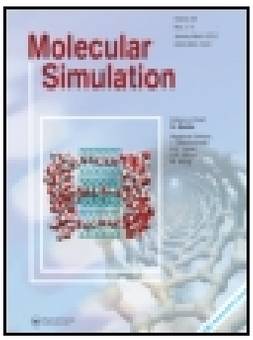

Molecular Simulation

ISSN: 0892-7022 (Print) 1029-0435 (Online) Journal homepage: https://www.tandfonline.com/loi/gmos20

Identification of DNA adduct formation of small molecules by molecular descriptors and machine learning methods

Hanbing Rao , Xianyin Zeng , Yanying Wang , Hua He , Feng Zhu , Zerong Li & Yuzong Chen

To cite this article: Hanbing Rao , Xianyin Zeng , Yanying Wang , Hua He , Feng Zhu , Zerong Li & Yuzong Chen (2012) Identification of DNA adduct formation of small molecules by molecular descriptors and machine learning methods, Molecular Simulation, 38:4, 259-273, DOI: 10.1080/08927022.2011.616891

To link to this article: https://doi.org/10.1080/08927022.2011.616891

Published online: 07 Oct 2011.

Molecular Simulation
Vol. 38, No. 4, April 2012, 259–273

Taylor & Francis
Taylor & Francis Group

Identification of DNA adduct formation of small molecules by molecular descriptors and machine learning methods

Hanbing Rao[a]*, Xianyin Zeng[a], Yanying Wang[a], Hua He[b], Feng Zhu[c], Zerong Li[d]* and Yuzong Chen[c]

[a]*College of Life and Science, Sichuan Agricultural University, Sichuan, Yaan 625014, P. R. China;* [b]*Animal Genetics and Breeding Institute of Sichuan Agricultural University, Sichuan, Yaan 625014, P. R. China;* [c]*Bioinformatics and Drug Design Group, Department of Pharmacy and Centre for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore;* [d]*College of Chemistry, Sichuan University, Chengdu 610065, P. R. China*

(*Received 12 April 2011; final version received 19 August 2011*)

In this study, we developed new computational DNA adduct prediction models by using significantly more diverse training data-set of 217 DNA adducts and 1024 non-DNA adducts, and applying five machine learning methods which include support vector machine (SVM), *k*-nearest neighbour, artificial neural networks, logistic regression and continuous kernel discrimination. The molecular descriptors used for DNA adduct prediction were selected from a pool of 548 descriptors by using a multi-step hybrid feature selection method combining Fischer-score and Monte Carlo simulated annealing method. Some of the selected descriptors are consistent with the structural and physicochemical properties reported to be important for DNA adduct formation. The y-scrambling method was used to test whether there is a chance correlation in the developed SVM model. In the meantime, fivefold cross-validation of these machine learning methods results in the prediction accuracies of 64.1–82.5% for DNA adducts and 95.1–97.6% for non-DNA adducts, and the prediction accuracies for external test set are 78.2–100% for DNA adducts and 92.6–98.4% for non-DNA adducts. Our study suggested that the tested machine learning methods are potentially useful for DNA adducts identification.

Keywords: machine learning method; DNA adducts; feature selection; Monte Carlo simulated annealing; applicability domain

1. Introduction

DNA adducts are molecules, particularly small molecules, that bind to DNA covalently, which frequently leads to carcinogenesis [1,2], and some DNA adducts have been used as anticancer agents by targeting the DNA of cancer cells as well as normal cells [3]. DNA adducts have also been used as cancer inducers and biomarkers for quantitative measure of cancer in subjects such as rats or other living animals [4]. It has been reported that cancers can be induced by DNA adducts through DNA damage, mutagenesis and impairment of DNA repair which happens naturally under normal circumstances (DNA repair) [5–9], and DNA adduct anticancer drugs produce their anticancer effects by inducing DNA damage in cancer cells followed by the DNA damage-induced apoptosis [3]. Identification of DNA adducts is important for finding carcinogens and for searching anticancer drugs. As experimental methods for DNA adducts identification are costly and time consuming [10], it is desirable to develop methods that lower the cost and time of DNA adducts identification without significant accuracy reduction.

Computational methods have been explored for identifying carcinogen DNA adducts partly based on the knowledge of the metabolic activation mechanism [11–13]. Vogel and Nivard [12] developed quantitative structure-activity relationship (QSAR) among tumorigenic potency, heritable genetic damage and structural elements of alkylating carcinogens. More recently, Coluci et al. [13] have applied principal component analysis (PCA), hierarchical clustering analysis and neural networks method to identify the carcinogenic activity of 81 polycyclic aromatic hydrocarbons at $>80\%$ accuracy level. The purpose of this study was to develop a new DNA adduct prediction model based on a more diverse training data-set using various machine learning methods.

One important step in developing machine learning DNA adduct prediction model is to compute and select appropriate molecular descriptors. A single or a standard set of descriptors according to experience may reflect adducting features to some extent, but cannot guarantee a full capture of the whole properties. In other words, there is no pre-knowledge on descriptors that are most relevant to DNA adduct prediction, so a priori feature selection is not feasible. In this study, we calculated as many molecular descriptors as possible and selected the appropriate descriptors by using feature selection algorithms (FSAs).

There are two major classes of FSAs: classifier independent and classifier dependent. A classifier-inde-

*Corresponding authors. Email: rhbsau@gmail.com or lizerong@scu.edu.cn

ISSN 0892-7022 print/ISSN 1029-0435 online

http://dx.doi.org/10.1080/08927022.2011.616891
http://www.tandfonline.com

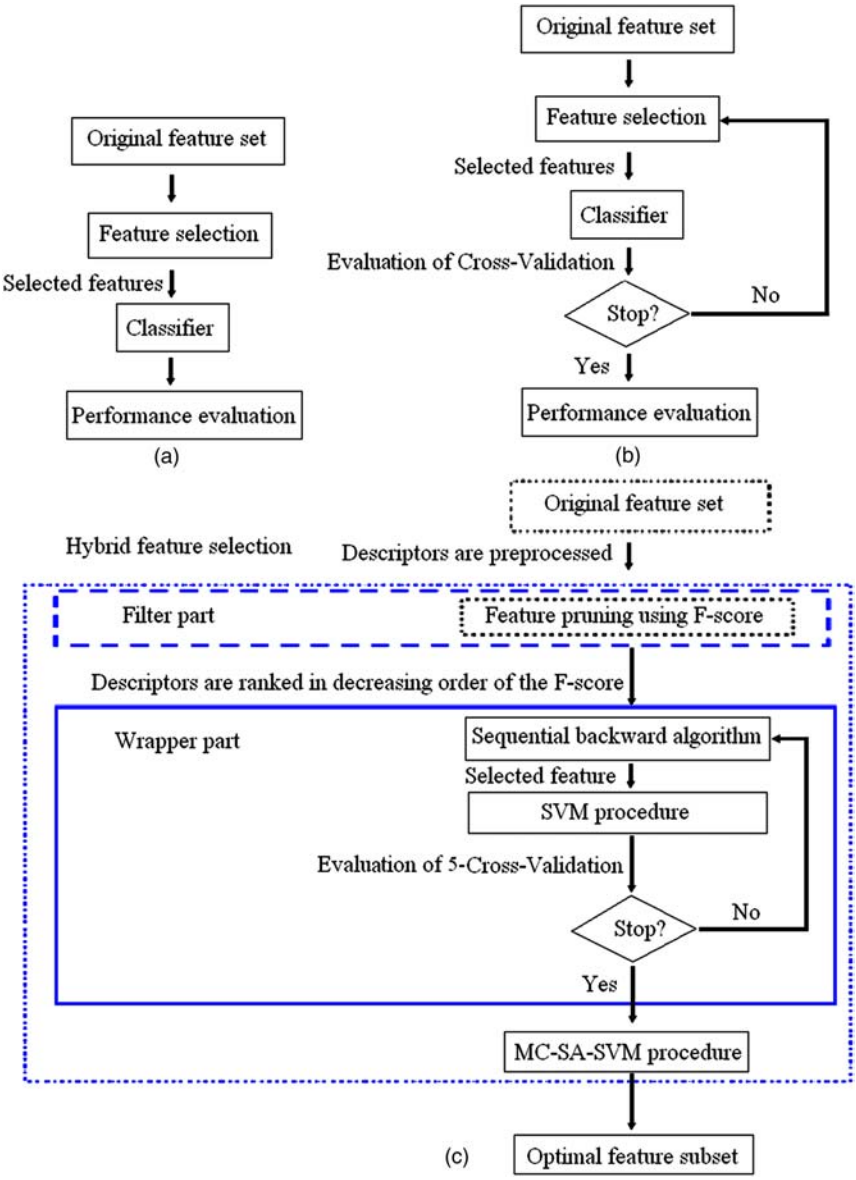

Figure 1. Comparison of feature selection method (a) filter method, (b) wrapper method and (c) hybrid method.

pendent approach is a filter method [14–16] as outlined in Figure 1(a), which is computationally efficient. The filter method attempts to identify relevant features by selecting a feature subset using a preprocessing step independent of the learning algorithm, which is less useful for redundant features and data with strongly correlated features. Classifier-dependent FSA is also called wrapper approach [17–20] as shown in Figure 1(b), which uses a specific learning algorithm, such as decision trees and support vector machines (SVMs) to evaluate the feature subset based on their contribution to the performance of the learner. Wrapper approach has the advantage of selecting features suitable to the specific learner, and hence generally results in higher learning performance than the filter method. In the wrapper approach, the selections of subset of features are imbedded in the classifier, such as recursive feature elimination [21], genetic algorithm (GA) [22] and simulated annealing (SA) method [23]. Comparing with filter method, the wrapper approach is much more computationally expensive, but is able to produce better results. A detailed introduction to the wrapper approach can be found in Ref. [24].

In this study, to overcome the computational cost of the wrapper approach and the low accuracy of the filter method, a multi-step hybrid FSA combining Fischer-score (*F*-score) and Monte Carlo simulated annealing (F-MC-SA), as shown in Figure 1(c), was used to select most relevant descriptors for DNA adduct prediction, which is different from other ranking algorithms like information-based method. *F*-score filter approach is capable of

calculating continuous features without discretising them. Moreover, MC-SA, a wrapper approach, is very efficient for searching global minimum. In the meantime, several DNA adduct prediction machine learning models were developed in combination with our hybrid feature selection method, which include SVM, *k*-nearest neighbour (*k*-NN), artificial neural networks (ANNs), logistic regression (LR) and continuous kernel discrimination (CKD). The performance of the developed models was further evaluated by different approaches: y-scrambling, five cross-validations and an external test data-set.

2. Materials and methods

2.1 Data-sets

We collected 217 DNA adducts from a number of published articles based on the criterion that there is reported experimental evidence to support the adducting property [8,25–29]. To sufficiently represent the chemical space of non-DNA adducts, we used 300 clinical trial drugs and 904 US Food and Drug Administration approved drugs from the drug information handbook [30] as the non-DNA adducts. Overall, a total of 217 DNA adducts and 1204 non-DNA adducts were used to develop and test machine learning models.

Geometry optimisation of each molecule was performed using the MM + force field of HyperChem7 (http://www.hyper.com) before computing the molecular descriptors. Molecular descriptors have been routinely used for quantitative description of the structural and physicochemical properties of molecules in the development of various QSAR models [31–34]. We used 548 1D and 2D descriptors (Supplementary Table S1) by the web-based software Model [35], which include 72 fingerprint descriptors [31], 30 constitutional descriptors [31], 92 molecular connectivity and molecular shape descriptors [31,36], 108 electro-topological state descriptors [31,37], 60 BCUT molecular descriptors [31,38] and 186 autocorrelation descriptors [31,39].

2.2 Data-set division and model validation

A recent study [40] has shown that an external test set is more appropriate for model validation than cross-validation method, and the training set should be diverse enough. Moreover, the representative points of the training set and the testing set should be close to each other. Massart and co-workers have compared four methods for selecting representative samples from training set, which include Kennard–Stone (KS) method, D-optimal design, Kohonen self-organisation mapping method and random selection method. They found that the KS method outperforms the other methods [41] because the samples chosen by the KS method can span the largest chemical space, so the prediction for most of the compounds in the test set will be interpolation and fall into the applicability domain of the chemical space covered by the training set and the model can have best prediction ability for unknown compounds. As the test set was not involved in the training process, it is independent and can be considered as an external independent test set. In this study, we divided our data-set into training and independent test sets by KS method [42–44]. In this splitting procedure, the training set has 162 DNA adducts and 768 non-DNA adducts agents, and the external test set has 55 DNA adducts and 256 non-DNA adducts agents. For comparison, a random method is also used to divide the data-set with the same procedure as the KS method.

Four methods were employed to validate the classification model. The first one is the fivefold cross-validation method. In fivefold cross-validation, the training set is divided into five subsets of approximately equal size, in which each compound in the training set appears only once. One subset of the compounds is withheld for testing, whereas the remaining subsets are used for training. This process is repeated five times for all five subsets, respectively, to provide predictions of all compounds when they are not included in the training. In this study, fivefold cross-validation is used to refine the classification models by selecting MC-SA to find the optimal subset of molecular descriptors, and optimise the model parameters, such as the sigma of the Gaussian kernel function, the number of hidden units of ANN and so on. The second model validation method is y-scrambling method, which is used to check chance correlation. The responses, i.e. the classification labels, are randomly permuted 30 times for the original data-set (including all descriptors) and 30 data-sets are given. Then the same procedure is applied to the data-sets as the above model optimisation step. The averaged prediction accuracies are given to indicate whether chance correlation has occurred. The resulting models obtained on the data-set with randomised responses should have significantly lower prediction accuracy than the models that are based on the real data because the relationship between the structure and response is broken [45,46].

The third model validation method is the application of selected descriptors to other machine learning approaches to see whether the selected descriptors are truly relevant to the discrimination of DNA adducts from non-DNA adducts. The fourth model validation method is the external test set method [47,48]. All compounds in the training set are used to train the models, while the prediction accuracies of the models are evaluated by the external test set. Because the compounds in the external test set did not involve in developing the models, it can provide a more rigorous validation of the model's predictive capability than the cross-validation method. Meanwhile, the domain of applicability is considered and discussed for compounds in the external test set.

2.3 Diversity of compounds in data-set

The diversity of the compounds used for modelling significantly affects the efficiency and robustness of the prediction models. The more diverse the compound data-set, the larger chemical space the applicability domain of model covers. Moreover, prediction accuracy of statistical learning systems is known to be strongly affected by the diversity of samples used in the training set [38,42]. The test set also needs to be diverse and representative of samples studied in order to accurately assess the capabilities of the prediction systems [45,49]. The diversity of compounds in a data-set can be estimated by using the diversity index (DI) value, which is the average values of the dissimilarity between all pairwise compounds in the data-set [50]:

$$\mathrm{DI} = \frac{\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}\mathrm{diss}(i,j)}{N(N-1)}, \quad (1)$$

where diss(i,j) is a measure of the dissimilarity between compounds i and j, and N is the number of compounds in the data-set. Dissimilarity is a complementary measure of similarity and usually defined as (1-similarity), so we have

$$\mathrm{DI} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}\mathrm{sim}(i,j)}{N(N-1)}, \quad (2)$$

where sim(i,j) is the similarity between compounds i and j. In this study, the similarity between any two compounds was computed by using commonly used similarity metric-Tanimoto coefficient [51,52]:

$$s(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}(x_{dj})^2 - \sum_{d=1}^{l} x_{di}x_{dj}}, \quad (3)$$

where l is the number of descriptors computed for the compounds in the data-set, x_{di} and x_{dj} are the values of dth descriptor for compounds i and j, respectively. It can be easily shown that $s(i,j) = 1$ when $x_{di} = x_{dj}$ ($d = 1,2, \ldots, l$) and $s(i,j) = -1/3$ when $x_{di} = -x_{dj}$ ($d = 1,2,\ldots,l$). Hence the Tanimoto coefficient ranges from -0.333 to 1. In this study, the Tanimoto coefficient is normalised to the range between 0 and 1 before substituting into Equation (2) by the following procedure:

$$\mathrm{sim}(i,j) = \frac{s(i,j) + 0.333}{1.333}, \quad (4)$$

where sim(i,j) is the normalised Tanimoto coefficient. Then the DI ranges between 0 and 1 and the diversity of a data-set increases with DI value. The value of 1 for DI denotes that each pair of compounds in the data-set has a zero-valued similarity, that is the data-set is sufficiently diverse for the given molecular descriptors, whereas the value of 0 for DI denotes that all of the compounds have the fully identical molecular descriptors. Clearly, the closer the DI value is to 1, the more diverse the data-set is. The computed value of DI is 0.7348 for the data-set of 217 DNA adducts and 1024 non-DNA adducts and 0.7344 and 0.7358 for the training and external test set, respectively. We use the NCI diversity set II [53] to find how DI can reflect the diversity of the data-set. The NCI diversity set II is a set of 1364 compounds selected from the original NCI-3D database of the almost 140,000 compounds based on their properties as unique three-point pharmacophores [54]. The computed DI value for the NCI diversity set II is 0.7614. The results are summarised in Table 1, which shows that the diversities of compounds in our data-sets are comparable to NCI diversity set II with high structural diversity.

Table 1. DI values of our data-sets and NCI diversity Set II.

Data-set	No. of compounds	DI value
The NCI diversity set II	1364	0.7614
Training set (in this study)	992	0.7344
External test set (in this study)	249	0.7358
The whole data-set (in this study)	1241	0.7348

2.4 Feature selection method

Not all of the descriptors calculated above are relevant to the discrimination of DNA adducts and non-DNA adducts. Elimination of the redundant descriptors can improve the accuracy of prediction, and facilitates the interpretation of the model by focusing on the most relevant descriptors. In this study, a multi-step hybrid feature selection method was used to find the optimal subset of features and the performance of the model, i.e. the fitness function was measured using the averaged overall prediction accuracy of fivefold cross-validation for each step of feature selection. The feature selection procedures are as follows:

Step 1. Processing. First, any descriptor that has identical values for more than 90% of the samples is removed. Second, any descriptor with the relative standard deviation (SD) less than 0.05 is removed. Finally, one of any two descriptors with the absolute value of Pearson correlation coefficient above 0.9 is removed.

Step 2. F-score ranking and backward selection. The preprocessed descriptors are ranked in decreasing order of the F-score. F-score is a simple filter technique that measures the discrimination ability of one feature [55,56]. The F-score of the ith feature can be defined as follows [55]:

$$F(i) = \frac{(\bar{\mathbf{x}}_i^{(+)} - \bar{\mathbf{x}}_i)^2 + (\bar{\mathbf{x}}_i^{(-)} - \bar{\mathbf{x}}_i)^2}{1/(n_+ - 1)\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{\mathbf{x}}_i^{(+)})^2 + 1/(n_- - 1)\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{\mathbf{x}}_i^{(-)})^2}, \quad (5)$$

where $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_i^{(+)}$ and $\bar{\mathbf{x}}_i^{(-)}$ are the averages of the ith feature of the whole, positive and negative data-sets, respectively; $x_{k,i}^{(+)}$ denotes the ith feature of kth positive sample and $x_{k,i}^{(-)}$ refers to the ith feature of kth negative sample. n_+ and n_- are the number of positive and negative samples, respectively. The numerator indicates the discrimination between the positive and negative data-sets, and the denominator indicates the

discrimination within each of the two data-sets. The higher the *F*-score, the more discriminative power this feature has [55]. In this study, features are ranked in decreasing order according to *F*-score, and the number of relevant descriptors is chosen by a sequential backward selection (SBS) algorithm. SBS starts from the initial set of all features, and each time the three lowest-ranked features are removed until the generation ability estimated by fivefold cross-validation reaches its maximum. Meanwhile, the model is optimised by a systematic search of the exponent parameter σ of the Gaussian kernel function in the SVM approach. After this step, the set of features and parameter σ are selected. *F*-score is simple and generally quite effective, but it ignores feature dependencies and the interaction with the classification. Therefore, the feature set may be further reduced by a wrapper approach.

Step 3. The wrapper approach applied in this study is Metropolis MC-SA selection, which helps to find most relevant molecular descriptors. The SA is the simulation of a physical process, and 'annealing', which involves heating the system to a high temperature and then gradually cooling it down to a preset temperature (e.g. room temperature). During this process, the possible configurations of the samples obey the Boltzmann distribution and hence the low energy states are the most populated at equilibrium. The implementation of MC-SA combined with SVM reported here is similar to that described in Ref. [57] and can be summarised as follows:

i) giving an initial exponent σ value for the Gaussian kernel function;
ii) setting the initial simulation temperature T;
iii) generating a trial solution to the underlying optimisation problem, i.e. an SVM model is built based on a random selection of descriptors;
iv) calculating the value of the fitness function, which characterises the quality of the trial solution to the underlying problem, i.e. the performance of the trial subset;
v) perturbing the trial solution to obtain a new solution and build a new MC-SA-SVM model for the new trial solution;
vi) calculating the value of the fitness function Q_{new} for the new trial solution;
vii) applying the optimisation criteria: if $Q_{\text{old}} < Q_{\text{new}}$, the new solution is accepted and used to replace the old trial solution; if $Q_{\text{old}} > Q_{\text{new}}$, the new solution is accepted only if the Metropolis criterion is satisfied, i.e.

$$\text{rnd} < \text{e}^{-(Q_{\text{old}} - Q_{\text{new}})/T}, \tag{6}$$

where rnd is a random number uniformly distributed between 0 and 1;
viii) lowering the simulation temperature T to a predetermined value and return to step (iii);
ix) systematically adjusting the σ value and going back to step (ii).

After these steps, an optimal subset of molecular descriptors and σ value will be obtained and the final MC-SA-SVM model will give the least generalisation error.

2.5 Machine learning methods

(1) SVM methods. The SVM method is a supervised machine learning technique for learning classification and regression rules from data. An introduction to SVM can be found in Refs [58–60]. There are two types of SVM algorithms, linear and nonlinear SVM. Linear SVM algorithm finds a hyper-plane separating two classes with a maximum margin. This hyper-plane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimises $\|\mathbf{w}\|^2$, which satisfies the following conditions: $\mathbf{w}\cdot\mathbf{x}_i + b \geq +1$, for $y_i = +1$ (DNA adducts) and $\mathbf{w}\cdot\mathbf{x}_i + b \leq -1$, for $y_i = -1$ (non-DNA adducts). Here $\mathbf{x}_i$ is the input feature vector, y_i is the class index and $\mathbf{w}$ is a vector normal to the hyper-plane. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified by using a positive or negative $f(\mathbf{x})$ value, indicating that the vector $\mathbf{x}$ belongs to the active or inactive class [61]. Nonlinear SVM maps feature vectors into a high-dimensional feature space implicitly by a kernel function such as Gaussian kernel $K(x_i, x_j) = \exp(-\|x_j - x_i\|^2/(2\sigma^2))$ [62,63] (also the kernel function chosen in this study) and the linear SVM procedure is then applied to the feature vectors in this feature space. The exponent σ of the Gaussian kernel is the model parameter of SVM approach.

(2) KNN method. The *k*-NN algorithm [64–66] is the most basic instance-based method. As one would expect from the name, this algorithm classifies $\mathbf{x}$ by examining the classes on the *k*-NNs and assigning it the class most frequently represented among the *k*-NNs. It assumes that all instances correspond to the points in the *n*-dimensional feature space. The nearest neighbours of an instance are defined in terms of the Euclidean distance [66–68]. In this study, the *k*-NN prediction accuracies are estimated through fivefold cross-validation with the same data-set and molecular descriptors selected in the SVM classification model.

(3) ANN method. The ANN is a mathematical tool that can be used for regression and classification, which was originally inspired by the neuron structure in the brain. It consists of a series of nodes (the analogy of neurons) that have multiple connections with other nodes. In this study, a back-propagation network employing a single layer of hidden units is used to find a classifier separating DNA adducts from non-DNA adducts. The gradient descent with momentum is used for the training. A three-layer architecture ANN is used in this study and the optimal number of neurons in the hidden layer is chosen by maximising the generalisation ability, which is the averaged

overall accuracies defined in Equation (20) estimated by a fivefold cross-validation [69].

(4) LR method. The LR method [70,71] is a supervised learning approach that attempts to distinguish K classes from each other using a weighted sum of some predictor variables X_i. For two-class classification problems, the probability u that an event belongs to the positive class is related to a set of explanatory variables in the form:

$$\log it(u) = \ln\left(\frac{u}{1-u}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta nXn = \beta^T X \tag{7}$$

or

$$u = \frac{1}{1+\exp(-\beta^T X)}, \tag{8}$$

where u is the probability that vector $\mathbf{x}$ belongs to the positive class, β_0 is the regression model constant and $\beta_1 - \beta_k$ are the coefficients corresponding to the descriptors $X_1 - X_k$. Then the probability that the event belongs to the negative class is $(1 - u)$. The parameters in the LR model are estimated by the maximum likelihood method. The natural logarithm of the likelihood function for a training set $\{(\mathbf{X}_i, y_i), i = 1,m)\}$ is

$$L(\beta) = \sum_{i=1}^{m} y_i \ln u + (1 - y_i)\ln(1 - u) - \frac{c}{2}\beta^T\beta, \tag{9}$$

where c is the regularised coefficient and is determined by maximising the generalisation ability for the validation set. The optimisation of the parameters $\beta_0, \beta_1, \ldots, \beta_n$, in this study is carried out through Newton–Raphson method [72]. In this study, the LR prediction accuracies are estimated through fivefold cross-validation with the same data-set and molecular descriptors selected in the SVM classification model.

(5) CKD method. The CKD method recently employed by Willett et al. [73–77] uses kernel density estimation to derive the conditional class probability density functions in Bayes classifier for conducting classification using continuous features. According to Bayes' theorem, the probabilities of a sample $\mathbf{x}$ in positive class and negative class are

$$P(y = +1|\mathbf{x}) = \frac{P(\mathbf{x}|y = +1)\cdot P(y = +1)}{P(\mathbf{x})}, \tag{10}$$

$$P(y = -1|\mathbf{x}) = \frac{P(\mathbf{x}|y = -1)\cdot P(y = -1)}{P(\mathbf{x})}, \tag{11}$$

where $p(\mathbf{x})$ is a constant for the positive class and negative class and can be obtained by the normalisation condition:

$$P(y = +1|\mathbf{x}) + P(y = -1|\mathbf{x}) = 1. \tag{12}$$

Hence,

$$P(\mathbf{x}) = P(\mathbf{x}|y = +1)\cdot P(y = +1) + P(\mathbf{x}|y = -1)\cdot P(y = -1). \tag{13}$$

$P(y = +1)$ and $P(y = -1)$ can be estimated from the proportions of the positive or the negative class in the training set:

$$P(y = +1) = \frac{l_+}{l}, \quad P(y = -1) = \frac{l_-}{l}, \tag{14}$$

where l_+ and l_- are the number of positive and negative samples, respectively. The conditional class probability density functions $P(\mathbf{x}|y = +1)$ and $P(\mathbf{x}|y = -1)$ are estimated from the training set by

$$f_+(\mathbf{x}) = \frac{1}{l_+}\sum_{i=1}^{l_+} K_h(\mathbf{x}, \mathbf{x}_i), \tag{15}$$

$$f_-(\mathbf{x}) = \frac{1}{l_-}\sum_{i=1}^{l_-} K_h(\mathbf{x}, \mathbf{x}_i), \tag{16}$$

where $K_{h_k}(\mathbf{x}, \mathbf{x}_i)$ is the kernel density function and h is the bandwidth of the Gaussian, which is optimised by an analysis of the training-set data. For binary variables, the kernel density function is usually in the form of Aitchison–Aitkin function [78]

$$K_h(\mathbf{x}, \mathbf{x}_i) = h^{m-d(\mathbf{x},\mathbf{x}_i)}(1-h)^{d(\mathbf{x},\mathbf{x}_i)}, \tag{17}$$

where $0.5 \le h \le 1$, m is dimension of the vector $\mathbf{x}$ and $d(\mathbf{x}, \mathbf{x}_i)$ is the squared Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} - \mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i). \tag{18}$$

Willett called this binary kernel discrimination. For continuous variables $\mathbf{x}$, the kernel density function is usually the Gaussian kernel:

$$K_h(\mathbf{x}, \mathbf{x}_i) = \frac{1}{h\sqrt{2\pi}}\mathrm{e}^{-((d(\mathbf{x}_i,\mathbf{x}))/2h^2)}, \tag{19}$$

where $0 < h < \infty$ and $d(\mathbf{x}, \mathbf{x}_i)$ is also the squared Euclidean distance. Willett called this CKD.

3. Performance measure

The effectiveness of machine learning methods can be measured by using true positive (TP), true negative (TN), false positive (FP), false negative (FN), sensitivity [SE = TP/(TP + FN)] and specificity [SP = TN/(TN + FP)] [79], where TP is the number of DNA adducts predicted correctly, TN is the number of non-DNA adducts predicted correctly, FN is the number of DNA adducts predicted as non-DNA adducts, FP is the number of non-DNA adducts predicted as DNA adducts. The overall prediction accuracy (Q) and Matthews correlation coefficient (MCC) [80] are used to measure the overall

prediction performance:

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{20}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}}. \tag{21}$$

4. Results and discussion

4.1 Influence of training set design on performance of machine learning methods

KS and random methods were, respectively, used to design the training set. The results are listed in Table 2. It is shown that the performance of the KS method generally outperforms the random method and this is because that the samples chosen by the KS method, which maximises the minimal Euclidean distances between already selected compounds and the remaining compounds, can span the largest chemical space, so most of the compounds in the test set will be in interpolation and fall into the applicability domain of the chemical space covered by the training set. However, the random method divides compounds randomly into training set and test set and this strategy is usually only effective for very densely populated samples, because there is a risk that compounds of some classes in the test set are not selected in the training set.

4.2 Feature selection and model development

The fivefold cross-validation performance of SVM models integrating three feature selection steps is given in Table 3. From Table 3, it can be seen that after the preprocessing step, the number of molecular descriptors is reduced from 548 to 171, indicating that the majority of descriptors have low information content or are highly correlated with other descriptors. After this preprocessing, the prediction

Table 2. Influence of training set design on performance of machine learning methods[a].

	SVM ($\sigma = 4$)			ANN ($n = 35$)			LR ($C = 1$)			K-NN ($k = 1$)			CKD ($h = 0.5$)		
Methods	SE (%)	SP (%)	Q (%)	SE (%)	SP (%)	Q (%)	SE (%)	SP (%)	Q (%)	SE (%)	SP (%)	Q (%)	SE (%)	SP (%)	Q (%)
KS	78.4	95.1	91.8	72.5	90.8	87.6	60.5	90.7	85.4	78.5	90.5	88.3	75.5	92.3	89.3
Random	70.5	85.3	82.7	65.5	85.8	82.2	50.5	80.7	75.4	70.5	80.4	78.7	68.5	94.3	89.7

[a] Averaged prediction accuracies over the five sets by fivefold cross-validation.

Table 3. Effect of feature selection on the performance of the SVM model.

			Prediction for DNA adducts			Prediction for non-DNA adducts				
Step (number of descriptors)[a]	Optimal σ[b]	Cross-validation set	TP	FN	SE(%)	TN	FP	SP(%)	Q(%)	C
Step 1 (171)	15	1	34	10	77.3	194	11	95.1	91.7	0.757
		2	31	13	70.5	197	8	96.1	90.8	0.698
		3	34	9	79.1	195	10	95.1	92.3	0.735
		4	36	7	83.7	194	11	95.1	92.7	0.757
		5	35	8	81.3	192	12	94.1	91.5	0.729
		Average			78.4			95.1	91.8	0.735
		SD[c]			5			0.72	0.73	0.024
Step 2 (143)	13	1	34	10	77.3	196	9	95.8	92.4	0.735
		2	32	12	72.7	200	5	97.6	93.2	0.734
		3	35	8	81.4	198	7	96.6	94	0.787
		4	37	6	84.1	197	8	96.1	94.4	0.807
		5	36	7	83.7	194	11	94.6	92.7	0.757
		Average			79.8			96.1	93.3	0.764
		SD[c]			4.8			7.4	0.92	0.032
Step 3 (61)	4	1	34	10	77.3	201	4	98	94.4	0.799
		2	34	10	77.3	202	3	98.5	94.8	0.813
		3	36	7	83.7	200	5	97.6	95.2	0.828
		4	38	5	88.4	202	3	98.5	96.8	0.886
		5	37	6	86	195	9	95.6	93.9	0.795
		Average			82.5			97.6	95	0.824
		SD[c]			5.1			4.5	1.2	0.053

[a] Step 1, preprocessing; Step 2, filter step through F-score ranking and backward selection; Step 3, MC-SA [b] σ, exponent of the Gaussian kernel of SVM
[c] SD, standard deviation.

accuracies for DNA adducts, non-DNA adducts and the all samples are 78.4%, 95.1% and 91.8%, respectively, and the MCC *C* is 0.735. In the second step, i.e. *F*-score ranking and backward selection, the number of molecular descriptors is reduced from 171 to 143 and the prediction accuracies for DNA adducts, non-DNA adducts and all the samples are improved to 79.8%, 96.1% and 93.3%, respectively, and the MCC *C* reaches 0.764. In the third step of feature selection, i.e. MC-SA, the number of molecular descriptors is further reduced to 61. The prediction accuracies for the DNA adducts, non-DNA adducts and the all samples are 82.5%, 97.6% and 95%, respectively, and the MCC *C* is 0.824. These results show that the multi-step hybrid feature selection method is capable of removing the non-relevant descriptors thereby reducing the noises and improving the performance of the SVM. Moreover, those 61 most relevant features identified were subsequently used to build five machine learning models.

The 61 selected descriptors are listed in Table 4. Most of the selected descriptors are the descriptors encoding the 1D or 2D structural information weighted by atomic physicochemical properties. These descriptors can be categorised into several classes: simple molecular properties, BCUT descriptors, electro-topological state, autocorrelation descriptors, molecular connectivity and shape descriptors. Simple descriptors are molecular weight, counts of special atoms, chemical bonds and sub-structures in the molecules. BCUT descriptors encode atomic properties relevant to intermolecular interaction [81–83].

Table 4. Selected molecular descriptors.

Descriptor class	Descriptions	*N*[a]
Simple molecular properties	Number of rotatable bonds, number of six-member non-aromatic rings, number of N atoms, number of seven-member rings, number of six-member aromatic rings, number of Cl atoms, average molecular weight	6
BCUT descriptors	The third highest eigenvalue of BCUT descriptors weighted by atomic electronegativity; the fifth highest eigenvalue of BCUT descriptors weighted by atomic E-state; the second lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity; the third lowest eigenvalue of BCUT descriptors weighted by atomic E-state; the fourth lowest eigenvalue of BCUT descriptors weighted by atomic polarisability; the third lowest eigenvalue of BCUT descriptors weighted by atomic polarisability; the first lowest eigenvalue of BCUT descriptors weighted by atomic polarisability; the fifth highest eigenvalue of BCUT descriptors weighted by atomic electronegativity; the fourth highest eigenvalue of BCUT descriptors weighted by atomic mass; the first highest eigenvalue of BCUT descriptors weighted by atomic polarisability; the fourth highest eigenvalue of BCUT descriptors weighted by atomic VDW volume; the second lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity; the second lowest eigenvalue of BCUT descriptors weighted by atomic VDW radius; the second lowest eigenvalue of BCUT descriptors weighted by atomic mass; the fourth lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity; the fourth lowest eigenvalue of BCUT descriptors weighted by atomic VDW radius; the fourth highest eigenvalue of BCUT descriptors weighted by atomic mass; the third lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity	18
Electro-topological state	Sum of estate of atom type aaaC; sum of estate of atom type sCH3; sum of estate of atom type ddsN; sum of estate of atom type dS; sum of estate of atom type ssO; sum of H estate of atom type HaaCH; sum of estate of all C atoms; sum of estate of atom type aaN; sum of estate of all C atoms; sum of estate of atom type dsN	11
Autocorrelation descriptors (2D)	Moreau-Broto autocorrelation of lag 5 weighted by atomic E-state indices, Moreau-Broto autocorrelation of lag 7 weighted by atomic E-state indices, Moran autocorrelation of lag 2 weighted by atomic electronegativity, Geary autocorrelation of lag 1 weighted by atomic mass, Geary autocorrelation of lag 3 weighted by atomic VDW radius, Geary autocorrelation of lag 4 weighted by atomic E-state indices, Geary autocorrelation of lag 2 weighted by atomic mass, Geary autocorrelation of lag 1 weighted by atomic electronegativity, Geary autocorrelation of lag 3 weighted by atomic mass, Geary autocorrelation of lag 2 weighted by atomic polarisability,Geary autocorrelation of lag 6 weighted by atomic electronegativity, Geary autocorrelation of lag 3 weighted by atomic E-state indices	12
Molecular connectivity and shape	The second solvation connectivity index; mean eccentricity deviation; dispersion; 0th order delta chi index; arithmetic topological index by Narumi	5
Fingerprint descriptors	Fingerprint for containing rings connected by seven non-ring edges, fingerprint for secondary ammonium, fingerprint for containing rings connected by four non-ring edges, fingerprint for diol (C(OH)—C(OH)—), fingerprint for containing rings connected by three non-ring edges, fingerprint for fused rings with two rings, fingerprint for organohalide, fingerprint for six-member aromatic rings, fingerprint for O heterocyclic rings, fingerprint for ketone (R—CO—R)	10

[a] The number of molecular descriptors.

The electro-topological state indices are numerical values computed for each atom in a molecule, which encode information about both the topological environments of that atom in the molecule [37,84]. Topological autocorrelation descriptors, including Moreau-Broto autocorrelation, Moran coefficient and Geary coefficient, are molecular descriptors encoding both molecular structure and physicochemical properties attributed to atoms as a vector [31]. Molecular connectivity and shape descriptors encode information about molecular size, shape, branching, un-saturation, heteroatom content and cyclicity [85,86]. The molecular fingerprints can be considered as another class of molecular descriptors. They have been successfully used in molecular similarity search [87], indicating that they can give more accurate description of the molecular structures.

As shown in Table 4, most of the selected descriptors are the distributed-based descriptors taking the atomic properties as the weighting. So, these descriptors are grouped according to atomic properties and their ratios in the descriptor set, which are given in Figure 2. The descriptors weighted by atomic properties account for 74.1% of all selected descriptors. The results indicate that these descriptors reflect the distribution of atomic properties along the E-state indices, electronegativity and polarisability play an important role for discriminating the diverse data-set, which is consistent with a linear discriminant analysis of structure-based descriptors for multidrug resistant (MDR) agents that showed that 60% of the molecular descriptors important for MDR are topological in nature [88]. Among the 61 selected descriptors, the 18 most important descriptors whose *F*-score values are greater than 0.2 are listed in Table 5. The higher the *F*-score, the more discriminative power this feature has. So these descriptors are very important for discriminating between the DNA-adducts and non-DNA adducts.

Some of our selected molecular descriptors for discriminating DNA adducts from non-DNA adducts are consistent with the structural and physicochemical properties, reported to be important for DNA adduct formation. For instance, the food mutagen 2-amino-3-methylimidazo[4,5-f]quinoline approaches DNA and forms DNA adduct by adopting a specific conformation via large variation of specific rotatable bond [89]. This feature is covered by our selected descriptor, number of rotatable bonds. The reaction of some DNA adducts such as cisplatin analogues [90] and *N*-hydroxyarylamins with DNA occurs at specific nitrogen atoms, which is reflected by our selected descriptor: number of N atoms. The presence of a benzo ring in a specific DNA adduct has been found to be important for stabilised DNA adduct formation and subsequent carcinogenesis [91]. Our selected descriptors, 'number of six-member aromatic rings', and 'number of seven-member aromatic rings' and fingerprint descriptors for rings, partially account for the presence of such rings.

4.3 Model validation through Y-scrambling

Y-scrambling was applied to exclude the possibility of chance correlation, i.e. fortuitous correlation without any predictive ability. The classification labels (TN) of the 480 compounds in the training set were reordered in a random manner. Afterwards, attempts were made to build SVM model with the scrambled activity data. A total of 30 randomisation runs were performed. The results of the y-scrambling test are given in Table S2 (Supporting Information). The average accuracies for the DNA adducts, non-DNA adducts and overall samples are 20.3–44.5%, 55.1–63.4% and 51.4–62.6%, respectively. In all cases, the obtained random models have much lower prediction accuracies than the model based on the real data, indicating no obvious chance correlation in the SVM model.

4.4 Development and test of other machine learning models

To test whether they are truly relevant to the discrimination between DNA adducts and non-DNA adducts, the 61 selected descriptors were used to develop ANN, *k*-NN, LR and CKD DNA adduct prediction models. The prediction

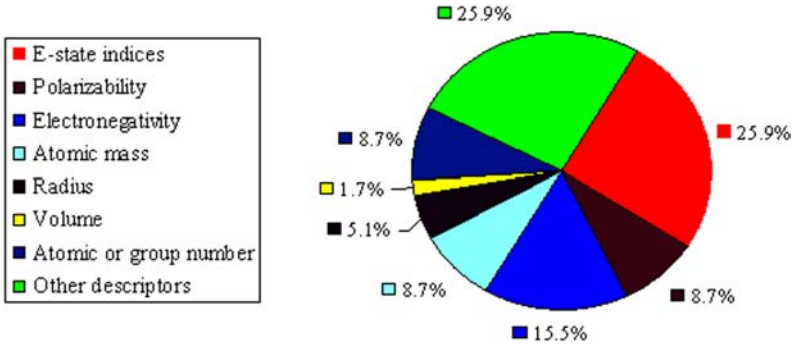

Figure 2. (Colour online) Classification of the selected descriptors.

Table 5. The Fisher-scores of the 18 most important descriptors.

No.	Molecular descriptors	Description	*F*-score value	Average value[a] DNA adduct	Average value[a] non-DNA adduct
1	$^2x^5$	The second solvation connectivity index	0.522	−0.779	0.165
2	M_{ed}	Mean eccentricity deviation	0.518	−0.807	0.171
3	λ_{H3}^{SE}	The third highest eigenvalue of BCUT descriptors weighted by atomic electronegativity	0.469	−0.897	0.189
4	I_5(ES)	Moreau-Broto autocorrelation of lag 5 weighted by atomic E-state indices	0.463	−0.716	0.152
5	λ_{H5}^{ES}	The fifth highest eigenvalue of BCUT descriptors weighted by atomic E-state indices	0.426	−0.714	0.152
6	I_7(ES)	Moreau-Broto autocorrelation of lag 7 weighted by atomic E-state indices	0.391	−0.667	0.141
7	ΣS(aaaC)	Sum of E-state of atom type aaaC	0.380	1.151	−0.244
8	λ_{L2}^{SE}	The second lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity	0.355	0.796	−0.169
9	D	Dispersion	0.345	−0.629	0.133
10	$\Delta^0\chi$	Zeroth delta chi index	0.322	−0.634	0.134
11	λ_{H3}^{ES}	The third lowest eigenvalue of BCUT descriptors weighted by atomic E-state indices	0.320	0.768	−0.163
12	λ_{L4}^{SE}	The fourth lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity	0.314	0.729	−0.154
13	λ_{L3}^{SE}	The third lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity	0.301	0.727	−0.154
14	λ_{L1}^{SE}	The fifth lowest eigenvalue of BCUT descriptors weighted by atomic electronegativity	0.246	0.704	−0.149
15	λ_{H5}^{SE}	The fifth highest eigenvalue of BCUT descriptors weighted by atomic electronegativity	0.237	−0.742	0.157
16	Nr	The number of six-member non-aromatic rings	0.234	−0.531	0.113
17	f	Fingerprint for containing rings connected by seven non-ring edges	0.231	−0.516	0.109
18	sCH3	Sum of estate of atom type sCH3	0.226	−0.532	0.113

[a] The values of descriptors have been standardised.

accuracies of these methods and the SVM method are given in Table 6. The prediction accuracies for DNA adducts, non-DNA adducts, total agents and MCC *C* are between 64.1 and 82.5%, 95.1 and 97.6%, 90.2 and 95%, and 0.636 and 0.824, respectively. SVM, *k*-NN and CKD were found to outperform LR and ANN. Our study suggests that the descriptors selected by our multi-step hybrid feature selection method in developing SVM DNA adduct prediction model are equally useful for developing other machine learning models for predicting DNA adducts. Therefore, these selected descriptors are likely relevant to the classification of DNA adducts from non-DNA adducts. Moreover, all the developed machine learning models show no apparent over-fitting phenomenon, which frequently occur in the application of wrapper methods (http://www.scss.tcd.ie/publications/tech-reports/reports.05/TCD-CS-2005-17.pdf).

4.5 Performance evaluation by the external test set

Before evaluating our developed SVM and other machine learning models by using the external test set, the applicability domain of these models needs to be determined. The applicability domain of a model is the chemical space covered by the training set, from which the model has been developed and is applicable to make predictions for new compounds. Ideally, the model should only be used to make predictions within that domain by interpolation, not extrapolation [92]. In mathematical terms, it means the estimation of interpolation regions in the multivariate space of training set, because, in general, interpolation is more reliable than extrapolation. The consideration of applicability domain is essential for proper applications and confidence assessment of the performance of machine learning methods. The applicability domain problem has been addressed by several groups [93,94]. There are four major approaches (range based, distance based, geometrical and probability density distribution based) to estimate interpolation regions in a multivariate space (http://ecb.jrc.ec.europa.eu/documents/QSAR/INFORMATIONSOURCES/applicability domain overview.pdf). In the study by Tropsha et al., the applicability domain was defined as the distance cutoff value $D_T = <D> + Zs$, where Z is a threshold parameter and $<D>$ and S are the average and SD of all Euclidean distance between each compound and its nearest neighbour for all compounds in the training set [93]. In this study,

Table 6. Performance of SVM and other machine learning methods using the selected 61 descriptors.

			Prediction for DNA adducts			Prediction for non-DNA adducts				
Methods	Parameter	Fivefold cross-validation	TP	FN	SE(%)	TN	FP	SP (%)	Q(%)	C
k-NN	$k = 1$	1	35	9	79.5	197	7	96.5	93.2	0.775
		2	33	11	77.2	197	7	96.5	92.4	0.743
		3	37	6	86.1	189	16	92.2	91.1	0.723
		4	34	9	79.1	196	9	95.6	92.7	0.747
		5	36	8	83.7	193	11	94.6	92	0.745
		Average			81.1			95.1	92.3	0.747
ANN	$n = 35$	1	35	9	79.5	199	6	97.1	94	0.788
		2	34	10	77.2	196	9	95.6	92.4	0.743
		3	34	9	79.1	196	9	95.6	92.7	0.747
		4	34	9	79.1	198	7	96.6	93.6	0.771
		5	31	12	72.1	192	12	94.1	90.3	0.662
		Average			77.4			95.8	92.6	0.742
LR	$C = 1$	1	27	17	61.4	196	9	95.6	89.6	0.618
		2	29	15	65.9	197	8	96.1	90.8	0.665
		3	29	15	67.4	194	11	94.6	89.9	0.629
		4	28	15	65.1	195	10	95.2	89.9	0.633
		5	26	17	60.5	198	7	97.1	90.7	0.636
		Average			64.1			95.7	90.2	0.636
CKD	$h = 0.4$	1	35	9	79.5	198	7	96.6	93.6	0.775
		2	34	10	77.3	198	7	96.6	93.2	0.760
		3	36	7	86	189	16	92.2	91.1	0.706
		4	34	9	79.1	196	9	95.6	92.7	0.747
		5	36	8	83.7	193	11	94.6	92.7	0.745
		Average			81.1			95.1	92.7	0.747
SVM	$\sigma = 4$	1	34	10	77.3	201	4	98	94.4	0.799
		2	34	10	77.3	202	3	98.5	94.8	0.813
		3	36	7	83.7	200	5	97.6	95.2	0.828
		4	38	5	88.4	202	3	98.5	96.8	0.886
		5	37	6	86	195	9	95.6	93.9	0.795
		Average			82.5			97.6	95	0.824

we define the applicability domain in a similar way using a similarity cutoff instead of distance cutoff

$$S_T = < S > - Z\sigma, \tag{22}$$

where $< S >$ and σ are the average and SD of the normalised Tanimoto similarity coefficient of all compounds in the training set to its nearest neighbour. Z is an arbitrary parameter to control the significance level. We set the default value of this parameter Z at 0.5, which formally places the allowed Tanimoto coefficient at one-half of the SD. Thus, if the normalised Tanimoto coefficient of an external compound to its nearest neighbour in the training set is less than this threshold, the compound is considered as beyond the application domain and the prediction is unreliable.

In this study, the applicability domain was estimated in the multivariate space defined by the 61 selected descriptors for the training set. Before the calculation of the applicability domain, each descriptor was auto-scaled to zero mean and unit variance, and PCA was applied to reduce the dimensionality and remove redundant information. In the end, the first 27 principal components with $>90\%$ of the information content were selected for the calculation of the similarity coefficient. For the training set, the average maximum and SD of the normalised Tanimoto coefficient are 0.857 and 0.131, respectively, and the similarity threshold S_T is 0.791. Only 3 of the 311 compounds in the external test set are not in the applicability domain and their normalised Tanimoto similarity coefficients to their nearest neighbours are 0.569, 0.557 and 0.474. The structures of these three 'non-applicable' compounds are shown in Figure 3.

Table 7 gives the DNA adduct prediction performance estimated by the external test set. The prediction accuracy for DNA adducts, non-DNA adducts, overall accuracy Q and MCC C are 78.2–100%, 92.6–98.4%, 91.9–96.5% and 0.726–0.893, respectively, for the five tested machine learning models SVM, CKD, k-NN, ANN and LR. In particular, CKD and SVM show slightly better overall performance than k-NN and ANN, which in turn are substantially better than LR. Overall, our testing results suggest that all the tested machine learning methods are potentially useful for predicting DBA-adducts at good prediction accuracy levels.

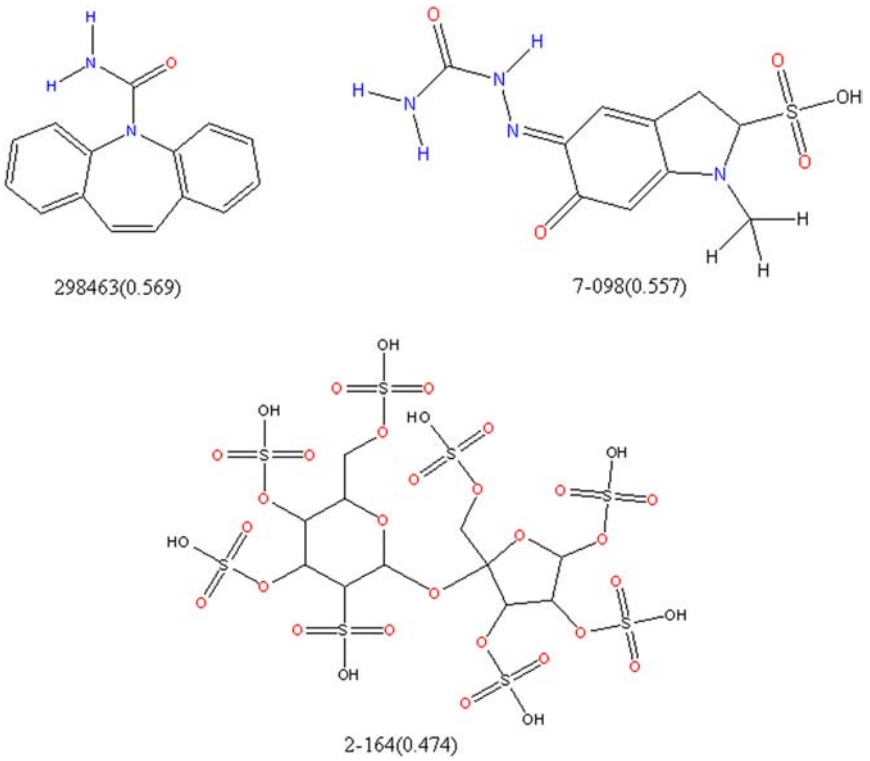

Figure 3. Structures of the compounds that are out of the application domain.

Table 7. Comparison of prediction accuracies of different machine learning approaches by independent test set with the selected molecular descriptors.

	External validation method							
Approaches	TP	FN	SE (%)	TN	FP	SP (%)	*Q* (%)	*C*
k-NN	55	0	100	237	19	92.6	93.9	0.830
ANN	54	1	98.2	237	19	92.6	93.4	0.817
LR	43	12	78.2	243	13	94.9	91.9	0.726
CKD	55	0	100	252	4	98.4	98.7	0.886
SVM	55	0	100	245	11	95.7	96.5	0.893

5. Conclusion

Our study demonstrated the usefulness of machine learning methods for predicting DNA adducts. The prediction performance of machine learning methods can be significantly improved by using molecular descriptors most relevant to classification of DNA adducts. We used a multi-step hybrid feature selection method to select the relevant descriptors from a large pool of molecular descriptors. The performance of SVM was found to be significantly improved by our selected descriptors, and the machine learning methods also showed good performance based on these selected descriptors. Moreover, some of the selected descriptors are consistent with the structural and physicochemical properties reported to be important for DNA adduct formation. These suggest that our multi-step hybrid feature selection method is efficient in selecting molecular descriptors relevant for classification of DNA adducts. Our study also reveals that the cross-validation method may be used to optimise the model parameters and select the relevant descriptors to overcome the over-fitting problem, and the hold-out method by designing a representative training set may be used to build the final reliable classification model, which can be virtually used in screening from large compound libraries and the prediction of DNA adducts compounds of the untested compounds.

Supporting Information

Supporting information including the all calculated descriptors are provided in Table S1 and the validation results of the proposed model by y-scrambling test are listed in Table S2 and the molecular structures of the DNA adducts and non-DNA adducts are given in sdf format.

Acknowledgements

This study was supported by a grant from the Two-Way Support Programs of Sichuan Agricultural University (Project No. 00770117) and the National Natural Science Foundation of China (Project No. 20973118).

References

[1] I. Al-Saleh, J. Arif, I. Ei-Doush, N. Al-Sanea, A. Abdul Jabbar, G. Billedo, N. Shinwari, A. Mashhour, and G. Mohamed, *Carcinogen DNA adducts and the risk of colon cancer: Case–control study*, Biomarkers 13 (2008), pp. 201–216.

[2] D.H. Phillips, *DNA adducts as markers of exposure and risk*, Mutation Res. 577 (2005), pp. 284–292.

[3] D. Wang and S.J. Lippard, *Cellular processing of platinum anticancer drugs*, Nat. Rev. Drug Discov. 4 (2005), pp. 307–320.

[4] K. Peltonen and A. Dipple, *Polycyclic aromatic hydrocarbons: Chemistry of DNA adducts formation*, J. Environ. Med. 37 (1995), pp. 52–58.

[5] E. Kriek, M. Rojas, K. Alexandrov, and H. Bartsch, *Polycyclic aromatic hydrocarbon-DNA adducts in humans: Relevance as biomarkers for exposure and cancer risk*, Mutation Res. 400 (1998), pp. 215–231.

[6] B.T. Golding and W.P. Watson, *Possible mechanisms of carcinogenesis after exposure to benzene*, IARC Sci. Publ. 150 (1999), pp. 75–88.

[7] M. Wu, S. Yan, D.J. Patel, N.E. Geacintov, and S. Broyde, *Relating repair susceptibility of carcinogen-damaged DNA with structural distortion and thermodynamic stability*, Nucleic Acids Res. 30 (2002), pp. 3422–3432.

[8] S.A. Kulkarni, D. Moir, and J. Zhu, *Influence of structural and functional modifications of selected genotoxic carcinogens on the metabolism and mutagenicity*, SAR QSAR Environ. Res. 18 (2007), pp. 459–514.

[9] L.J. Marnett, *Lipid peroxidation-DNA damage by malondialdehyde*, Mutation Res. 424 (1999), pp. 83–95.

[10] E.J. Carrier, V. Amarnath, J.A. Oates, and O. Boutaud, *Characterization of covalent adducts of nucleosides and DNA formed by reaction with levuglandin*, Biochemistry 48 (2009), pp. 10775–10781.

[11] R. Benigni, L. Conti, R. Crebelli, A. Rodomonte, and M.R. Vari, *Simple and α,β-unsaturated aldehydes: Correct prediction of genotoxic activity through structure–activity relationship models*, Environ. Mol. Mutagen. 19 (2005), pp. 338–345.

[12] E.W. Vogel and M.J.M. Nivard, *The subtlety of alkylating agents in reactions with biological macromolecules*, Mutation Res. 395 (1994), pp. 13–32.

[13] V.R. Coluci, R. Vendrame, R.S. Braga, and D.S. Galvao, *Identifying relevant molecular descriptors related to carcinogenic activity of polycyclic aromatic hydrocarbons (PAHs) using pattern recognition methods*, J. Chem. Inf. Comput. Sci. 42 (2002), pp. 1479–1489.

[14] M. Dash, K. Choi, P. Scheuermann, H. Liu, *Feature selection for clustering – a filter solution*, Proceedings of the Second International Conference on Data Mining, 2000, pp. 115–122.

[15] M.A. Hall, *Correlation-based feature selection for discrete and numeric class machine learning*, Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 359–366.

[16] H. Liu and R. Setiono, *A probabilistic approach to feature selection – a filter solution*, Proceedings of the 13th International Conference on Machine Learning (ICML'96), Bari, Italy, 1996, pp. 319–327.

[17] R. Caruana, D. Freitag, *Greedy attribute selection*, Proceedings of the 11th International Conference on Machine Learning, 1994, pp. 28–36.

[18] J.G. Dy and C.E. Brodley, *Feature subset selection and order identification for unsupervised learning*, Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000, pp. 247–254.

[19] Y. Kim, W. Street, and F. Menczer, *Feature selection in unsupervised learning via evolutionary search*, AAAI: Proceedings of the Sixth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2000, pp. 365–369.

[20] Y. Leung and Y. Hung, *A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification*, IEEE/ACM Trans. Comput. Biol. Bioinform. 7 (2010), pp. 108–117.

[21] H. Li, C.Y. Ung, C.W. Yap, Y. Xue, Z.R. Li, Z.W. Cao, and Y.Z. Chen, *Prediction of genotoxicity of chemical compounds by statistical learning*, Chem. Res. Toxicol. 18 (2005), pp. 1071–1080.

[22] S.W. Chen, Z.R. Li, and X.Y. Li, *Prediction of antifungal activity by support vector machine approach*, J. Mol. Struct. (THEOCHEM) 731 (2005), pp. 73–81.

[23] S. Ajmani, K. Jadhav, and S.A. Kulkarni, *Three-dimensional QSAR using the k-nearest neighbor method and its interpretation*, J. Chem. Inf. Model. 46 (2006), pp. 24–31.

[24] R. Kohavi and G.H. John, *Wrappers for feature subset selection*, Artif. Intell. 97 (1997), pp. 273–324.

[25] T. Matsuda, I. Terashima, Y. Matsumoto, H. Yabushita, S. Matsui, and S. Shibutani, *Effective utilization of N2-ethyl-2′-deoxyguanosine triphosphate during DNA synthesis catalyzed by mammalian replicative DNA polymerases*, Biochemistry 38 (1999), pp. 929–935.

[26] H.F. Chang, D.M. Huffer, M.P. Chiarelli, L.R. Blankenship, S.J. Culp, and B.P. Cho, *Characterization of DNA adducts derived from syn-benzo[ghi]fluoranthene-3,4-dihydrodiol-5,5a-epoxide and comparative DNA binding studies with structurally-related anti-diolepoxides of benzo[ghi]fluoranthene and benzo[c]phenanthrene*, Chem. Res. Toxicol. 15 (2002), pp. 198–208.

[27] S. Eberhard, I. Karen, E. Hirsch, S. Ekkehard, F.K. Georg, and F. Heidi, *Metabolism of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) in primary cultures of rat alveolar type II cells*, Drug Metab. Dispos. 28 (2000), pp. 180–184.

[28] S.S. Hecht, *Preparation of pyridine-N-glucuronides of tobacco-specific nitrosamines*, Chem. Res. Toxicol. 14 (2001), pp. 555–561.

[29] M.J. Newman, B.A. Light, A. Weston, D. Tollurud, J.L. Clark, D.L. Mann, J.P. Blackmon, and C.C. Harris, *Detection and characterization of human serum antibodies to polycyclic aromatic hydrocarbon diol-epoxide DNA adducts*, J. Clin. Invest. 82 (1988), pp. 145–153.

[30] C.F. Lacy, L.L. Armstrong, M.P. Goldman, and L.L. Lance, *Drug Information Handbook*, Vol. 2003, 10th Anniversary ed., Lexicomp, Hudson, Cleveland, 2004.

[31] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000, pp. 50–70.

[32] A.R. Katritzky and E.V. Gordeeva, *Electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research*, J. Chem. Inf. Comput. Sci. 33 (1993), pp. 835–857.

[33] L.B. Kier and L.H. Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, 1999.

[34] M. Karelson, V.S. Lobanov, and A.R. Katritzky, *Quantum-chemical descriptors in QSAR/QSPR studies*, Chem. Rev. 96 (1996), pp. 1027–1043.

[35] Z.R. Li, L.Y. Han, Y. Xue, C.W. Yap, H. Li, L. Jiang, and Y.Z. Chen, *Model-molecular descriptor lab: A web-based sever for computing structural and physicochemical feature of compounds*, Biotechnol. Bioeng. 97 (2007), pp. 389–396.

[36] H.P. Schultz, *Topological organic chemistry. 1. Graph theory and topological indices of alkanes*, J. Chem. Inf. Comput. Sci. 29 (1989), pp. 227–228.

[37] L.H. Hall and L.B. Kier, *Electrotopological state indices for atom types: A novel combination of electronic, topological and valence state information*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 1039–1045.

[38] R.S. Pearlman and K.M. Smith, *Novel software tools for chemical diversity*, Persp. Drug Disc. Des. 9–11 (1998), pp. 339–353.

[39] J. Caballero and F.D.F.M. González-Nilo, *Structural requirements of pyrido[2,3-d]pyrimidin-7-one as CDK4/D inhibitors: 2D autocorrelation, CoMFA and CoMSIA analyses*, Bioorg. Med. Chem. 16 (2008), pp. 6103–6115.

[40] A. Golbraikh and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*, J. Comput. Aid. Mol. Des. 16 (2002), pp. 357–369.

[41] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, and K.A. Pebble, *Artificial neural networks in classification of NIR spectral data: Design of the training set*, Chemometr. Intell. Lab. Syst. 33 (1996), pp. 35–46.

[42] R.W. Kennard and L.A. Stone, *Computer aided designs of experiments*, Technometrics 11 (1969), pp. 137–148.
[43] B. Bourguignon, P.F. De Aguiar, K. Thorre, and D.L. Massart, *Optimization in irregularly shaped regions: pH and solvent strength in reversed-phase high-performance liquid chromatography separations*, J. Chromatogr. Sci. 32 (1994), pp. 144–152.
[44] D.D. Claeys, T. Verstraelen, E. Pauwels, C.V. Stevens, M. Waroquier, and V. Van Speybroeck, *Conformational sampling of macrocyclic alkenes using a Kennard–Stone-based algorithm*, J. Phys. Chem. A. 114 (2010), pp. 6879–6887.
[45] H. Liu, E. Papa, and P. Gramatica, *QSAR prediction of estrogen activity for a large set of diverse chemicals*, Chem. Res. Toxicol. 19 (2006), pp. 1540–1548.
[46] A. Tropsha, P. Gramatica, and V.K. Gombar, *The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.
[47] H. Liu and P. Gramatica, *QSAR study of selective ligands for the thyroid hormone receptor β*, Bioorgan. Med. Chem. 15 (2007), pp. 5251–5261.
[48] P. Gramatica, *Principles of QSAR models validation: Internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.
[49] B. Bourguignon, P.F. De Aguiar, M.S. Khots, and D.L. Massart, *Optimization in irregularly shaped regions: pH and solvent strength in reversed-phase HPLC separations*, Anal. Chem. 66 (1994), pp. 893–904.
[50] J.J. Perez, *Managing molecular diversity*, Chem. Soc. Rev. 34 (2005), pp. 143–152.
[51] P. Willett, J.M. Barnard, and G.M. Downs, *Chemical similarity searching*, J. Chem. Inf. Comput. Sci. 38 (1998), pp. 983–996.
[52] P. Willett and V.A. Winterman, *Comparison of some measures for the determination of intermolecular structural similarity*, Quant. Struct. Act. Relat. 5 (1986), pp. 18–25.
[53] The NCI Diversity Set II, Available at http://dtp.nci.nih.gov/branches/dscb/div2_explanation.html (accessed on January 9, 2010)
[54] C. Li, L. Xu, D.W. Wolan, I.A. Wilson, and A.J. Olson, *Virtual screening of human-aminoimidazole-4-carboxamide ribonucleotide transformylase against the NCI diversity set by use of AutoDock to identify novel nonfolate inhibitors*, J. Med. Chem. 47 (2004), pp. 6681–6690.
[55] Y.W. Chen and C.J. Lin, *Combining SVMs with various feature selection strategies*, (2005), Available at http://www.csie.ntu.edu.tw/∼cjlin/papers/features.pdf
[56] C.L. Huang, M.C. Chen, and C.J. Wang, *Credit scoring with a data mining approach based on support vector machines*, J. Expert. Syst. Appl. 4 (2007), pp. 2870–2878.
[57] M.W.B. Trotter, B.F. Buxton, and S.B. Holden, *Support vector machines in combinatorial chemistry*, Meas. Contr. 34 (2001), pp. 235–239.
[58] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, and Y.Z. Chen, *Prediction of p-glycoprotein substrates by support vector machine approach*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 1497–1505.
[59] C.J.C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Min. Knowl. Disc. 2 (1998), pp. 127–167.
[60] A.R. Katritzky and E.V. Gordeeva, *Raditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research*, J. Chem. Inf. Comput. Sci. 33 (1993), pp. 835–857.
[61] S. Ajmani, K. Jadhav, and S.A. Kulkarni, *Three-dimensional QSAR using the k-nearest neighbor method and its interpretation*, J. Chem. Inf. Model. 46 (2006), pp. 24–31.
[62] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, *Drug design by machine learning: Support vector machines for pharmaceutical data analysis*, Comput. Chem. 26 (2001), pp. 5–14.
[63] R. Czerminski, A. Yasri, and D. Hartsough, *Use of support vector machine in pattern application to QSAR studies*, Quant. Struct. Act. Relat. 20 (2001), pp. 227–240.
[64] C.J. Huberty, *Applied Discriminant Analysis*, John Wiley & Sons, New York, 1994.
[65] E. Fix and J.L. Hodges, *Discriminatory analysis: Nonparametric discrimination: Consistency properties*, USAF School of Aviation Medicine, Randolph Field, TX, 1951, pp. 261–270.
[66] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1982.
[67] J. Yuan and W. Chen, *A gamma dose distribution evaluation technique using the k-d tree for nearest neighbor searching*, Med. Phys. 37 (2010), pp. 4868–4873.
[68] H. Wang, *Anal Mach Intell Nearest neighbors by neighborhood counting*, IEEE Trans. Patt. 28 (2006), pp. 942–953.
[69] A. Givehchi and G. Schneider, *Impact of descriptor vector scaling on the classification and nondrugs with artificial neural networks*, J. Mol. Model. 10 (2004), pp. 204–211.
[70] W. Vach, R. Robner, and M. Schumacher, *Neural networks and logistic regression: Part I*, Comput. Stat. Data Anal. 21 (1996), pp. 683–701.
[71] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley, New York, 1989.
[72] H.A. Bruck, S.R. McNeill, M.A. Sutton, and W.H. Peters, *Digital-image-correlation using Newton-Raphson method for partial differential correction*, Exp. Mech. 29 (1989), pp. 261–267.
[73] B. Chen, R.F. Harrison, G. Papadatos, P. Willett, D.J. Wood, X.Q. Lewell, P. Greenidge, and N. Stiefl, *Evaluation of machine-learning methods for ligand-based virtual screening*, J. Comput. Aided Mol. Des. 21 (2007), pp. 53–62.
[74] J. Hert, P. Willett, and D.J. Wilton, *New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching*, J. Chem. Inf. Model. 46 (2006), pp. 462–470.
[75] P. Willett and D. Wilton, *Virtual screening using binary kernel discrimination: Analysis of pesticide data*, J. Chem. Inf. Model. 46 (2006), pp. 471–477.
[76] B. Chen, R.F. Harrison, K. Pasupa, P. Willett, D.J. Wilton, D.J. Wood, and X.Q. Lewell, *Virtual screening using binary Kernel discrimination: Effect of noisy training data and the optimization of performance*, J. Chem. Inf. Model. 46 (2006), pp. 478–486.
[77] P. Willett and D. Wilton, *Prediction of ion channel activity using binary Kernel discrimination*, J. Chem. Inf. Model. 47 (2007), pp. 1961–1966.
[78] J. Aitchison and C.G.G. Aitken, *Multivariate binary discrimination by the kernel method*, Biometrika 63 (1976), pp. 413–420.
[79] J.E. Roulston, *Screening with tumor markers: Critical issues*, Mol. Biotechnol. 20 (2002), pp. 153–162.
[80] B.W. Matthews, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochim. Biophys. Acta. 405 (1975), pp. 442–451.
[81] A.R. Leach and V.J. Gillet, *An Introduction to Chemoinformatics*, 2nd ed., Springer, Netherlands, 2007.
[82] R.S. Pearlman, K.M. Smith, H. Kubingi, T. Martin, and G. Folkers (eds.), *3D-QSAR and Drug Design: Recent Advances*, Kluwer Academic, Dordrecht, Netherlands, 1997.
[83] F.R. Burden, *Molecular identification number for substructure searches*, J. Chem. Inf. Comput. Sci. 29 (1989), pp. 225–227.
[84] L.H. Hall, B.K. Mohney, and L.B. Kier, *The electrotopological state: Structure information at the atomic level for molecular graphs*, J. Chem. Inf. Comput. Sci. 31 (1991), pp. 76–82.
[85] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press, Wiley, Letchworth, Hertfordshire; New York, 1986.
[86] L.H. Hall and L.B. Kier, *The molecular connectivity chi indices and kappa shape indices in structure-property modeling*, in *Reviews of Computational Chemistry*, K.B. Lipkowitz, and D.B. Boyd, eds., Vol. 2, VCH Publishers, New York, 1991, pp. 367–412.
[87] T. Kogej, O. Engkvist, N. Blomberg, and S. Muresan, *Multi-fingerprint based similarity searches for targeted class compound selection*, J. Chem. Inf. Model. 46 (2006), pp. 1201–1213.
[88] G.A. Bakken and P.C. Jurs, *Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis*, J. Med. Chem. 43 (2000), pp. 4534–4541.
[89] F. Wang, C.E. Elmquist, J.S. Stover, C.J. Rizzo, and M.P. Stone, *DNA sequence modulates the conformation of the food mutagen 2-amino-3-methylimidazo[4,5-f]quinoline in the recognition*

sequence of the NarI restriction enzyme, Biochemistry 24 (2007), pp. 8498–8516.

[90] M.H. Baik, R.A. Friesner, and S.J. Lippard, *Theoretical study of cisplatin binding to purine bases: Why does cisplatin prefer guanine over adenine as substrate?* J. Am. Chem. Soc. 125 (2003), pp. 14082–14092.

[91] A. Luch, *On the impact of the molecule structure in chemical carcinogenesis*, EXS 99 (2009), pp. 151–179.

[92] N. Nikolova-Jeliazkova and J. Jaworska, *An approach to determining applicability domains for QSAR group contribution models: An analysis of SRC KOWWIN*, Altern. Lab. Anim. 33 (2005), pp. 461–470.

[93] A. Tropsha and A. Golbraikh, *Predictive QSAR modeling workflow, model applicability domains, and virtual screening*, Curr. Pharm. Des. 13 (2007), pp. 3494–3504.

[94] L. Eriksson, J. Jaworska, A. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs*, Environ. Health Persp. 111 (2003), pp. 1351–1375.