-nitrosoethanamine (**27**) containing a chloroethyl fragment are two highly toxic agents with LD_{50} values of 45 and 22 mg/kg, respectively. It has been mentioned in a previous study that nitrile was a potentially toxic fragment [38]. Nitrile compounds (e.g., acetonitrile, acrylonitrile, and propionitrile) can release the cyanide anions through hydrolysis to exert their high toxicity [38]. The cyanide anion could affect the central nervous system and the heart by inhibiting cytochrome c oxidase. Hetero N nonbasic can be defined as an aromatic nitrogen atom having two further total connections or an aromatic nitrogen atom affording a charge of +1 with three further total connections. While another opinion suggested that hetero N and heterocycle might be only the background noise of models, or they may be parts of some toxic substructures not defined in the fingerprint [39]. The alkylchlorides are potentially alkylating agents towards DNA. In compounds containing these fragments, the electron withdrawing effect of the Cl atom increases the electrophilic character of the carbon, followed by forming carbocations and resulting in DNA damage. For example, chloroethylnitrosoureas are an important type of anticancer agents, they exert anticancer activity through chloroethylating DNA guanine and ultimately produce G–C interstrand crosslinks [40–43]. Other toxic compounds containing alkylchlorides include nitrogen mustards, epichlorohydrin, dichloromethane, dichloroethane and so on. Tertiary aliph amine compounds usually undergo metabolic activation to generate a number of oxidative products including *N*-dealkylation, ring hydroxylation, α-carbonyl formation, *N*-oxygenation, and ring opening metabolites through the formation of iminium ion intermediates [44]. Some environmental pollutants and therapeutic pharmaceuticals and their related metabolites containing a tertiary amine structure have the potential to form iminium intermediates that are reactive toward nucleophilic macromolecules, including the piperazines, piperidines and related compounds, pyrrolidines and *N*-alkyltetrahydroquinolines [44]. The substructure fragments were also analyzed by the MoSS module in KNIME [45]. The results indicated that 41 fragments were obtained for acute oral toxicity of NNCs. The detailed results are listed in Table S4 in the Supplementary Materials. Pyridine (Hetero N nonbasic) and nitrile derivatives have a larger proportion in Moss results, which is consistent with the IG results. The unique substructure characteristics detected by MoSS are imine and hydrazine fragments. Imine derivatives (Schiff base) are unstable and undergo hydrolysis to give the corresponding amine and carbonyl compounds, in which the latter (e.g., aldehydes or ketones) contain potential carbocations which act as electrophiles to form adducts with DNA. Compounds containing hydrazine fragments can be activated by endogenous substances such as metal ions or enzymes (e.g., cytochrome P450-dependent oxidases and flavin monooxygenases) to form carbocations and carbon-centered radicals, resulting in reactive radical species that cause DNA damage [33].

Table 5. Privileged substructures in compounds with high toxicity identified by information gain and frequency analysis method.

No.	Description	SMARTS	General Structures	Representative Compounds	IG	F_H
SubFP133	Nitrile	[NX1]#[CX2]			0.048	3.64
SubFP181	Hetero N nonbasic	[nX2,nX3+]			0.037	2.73
SubFP184	Heteroaromatic	[a;!c]			0.037	2.73
SubFP8	Alkylchloride	[ClX1][CX4]			0.024	3.64
SubFP26	Tertiary aliph amine	[NX3H0+0,NX4H1+;!$([N][!C]);!$([N]*~[#7,#8,#15,#16])]			0.024	3.64

3. Materials and Methods

3.1. QSAR Study

3.1.1. Data Preparation

The in vivo toxicity data of 80 NNCs were carefully collected from the US National Library of Medicine TOXNET ChemIDplus database in terms of 50% lethal dose concentration (LD$_{50}$) [46]. We selected oral LD$_{50}$ values in rats as the endpoint in this study, since most of experiments chose the oral route to estimate the toxicity [2,47]. Compounds that contain at least 1 *N*-nitroso group substituent were collected from the database. To date, this is the largest dataset that contains rodent toxicity data for NNCs as far as we know. Most regression algorithms depend on normally distributed data, so if the data are not normally distributed, a numerical transformation should be performed to obtain a normal distribution. In this study, all the original LD$_{50}$ values were converted into the corresponding $-$logLD$_{50}$ values and were used as the dependent variables in QSAR analysis. The –LogLD$_{50}$ values for the dataset range from 2.12 to 5.00, suggesting the data are adequately distributed for QSAR study. The name, CAS no. and toxicity values of NNCs are listed in Table 6.

3.1.2. Calculation of Descriptors

Quantum chemistry calculations were prevalently used in the study of QSAR modeling [48–50]. The density functional theory (DFT) level of approximation for chemistry is suitable for many applications because of the better accuracy and the relative computational efficiency [51–53]. In the present study, before calculating molecular descriptors, all chemical structures of NNCs were generated by using the Gaussview 5.0 software (Gaussian, Inc., Pittsburgh, PA, USA), and then were optimized by DFT method using the Gaussian 09 program [54] at the B3LYP functional (the standard Becke's three-parameter exchange potential and the Lee-Yang-Parr correlation functional, Gaussian, Inc., Wallingford, CT, USA) and 6-311++G(d,p) basis set. Frequency analyses on the optimized geometries ensure the geometry is an accurate saddle point rather than a transition state. A set of quantum chemical descriptors were calculated after the geometry optimization, such as dipole moment (μ), total energy (E), the highest occupied molecular orbital energy (E_{HOMO}), the lowest unoccupied molecular orbital energy (E_{LUMO}), $E_{\mathrm{LUMO}} - E_{\mathrm{HOMO}}$ gap, the bond lengths (B) and the bond angles (A). The DRAGON [55] software (version 7.0) was used to obtain the 0-2D (two-dimension) molecular descriptors. As most 3D descriptor groups encoding 3D structures were found to be sensitive to the quantum chemical calculation method [56] which can influence the accuracy of QSAR model, we therefore excluded the 3D descriptors. The total number of 0-2D descriptors was 3822. Finally, the quantum chemistry descriptors were combined with the 0-2D descriptors generated by DRAGON software to establish the QSAR models. The wide range of descriptors will facilitate the finding of hidden important variables.

Table 6. Names, CAS no. and corresponding toxicity values of *N*-nitroso compounds used in this study.

No.	Name	CAS No.	LD$_{50}$ mg/kg	Log (LD$_{50}$)$^{-1}$	Predicted log(LD$_{50}$)$^{-1}$
1	Diallylnitrosamine [a]	16338-97-9	800 (L) [b]	3.10	3.20
2	Dipentylnitrosamine	13256-06-9	1750 (L)	2.76	2.72
3	*N*-Methyl-*N*,4-dinitrosoaniline	99-80-9	1370 (L)	2.86	3.20
4	Nitroso-*N*-methyl-*N*-(2-phenyl) ethylamine	13256-11-6	48 (H) [b]	4.32	4.30
5	*N*-Nitroso(2,2,2-trifluoroethyl)ethylamine [a]	82018-90-4	960 (L)	3.02	3.20
6	Nitrosodibutylamine	924-16-3	1200 (L) [b]	2.92	3.17
7	*N*-Nitrosodipropylamine	621-64-7	480 (L) [b]	3.32	3.25
8	Nitrosoethylmethylamine	10595-95-6	90 (H) [b]	4.05	4.34
9	2-Nitrosomethylaminopyridine [a]	16219-98-0	60 (H)	4.22	4.23
10	Nitrosomethylaniline	614-00-6	225 (L)	3.65	4.15
11	Diisopropylnitrosamine	601-77-4	850 (L)	3.07	2.68
12	*N*-Nitrosobis(2,2,2-trifluoro ethyl)amine	625-89-8	300 (L)	3.52	3.46
13	*N*-Ethyl-*N*-*tert*-butylnitrosamine	3398-69-4	1600 (L) [b]	2.80	2.71

Table 6. *Cont.*

No.	Name	CAS No.	LD_{50} mg/kg	Log $(LD_{50})^{-1}$	Predicted $\log(LD_{50})^{-1}$
14	*N*-Nitrosomethylaminoacetonitrile	3684-97-7	45 (H)	4.35	4.25
15	*N*-Butyl-*N*-(4-hydroxybutyl) nitro samine	3817-11-6	1800 (L) [b]	2.74	2.34
16	*N*-Nitrosomethylvinylamine	4549-40-0	24 (H)	4.62	4.51
17	*N*-Nitroso-*N*-methylallylamine	4549-43-3	340 (L)	3.47	3.55
18	*N*-Ethyl-*N*-butylnitrosamine	4549-44-4	380 (L) [b]	3.42	3.71
19	*N*-Nitrosodibenzylamine	5336-53-8	900 (L)	3.05	2.92
20	*N*-Nitroso-*N*-methylcyclohexylamine [a]	5432-28-0	30 (H) [b]	4.52	3.92
21	Nitrosomethyl-n-butylamine	7068-83-9	130 (H)	3.89	3.99
22	*N*-Ethyl-*N*-hydroxyethylnitrosamine	13147-25-6	7500 (L)	2.12	2.53
23	*N*-Amyl-*N*-methylnitrosamine	13256-07-0	120 (H)	3.92	3.85
24	Dinitrosodimethylethylenediamine	13256-12-7	125 (H) [b]	3.90	3.90
25	Vinylethylnitrosamine	13256-13-8	88 (H)	4.06	3.68
26	*N*-Nitrososarcosine	13256-22-9	5000 (L)	2.30	2.68
27	2-Chloro-*N*-methyl-*N*-nitrosoethanamine	16339-16-5	22 (H)	4.66	4.10
28	*N*-Methyl(methoxymethyl)nitrosamine	39885-14-8	700 (L)	3.15	3.34
29	Methyl(acetoxymethyl)nitrosamine [a]	56856-83-8	130 (H)	3.89	3.83
30	Acetoxymethylbutylnitrosamine [a]	56986-36-8	1500 (L)	2.82	2.89
31	1-Methoxy-ethyl-ethylnitrosamine	61738-03-2	1000 (L) [b]	3.00	2.84
32	Methoxymethyl-ethylnitrosamine	61738-04-3	540 (L)	3.27	3.12
33	1-Methoxy-ethyl-methylnitrosamine	61738-05-4	240 (L)	3.62	3.35
34	Acetoxymethylpropylnitrosamine	66017-91-2	1000 (L)	3.00	3.05
35	Methyl(butyroxymethyl)nitrosamine	67557-56-6	800 (L) [b]	3.10	3.20
36	Acetoxymethyltrideuteromethylnitrosamine	67557-57-7	120 (H)	3.92	3.88
37	*N*-Nitroso-*N*-phenylhydroxylamine	148-97-0	490 (L) [b]	3.31	3.53
38	*N*-methyl-n-benzylnitrosamine	937-40-6	18 (H) [b]	4.74	4.22
39	4-(Methylnitrosoamino)benzaldehyde [a]	7431-19-8	2000 (L)	2.70	2.76
40	3-(*N*-Nitrosomethylamino)sulfolan	13256-21-8	750 (L)	3.12	2.88
41	Aethyl-4-picolylnitrosamin	13256-23-0	40 (H)	4.40	4.02
42	*N*,*N*′-Dimethylnitrosourea	13256-32-1	280 (L)	3.55	3.50
43	*N*-Nitrososarcosine ethyl ester	13344-50-8	4000 (L)	2.40	2.65
44	4-Nitrosomethylaminopyridine	16219-99-1	200 (L)	3.70	3.81
45	*N*-Nitrosoethylisopropylamine	16339-04-1	1100 (L) [b]	2.96	3.19
46	*N*-Nitrosotrimethylhydrazine	16339-14-3	95 (H) [b]	4.02	4.05
47	*N*-Nitrosodiacetonitrile	16339-18-7	163 (H)	3.79	3.89
48	*N*-Nitroso-*N*-ethylbenzylamine	20689-96-7	250 (L) [b]	3.60	3.66
49	*N*-Nitroso-*O*,*N*-diethylhydroxylamine	56235-95-1	1000 (L) [b]	3.00	2.79
50	*N*-Nitroso-*N*-(2-methylbenzyl)methylamine	62783-48-6	90 (H)	4.05	3.96
51	*N*-Methyl-*N*-nitroso-(3-methylphenyl)methylamine	62783-49-7	600 (L)	3.22	3.41
52	*N*-Methyl-*N*-nitroso-(4-methylphenyl)methylamine	62783-50-0	400 (L) [b]	3.40	3.89
53 [c]	*N*-Nitroso-*N*-methyl-1(1-phenyl)-ethylamine [a]	68690-89-1	600 (L)	3.22	4.00
54	*N*-Nitroso-*N*-methyl-2-(2-phenyl)-propylamine	68690-90-4	2100 (L)	2.68	2.82
55	3-Nitrosomethylaminopyridine	69658-91-9	10 (H)	5.00	4.40
56	*N*-Nitrosodiethylamine [a]	55-18-5	220 (L)	3.66	3.62
57	*N*-Nitrosodimethylamine	62-75-9	37 (H)	4.43	4.53
58	*N*-Nitrosodiphenylamine	86-30-6	1825 (L)	2.74	2.74
59	*N*-Nitroso-3,6-dihydro-1,2-oxazine	3276-41-3	900 (L)	3.05	3.05
60	*R*(−)-*N*-Nitroso-2-methylpiperidine	14026-03-0	600 (L)	3.22	2.94
61	*S*(+)-*N*-Nitroso-2-methylpiperidine	36702-44-0	600 (L)	3.22	3.00
62	*N*-Nitrosoheptamethyleneimine [a]	20917-49-1	283 (L)	3.55	3.58
63	*N*-Nitrosomorpholine	59-89-2	282 (L)	3.55	3.23
64	*N*-Nitrosopyrrolidine	930-55-2	900 (L)	3.05	3.35
65	1-Nitrosopiperazine	5632-47-3	2260 (L)	2.65	3.39
66	*N*-Nitrosopiperidine	100-75-4	200 (L)	3.70	3.39
67	*N*-Nitroso-tetrahydro-1,2-oxazine	40548-68-3	830 (L) [b]	3.08	2.95
68	*N*-Nitrosoperhydroazepine [a]	932-83-2	336 (L)	3.47	3.51
69	*N*-Nitrosoindoline	7633-57-0	320 (L)	3.49	3.40
70	*N*-Nitroso-*N*′-methylpiperazine [a]	16339-07-4	100 (H) [b]	4.00	3.51
71	*N*-Nitrosoazacyclononane	20917-50-4	566 (L) [b]	3.25	3.40
72	3-Nitrosotetrahydro-1,3-oxazine	35627-29-3	600 (L)	3.22	3.29
73	*N*-Nitroso-1,3-oxazolidine	39884-52-1	1500 (L)	2.82	2.92
74	1-Amyl-1-nitrosourea [a]	10589-74-9	560 (L)	3.25	3.30
75	*N*-Nitroso-*N*-butylurea	869-01-2	400 (L) [b]	3.40	3.49
76	*N*-Nitroso-*N*-ethylurea	759-73-9	300 (L)	3.52	3.46
77	*N*-Nitroso-*N*-methylurea	684-93-5	110 (H)	3.96	4.27
78	Propylnitrosourea	816-57-9	480 (L)	3.32	3.13
79	*N*-Nitroso-*N*-methylbiuret [a]	13860-69-0	450 (L) [b]	3.35	3.73
80	Ethylnitrosobiuret [a]	32976-88-8	1050 (L)	2.98	3.53

[a] Test set in QSAR study; [b] Test set in classification study; [c] Outlier in the best GA-MLR-based QSAR model.

3.1.3. QSAR Modeling and Model Evaluation

QSARINS 2.2.2 software (Varese, Italy) [57,58] was used to develop QSAR models by means of GA and MLR methods. After all types of molecular descriptors were generated, we performed the pre-filtration prior to modeling. The constant or near-constant values (>80%) and the highly inter-correlated descriptors (>95%) were eliminated due to statistical insignificance. All the compounds were ranked according to the toxicity value ($-\log LD_{50}$), then one was selected as the test set every five compounds, and the remaining compounds were used as the training set. A training set was used for constructing QSAR models, whereas a test set was used for evaluating the external predictive ability of the models. All subsets and GA tools of QSARINS 2.2.2 software were utilized for descriptor selection. First, all low-dimensional models (up to 2–3 descriptors) were calculated using the all subset facility to gain an insight into the best descriptors encoding the effect and to avoid a completely random start of the GA. The core of chromosomes of the initial population for the GA was the best subset of descriptors determined at this step. Then, GA was utilized to detect the solution space by maximizing the leave-one-out (LOO) cross-validation correlation coefficient (Q^2_{loo}) as the fitness function. To obtain the best variables, the population size, mutation rate and number of generations were set as 200, 20 and 2000, respectively [23,56]. Q^2_{loo} was chosen as it provides a measurement of model stability and robustness. Following this procedure repeatedly, a population of good models was generated.

The statistical quality and internal predictive ability of QSAR models were evaluated using the coefficient of determination R^2 and modified form R^2_{adj}, root mean square error (RMSE) and Q^2_{loo}. The QUIK rule (Q Under Influence of K) [59] was used to test the inter-correlation among descriptors and was set to 0.05 to eliminate models with high multicollinearity. The external predictive ability of the models was assessed through the test set and evaluated by Q^2_{ext}, $Q^2_{ext} = 1 - \text{PRESS}/\text{SD}$, where PRESS is the sum of squared deviations between the experimental values and the predicted value for each molecule in the test set, and SD is the sum of squared deviations between the experimental values of the test set molecules and the mean experimental value of the training set molecules [25]. Q^2_{F1} [60], Q^2_{F2} [61], Q^2_{F3} [62,63], Concordance Correlation Coefficient (CCC) [64,65], CCC_{ext} [66,67] and $RMSE_{ext}$ are also involved. A Y-scrambling procedure (2000 iterations to check the fitting of the randomly reordered Y-data) was also performed to evaluate the possibility of the chance correlation in the QSAR models. The dependent variables ($-\text{Log}LD_{50}$) were randomly shuffled and new QSAR models were established using the original independent variable matrix. If the QSAR model obtained by shuffling the $-\text{Log}LD_{50}$ values gave significantly lower coefficients of determination than the original model, we considered that the proposed QSAR model was not obtained casually. These parameters were calculated according to the following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3}$$

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right) \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{5}$$

$$Q^2_{loo} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{6}$$

where y_i and $\overline{y}$ are the actual and average activities and $\hat{y}_i$ are predictive activities.

The Multi-Criteria Decision Making (MCDM) method included in QSARINS 2.2.2 software was used to summarize the model performances relevant to internal and external validations as scores [56,57]. The scores range from 0 to 1, where 0 and 1 represent the worst and the best validation criteria, respectively. After numerous rounds of trials, models were finally selected with the best

MCDM score, fulfilling the statistical thresholds for fitting, internal and external validation, and with the least possible number of descriptors [66,68].

3.1.4. Application Domain

To consider the scope and limitations of the proposed models, the applicability domain (AD) was considered. In other words, the AD describes the range of chemical structures for which the models are considered to be applicable. The predicted values are reliable only for those compounds fall on the AD. The AD of each model was evaluated by the leverage approach [69]. Williams plot, which is a plot of standardized cross-validated residuals versus leverages (hat values, h), was used to visualize the outliers in both the structural and the response spaces. The critical hat value of structural threshold was set as $h^* = 3(p + 1)/n$, where p is the number of descriptors of the model and n is the number of training compounds. If $h > h^*$, a compound will be identified as an outlier. For the training set, compounds with $h > h^*$ seriously affect the statistical parameters of models, so they were removed, and the model was calibrated again. For the test set, if compounds are observed with $h > h^*$, their predicted values were unreliable. A critical value of 3 for the standardized residual in response space is usually used to identify statistical outliers. Response outliers in MLR models were identified if its predicted value is higher than ±3 standardized residuals.

3.2. Classification Study

3.2.1. Data Preparation

Before the classification study, we conducted a preliminary test to determinate the classification criterion. The same dataset of 80 NNCs used in QSAR studies was divided into three different levels of toxicity (50, 100, and 200 mg/kg, respectively). The results obtained from the preliminary test indicated that a toxic level with 200 mg/kg as the classification criterion had the best performance of classification. Finally, a dataset containing 22 compounds with high toxicity and 58 compounds with low toxicity was obtained. All these compounds were then randomly divided into a training set and a test set with a ratio of 7:3. A complete list of the compounds' classification is presented in Table 6.

3.2.2. Molecular Fingerprints

Molecular fingerprints are developed to describe chemical structures in a chemical database and widely used in similarity searching and classification. Therefore, substructure features in each fingerprint dictionary are defined to cover full of representative substructures. In this case, a molecule was described as a binary string of structural keys. SMiles Arbitrary Target Specification (SMARTS) is a language used for describing molecular patterns and properties using rules that are extensions of simplified molecular input line entry specification (SMILES) [70]. Different substructure patterns with SMARTS lists were predefined in a dictionary. For a SMARTS pattern, if a substructure existed in the given molecule, the corresponding bit was set to "1" and otherwise set to "0" [70]. Four fingerprints were used in our study, including the Estate fingerprint (Est, 79 bits), MACCS keys (166 bits), PubChem fingerprints (881 bits), and Substructure fingerprint (SubFP, 307 bits). All these four fingerprints were calculated by the PaDEL-Descriptor program [71].

3.2.3. Machine Learning Methods

Seven machine learning methods were used to build the classification models. They are *k*-nearest neighbor (*k*NN), Logistic Regression (LR), Naïve Bayes (NB), Artificial Neural Network (ANN), Random Forest (RF), Support vector machine (SVM), and Tree. The seven methods were performed using Orange Canvas 3.11 software (freely available at https://orange.biolab.si/).

k-nearest neighbor (*k*NN): *k*NN is a nonparametric method to classify objects based on nearest training samples in the feature space. For each test sample Z = (x′, y′), the list of its nearest neighbor was determined by the algorithm calculated the distance or similarity between each training example (x, y) [72]. After that it can be classified on the basis of the majority of the nearest neighbors. In order to reduce the impact of k (the number of nearest neighbors) value, a distance-weighted method was utilized. In this study, we chose the Euclidean distance and distance-weighted parameters and the k value was set to 5.

Logistic Regression (LR): LR was developed by statistician David Cox in 1958 [73,74], which has usually been applied to a binary dependent variable. The two possible dependent variable values can be labeled as symbols of "0" and "1", which represent results such as pass/fail, win/lose, alive/dead or yes/no, respectively.

Naïve Bayes (NB): The NB classifier method is a simple classification method based on the Bayes rule for the conditional probability [75]. This method allows users to categorize compounds in a data set based on the equal and independent contribution of their attributes. The prior probability can be directly estimated from the training set since it is the same to all of the classes, while the marginal probability is ignored. In this study, the default settings in Orange were applied to perform the NB classification.

Artificial Neural Network (ANN): ANN has become a prevalent method which can be used for identifying complex nonlinear relationship for classification and regression [76]. The network consisted of three layers containing one input layer, one hidden layer, and one output layer. The ANN method in Orange 3.11 is a multi-layer perceptron (MLP) algorithm with backpropagation. In this work, the number of neurons per hidden layer was set to 200, and the rectified linear unit function (ReLu) was chosen as activation function for the hidden layer.

Random Forest (RF): RF was developed by Breiman, which is an ensemble learning method for classification and regression [77]. The forest is assembled by trees. Each tree is developed from a bootstrap sample from the training set. The tree grows up to maximum size without pruning. When developing individual trees, an arbitrary subset of attributes is achieved (hence the term "Random"), from which the best attribute for the split is selected. The final model is based on the majority of individually developed trees in the forest. The number of trees in the forest was set to 20.

Support vector machine (SVM): SVM is a machine learning technique that separates the attribute space with a hyperplane, thus maximizing the margin between the instances of different classes or class values. It was first developed by Vapnik and co-workers in 1995, which is a kernel-based algorithm for binary data classification and regression [78]. Polynomial kernel, Gaussian radial basis function kernel (RBF) and sigmoid kernel are the generally used kernel functions. The penalty coefficient C and slack variable γ should be introduced to make a compromise between linear separability and maximal margin. In this study, the RBF kernel was chosen, and the parameters C and γ were tuned on the training set by 10-fold cross-validation. Orange embeds a popular implementation of SVM from the LIBSVM package [79]. The linear function was chosen, and the cost was set to 1.00.

Tree: Tree is a simple algorithm that splits the data into nodes by class purity. It is a precursor to RF. Tree in Orange is designed in-house and can handle both discrete and continuous datasets. It includes decision nodes, branches, and leaves. A decision tree inputs an object or situation described by a number of properties and outputs a yes/no decision. An instance is classified by beginning at the root node of the decision tree, testing the attribute specified by this node, followed by moving down to the tree branch according to the value of the attribute [80]. In the pre-pruning process, the minimal instance in leaves is 3, and stops splitting nodes with fewer instances than 5. Other parameters of tree were used with the default values in Orange.

3.2.4. Performance Evaluation

The 10-fold cross-validation and test set were used to evaluate the performance of all the established models. For 10-fold cross-validation, the training set was further divided in to ten subsets, nine of which were chosen as training sets and one subset as a test set in each run. After ten runs, each subset was used as a test set and the entire dataset was predicted. All models were evaluated by counting the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative

(FN) compounds. Further, the classification accuracy (CA), sensitivity (SE), and specificity (SP) were also calculated by the following equations:

$$\mathrm{CA} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \tag{7}$$

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{8}$$

$$\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \tag{9}$$

The CA is the total percentage of both high toxic and low toxic compounds that were correctly predicted. The SE is the predictive accuracy of the high toxic compounds and the SP means the predictive accuracy of low toxicity. Further, the receiver operating characteristic (ROC) curve where the TP rate (or sensitivity) against the FP rate (1-specificity) was plotted. The area under the ROC curve (AUC) was also calculated. The values of AUC range from 0.5 to 1.0 [81], where 1 indicates a perfect classifier, 0.5 means the classifier has no discriminative power.

3.2.5. Analysis of Privileged Substructures

The information gain (IG) [70] and substructure fragment analysis [82,83] were used to identify the privileged substructure fragments and the structural alerts. If a substructure was more frequently presented in the class of compounds with high toxicity, this substructure could be regarded as a privileged substructure involved in chemical toxicity. The frequency of a fragment in high toxic compounds was defined as follows:

$$\text{Frequency of a fragment} = \frac{N^{H}_{fragment} \times N_{total}}{N_{fragment\_total} \times N_{H}} \tag{10}$$

where $N^{H}_{fragment}$ is the number of compounds containing the fragment in the class of high toxic compounds; N_{total} is the total number of compounds; $N^{H}_{fragment\_total}$ is the total number of compounds containing the fragment; and N_{H} is the number of high toxicity compounds.

In addition, the MoSS module in KNIME (available online: http://www.knime.org/) was also used to search for substructure fragments that are frequently presented in a set of molecules. In the MoSS module, the “minimum fragment size” and “minimum focus support in %” values are important for fragment search. In our study, the two values were finally set to 4 and 3, respectively.

4. Conclusions

In this study, we developed the QSAR and classification models of a large set of 80 NNCs with their rat acute oral toxicity. All QSAR models were established by GA-MLR methods. A reasonable correlation ($Q^2{}_{\mathrm{loo}}$ = 0.7533, R^2 = 0.8071, $Q^2{}_{\mathrm{ext}}$ = 0.7041, $R^2{}_{\mathrm{ext}}$ = 0.7195) was obtained between experimental and predicted toxicity values for the NNCs studied in the best QSAR model with eight molecular descriptors. The robustness and fitting goodness of QSAR models were evaluated using LOO cross-validation, while the test set was used to assess the external predictive power. The QUIK rule was used to eliminate models with high predictor collinearity. The possibility of chance correlation of the best model was checked by a Y-scrambling procedure. All the classification models were obtained by four molecular fingerprints (Est, MACCS, PubChem and SubFP) combined with seven machine learning methods (*k*NN, LR, NB, ANN, RF, SVM and Tree). All these models were examined by 10-fold cross-validation and external test sets to evaluate their internal and external predictive performance. The best classification model was the MACCS-ANN model with Q and AUC values of 0.821, 0.905 and 0.792, 0.992 for the training set and external test set, respectively. Analysis of privileged substructures performed by IG and frequency analysis methods can identify some substructures (fragments) as structural alerts for acute oral toxicity of NNCs. The substructures were further tested and verified by MoSS analysis. From the results of GA-MLR-based QSAR and classification models, we can conclude that the polarizability, IP, the presence/absence and frequency of C-O bond, Nitrile, Hetero N nonbasic, Alkylchloride, Tertiary aliph amine can be regarded as main attributes for assessing in vivo toxicity of NNCs. We believe that the models we developed reflect major contributions to our knowledge of the toxicity of NNCs. Compared with GA-MLR-based QSAR models, the semi-quantitative classification models could determine toxic severity of compounds with high accuracy directly. All the proposed models can provide useful insights into the structural features responsible for the acute oral toxicity of NNCs and therefore could help to improve our understanding of the toxicity mechanisms in vivo for this class of compounds. In summary, our study not only provides useful tools for predicting the in vivo toxicity of NNCs quantitatively or semi-quantitatively, but is also helpful to estimate acute toxicity in assessment of environmental safety.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/19/10/3015/s1.

Author Contributions: T.F., G.S. and L.Z. conceived and designed the experiments; T.F. and G.S. performed the experiments and analyzed the data; L.Z., X.C. and R.Z. contributed analysis tools and helped in the "Results and Discussion Section"; T.F., G.S. and L.Z. wrote the paper. All authors read and approved the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (No. 21778011), Beijing Natural Science Foundation (Nos. 7184192 and 7162015), China Postdoctoral Science Foundation funded project (No. 2017M620567), Beijing Postdoctoral Research Foundation (No. 2018-ZZ-022), Education Commission Science and Technology Project of Beijing Municipality (No. PXM2015_014204_500175).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lijinsky, W. *N*-nitroso compounds in the diet. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* **1999**, *443*, 129–138. [CrossRef]
2. Druckrey, H.; Preussmann, R.; Ivankovic, S.; Schmahl, D. Organotropic carcinogenic effects of 65 various *N*-nitroso- compounds on BD rats. *Z. Krebsforsch.* **1967**, *69*, 103–201. [CrossRef] [PubMed]
3. Lijinsky, W.; Andrews, A.W. *N*-nitrosamine mutagenicity using the Salmonella/Mammalian-microsome mutagenicity assay. In *Genotoxicology of N-Nitroso Compounds*; Rao, T.K., Lijinsky, W., Epler, J.L., Eds.; Plenum Press: New York, NY, USA, 1984; pp. 13–43.
4. Lijinsky, W. Structure-activity relations in carcinogenesis by *N*-nitroso compounds. *Cancer Metast. Rev.* **1987**, *6*, 301–356. [CrossRef]
5. Bartsch, H.; Oshima, H.; Pignatelli, B.; Malaveille, C.; Friesen, M. Nitrite-reactive phenols present in smoked foods and amino-sugars formed by the Maillard reaction as precursors of genotoxic arenediazonium ions or nitroso compounds. In *Mutagens in Food: Detection and Prevention*; Hikoya, H., Ed.; CRC Press: Boca Raton, FL, USA, 1991; pp. 87–100.
6. Lijinsky, W. *Chemistry and Biology of N-Nitroso Compounds*; Cambridge University Press: Cambridge, UK, 1992.
7. Yuan, J.T.; Pu, Y.P.; Yin, L.H. Predicting carcinogenicity and understanding the carcinogenic mechanism of *N*-nitroso compounds using a TOPS-MODE approach. *Chem. Res. Toxicol.* **2011**, *24*, 2269–2279. [CrossRef] [PubMed]
8. Laires, A.; Gaspar, J.; Borba, H.; Proenca, M.; Monteiro, M.; Rueff, J. Genotoxicity of nitrosated red wine and of the nitrosatable phenolic-compounds present in wine-tyramine, quercetin and malvidine-3-glucoside. *Food Chem. Toxicol.* **1993**, *31*, 989–994. [CrossRef]
9. Gaspar, J.; Laires, A.; Va, S.; Pereira, S.; Mariano, A.; Quina, M.; Rueff, J. Mutagenic activity of glycine upon nitrosation in the presence of chloride and human gastric juice: A possible role in gastric carcinogenesis. *Teratog. Carcinog. Mutagen.* **1996**, *16*, 275–286. [CrossRef]
10. Duarte, M.P.; Laires, A.; Gaspar, J.; Oliveira, J.S.; Rueff, J. Genotoxicity of instant coffee and of some phenolic compounds present in coffee upon nitrosation. *Teratog. Carcinog. Mutagen.* **2000**, *20*, 241–249. [CrossRef]
11. Bartsch, H.; Ohshima, H.; Pignatelli, B. Inhibitors of endogenous nitrosation—Mechanisms and implications in human cancer prevention. *Mutat. Res.* **1988**, *202*, 307–324. [CrossRef]

12. Tratnyek, P.G.; Bylaska, E.J.; Weber, E.J. In silico environmental chemical science: Properties and processes from statistical and computational modelling. *Environ. Sci. Process. Impacts* **2017**, *19*, 188–202. [CrossRef] [PubMed]
13. Card, M.L.; Gomez-Alvarez, V.; Lee, W.; Lynch, D.G.; Orentas, N.S.; Lee, M.T.; Wong, E.M.; Boethling, R.S. History of EPI Suite((TM)) and future perspectives on chemical property estimation in US Toxic Substances Control Act new chemical risk assessments. *Environ. Sci. Process. Impacts* **2017**, *19*, 203–212. [CrossRef] [PubMed]
14. Cronin, M.; Walker, J.D.; Jaworska, J.S.; Comber, M.; Watts, C.D.; Worth, A.P. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environ. Health Perspect.* **2003**, *111*, 1376–1390. [CrossRef] [PubMed]
15. Combes, R.; Grindon, C.; Cronin, M.; Roberts, D.W.; Garrod, J.F. Integrated decision-tree testing strategies for acute systemic toxicity and toxicokinetics with respect to the requirements of the EU REACH legislation. *Atla-Altern. Lab. Anim.* **2008**, *36*, 45–63.
16. Cronin, M.; Madden, J.; Enoch, S.; Roberts, D. Evaluation of categories and read-across for toxicity prediction allowing for regulatory acceptance. In *Chemical Toxicity Prediction: Category Formation and ReadAcross*; The Royal Society of Chemistry: Cambridge, UK, 2013; pp. 155–167.
17. Cronin, M.T. (Q)SARs to predict environmental toxicities: Current status and future needs. *Environ. Sci. Process. Impacts* **2017**, *19*, 213–220. [CrossRef] [PubMed]
18. Dai, Q.H.; Zhong, R.G.; Gao, X.M. Pattern recognition data for structure-carcinogenic activity relationship of *N*-nitroso compounds based upon di-region theory. *Environ. Chem.* **1987**, *6*, 1–12.
19. Luan, F.; Zhang, R.S.; Zhao, C.Y.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Classification of the carcinogenicity of *N*-nitroso compounds based on support vector machines and linear discriminant analysis. *Chem. Res. Toxicol.* **2005**, *18*, 198–203. [CrossRef] [PubMed]
20. Helguera, A.M.; Gonzalez, M.P.; Dias Soeiro Cordeiro, M.N.; Cabrera Perez, M.A. Quantitative structure—Carcinogenicity relationship for detecting structural alerts in nitroso compounds: Species, rat; Sex, female; Route of administration, Gavage. *Chem. Res. Toxicol.* **2008**, *21*, 633–642. [CrossRef] [PubMed]
21. Helguera, A.M.; Cordeiro, M.N.D.S.; Pérez, M.Á.C.; Combes, R.D.; González, M.P. Quantitative structure carcinogenicity relationship for detecting structural alerts in nitroso-compounds☆Species: Rat; Sex: Male; Route of administration: Water. *Toxicol. Appl. Pharm.* **2008**, *231*, 197–207. [CrossRef] [PubMed]
22. Rogers, D.; Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866. [CrossRef]
23. Wu, X.; Zhang, Q.; Hu, J. QSAR study of the acute toxicity to fathead minnow based on a large dataset. *SAR QSAR Environ. Res.* **2016**, *27*, 147–164. [CrossRef] [PubMed]
24. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **2010**, *29*, 476–488. [CrossRef] [PubMed]
25. Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graph. Model.* **2002**, *20*, 269–276. [CrossRef]
26. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2008; Volume 11.
27. Cassotti, M.; Ballabio, D.; Consonni, V.; Mauri, A.; Tetko, I.V.; Todeschini, R. Prediction of acute aquatic toxicity toward daphnia magna by using the GA-*k*NN method. *Atla-Altern. Lab. Anim.* **2014**, *42*, 31–41.
28. Zhang, J.; Ji, L.; Liu, W. In Silico Prediction of Cytochrome P450-mediated biotransformations of xenobiotics: A case study of epoxidation. *Chem. Res. Toxicol.* **2015**, *28*, 1522–1531. [CrossRef] [PubMed]
29. Galvez, J.; Garciadomenech, R.; Dejulianortiz, V.; Soler, R. Topological approach to analgesia. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1198–1203. [CrossRef] [PubMed]
30. Galvez, J.; Garciadomenech, R.; Dejulianortiz, J.V.; Soler, R. Topological approach to drug design. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272–284. [CrossRef] [PubMed]
31. Roberto, T.; Consonni, V. Molecular Descriptors for Chemoinformatics. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., KubiAnyi, H., Folkers, G., Eds.; WILEY-VCH: Weinheim, Germany, 2009; pp. 617–618.
32. Ghose, A.K.; Crippen, G.M. Atomic physicochemical parameters for 3-dimensional structure-directed quantitative structure-activity-relationships I. Partition-coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577. [CrossRef]

33. Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P.W.; Tang, Y. In silico prediction of chemical Ames mutagenicity. *J. Chem. Inf. Model.* **2012**, *52*, 2840–2847. [CrossRef] [PubMed]
34. Du, H.; Cai, Y.; Yang, H.; Zhang, H.; Xue, Y.; Liu, G.; Tang, Y.; Li, W. In silico prediction of chemicals binding to aromatase with machine learning methods. *Chem. Res. Toxicol.* **2017**, *30*, 1209–1218. [CrossRef] [PubMed]
35. Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **1996**, *49*, 1225–1231. [CrossRef]
36. Li, X.; Chen, L.; Cheng, F.; Wu, Z.; Bian, H.; Xu, C.; Li, W.; Liu, G.; Shen, X.; Tang, Y. In silico prediction of chemical acute oral toxicity using multi-classification methods. *J. Chem. Inf. Model.* **2014**, *54*, 1061–1069. [CrossRef] [PubMed]
37. Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P.W.; Tang, Y. In silico prediction of drug induced liver toxicity using substructure pattern recognition method. *Mol. Inform.* **2016**, *35*, 136–144. [CrossRef] [PubMed]
38. Bhattacharya, R.; Satpute, R.M.; Hariharakrishnan, J.; Tripathi, H.; Saxena, P.B. Acute toxicity of some synthetic cyanogens in rats and their response to oral treatment with alpha-ketoglutarate. *Food Chem. Toxicol.* **2009**, *47*, 2314–2320. [CrossRef] [PubMed]
39. Lei, T.; Li, Y.; Song, Y.; Li, D.; Sun, H.; Hou, T. ADMET evaluation in drug discovery: 15. Accurate prediction of rat oral acute toxicity using relevance vector machine and consensus modeling. *J. Cheminform.* **2016**, *8*, 1–19. [CrossRef] [PubMed]
40. Sun, G.H.; Zhao, L.J.; Fan, T.J.; Li, S.S.; Zhong, R.G. Investigations on the effect of O-6-benzylguanine on the formation of dG-dC interstrand cross-links induced by chloroethylnitrosoureas in human glioma cells using stable isotope dilution high-performance liquid chromatography electrospray ionization tandem mass spectrometry. *Chem. Res. Toxicol.* **2014**, *27*, 1253–1262. [PubMed]
41. Sun, G.H.; Zhang, N.; Zhao, L.J.; Fan, T.J.; Zhang, S.F.; Zhong, R.G. Synthesis and antitumor activity evaluation of a novel combi-nitrosourea prodrug: Designed to release a DNA cross-linking agent and an inhibitor of O^6-alkylguanine-DNA alkyltransferase. *Bioorg. Med. Chem.* **2016**, *24*, 2097–2107. [CrossRef] [PubMed]
42. Sun, G.; Fan, T.; Zhao, L.; Zhou, Y.; Zhong, R. The potential of combi-molecules with DNA-damaging function as anticancer agents. *Future Med. Chem.* **2017**, *9*, 403–435. [CrossRef] [PubMed]
43. Sun, G.H.; Zhao, L.J.; Zhong, R.G.; Peng, Y.Z. The specific role of O6-methylguanine-DNA methyltransferase inhibitors in cancer chemotherapy. *Future Med. Chem.* **2018**, *16*, 1971–1996. [CrossRef] [PubMed]
44. Ma, I.; Peterlin, L. Role of cyclic tertiary amine bioactivation to reactive iminium species: Structure toxicity relationship. *Curr. Drug Metab.* **2011**, *12*, 35–50.
45. Borgelt, C.; Meinl, T.; Berthold, M. MoSS: A program for molecular substructure mining. In Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations, Chicago, IL, USA, 21 August 2005.
46. TOXNET-ChemIDplus. Available online: https://chem.nlm.nih.gov/chemidplus/ (accessed on 3 June 2017).
47. Schweinsberg, F.; Schottkollat, P.; Burkle, G. Change of toxicity and carcinogenicity of *N*-methyl-*N*-nitrosobenzylamine in rats by methylsubstitution in phenylresidue. *Z. Krebsforsch.* **1977**, *88*, 231–236. [CrossRef]
48. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027–1043. [CrossRef] [PubMed]
49. Karabulut, S.; Sizochenko, N.; Orhan, A.; Leszczynski, J. A DFT-based QSAR study on inhibition of human dihydrofolate reductase. *J. Mol. Graph. Model.* **2016**, *70*, 23–29. [CrossRef] [PubMed]
50. Cheng, Y.; Luo, F.; Zeng, Z.; Wen, L.; Xiao, Z.; Bu, H.; Lv, F.; Xu, Z.; Lin, Q. DFT-based quantitative structure-activity relationship studies for antioxidant peptides. *Struct. Chem.* **2015**, *26*, 739–747. [CrossRef]
51. Nendza, M.; Mueller, M.; Wenzel, A. Classification of baseline toxicants for QSAR predictions to replace fish acute toxicity studies. *Environ. Sci. Process. Impacts* **2017**, *19*, 429–437. [CrossRef] [PubMed]
52. Enoch, S.J.; Cronin, M.T.D.; Schultz, T.W.; Madden, J.C. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition. *Chem. Res. Toxicol.* **2008**, *21*, 513–520. [CrossRef] [PubMed]
53. Pasha, F.A.; Muddassar, M.; Beg, Y.; Cho, S.J. DFT-based de novo QSAR of phenoloxidase inhibitors. *Chem. Biol. Drug Des.* **2008**, *71*, 483–493. [CrossRef] [PubMed]
54. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; et al. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, USA, 2009.

55. Kode Srl. Dragon (Software for Molecular Descriptor Calculation) V 7.0.6. Available online: https://chm.kode-solutions.net/ (accessed on 3 September 2017).
56. Onlu, S.; Turker, S.M. Impact of geometry optimization methods on QSAR modelling: A case study for predicting human serum albumin binding affinity. *SAR QSAR Environ. Res.* **2017**, *28*, 491–509. [CrossRef] [PubMed]
57. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J. Comput. Chem.* **2013**, *34*, 2121–2132. [CrossRef]
58. Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *J. Comput. Chem.* **2014**, *35*, 1036–1044. [CrossRef] [PubMed]
59. Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index: Theory development and its application in chemometrics. *Chemometr. Intell. Lab. Syst.* **1999**, *46*, 13–29. [CrossRef]
60. Shi, L.M.; Fang, H.; Tong, W.D.; Wu, J.; Perkins, R.; Blair, R.M.; Branham, W.S.; Dial, S.L.; Moland, C.I.; Sheehan, D.M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186–195. [CrossRef] [PubMed]
61. Schueuermann, G.; Ebert, R.; Chen, J.; Wang, B.; Kuehne, R. External validation and prediction employing the predictive squared correlation coefficient-test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140–2145. [CrossRef] [PubMed]
62. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the q(2) parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678. [CrossRef] [PubMed]
63. Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemometr.* **2010**, *24*, 194–201. [CrossRef]
64. Lin, L. A Concordance correlation-coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255–268. [CrossRef] [PubMed]
65. Lin, L. Assay validation using the concordance correlation-coefficient. *Biometrics* **1992**, *48*, 599–604. [CrossRef]
66. Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320–2335. [CrossRef] [PubMed]
67. Chirico, N.; Gramatica, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044–2058. [CrossRef] [PubMed]
68. Roy, K.; Das, R.N.; Ambure, P.; Aher, R.B. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometr. Intell. Lab.* **2016**, *152*, 18–33. [CrossRef]
69. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701. [CrossRef]
70. Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem Inf. Model.* **2010**, *50*, 1034–1041. [CrossRef] [PubMed]
71. Yap, C.W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef] [PubMed]
72. Kauffman, G.W.; Jurs, P.C. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553–1560. [CrossRef] [PubMed]
73. Cox, D. The regression analysis of binary sequences. *J. R. Stat. Soc.* **1958**, *2*, 215–242.
74. Walker, S.H.; Duncan, D.B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **1967**, *54*, 167–179. [CrossRef] [PubMed]
75. Sun, H. A Naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48*, 4031–4039. [CrossRef] [PubMed]
76. Parhizgar, H.; Dehghani, M.R.; Khazaei, A.; Dalirian, M. Application of neural networks in the prediction of surface tensions of binary mixtures. *Ind. Eng. Chem. Res.* **2012**, *51*, 2775–2781. [CrossRef]
77. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
78. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
79. Chang, C.; Lin, C. LIBSVM: A library for support vector machines. *ACM. Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27. [CrossRef]

80. Plewczynski, D.; Spieser, S.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098–1106. [CrossRef] [PubMed]
81. Pérez-Garrido, A.; Helguera, A.M.; Borges, F.; Cordeiro, M.N.D.S.; Rivero, V.; Escudero, A.G. Two new parameters based on distances in a receiver operating characteristic chart for the selection of classification models. *J. Chem. Inf. Model.* **2011**, *51*, 2746–2759. [CrossRef] [PubMed]
82. Horton, D.A.; Bourne, G.T.; Smythe, M.L. The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* **2003**, *103*, 893–930. [CrossRef] [PubMed]
83. Jensen, B.F.; Vind, C.; Brockhoff, P.B.; Refsgaard, H.H.F. In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using gaussian kernel weightedk-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511. [CrossRef] [PubMed]