Double Min-Score (DMS) Algorithm for Automated Model Selection in Predictive Toxicology

Anna Wojak, Daniel Neagu and Mick Ridley
School of Computing, Informatics and Media, University of Bradford

***Abstract*— *In silico* modelling is considered a cost efficient alternative to *in vivo - in vitro* testing and it allows the identification of toxic effects in early stages of product development. To speed up this discovery process while reducing the need of lab tests, a collection of high quality data and models is required. Data integration, data quality assessment and model management are current challenges for predictive toxicology in order to make data and models re-usable sources of information. One of model management components is model selection. This paper presents a novel concept of an automated model selection framework that leads to an efficient re-usage of existing models in predictive toxicology. This study demonstrates that the new classification methods, used for model selection, improve the quality of the test set, and unseen chemical compound prediction.**

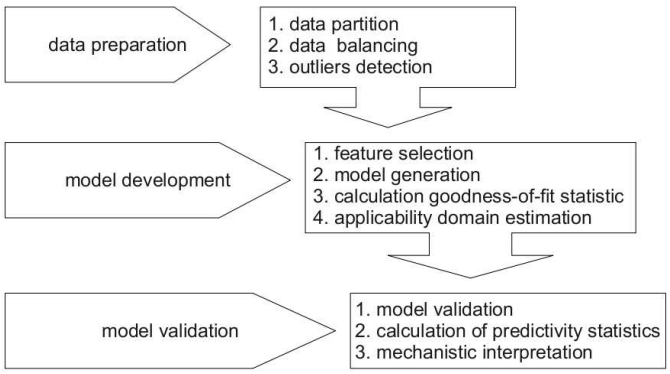
I. Introduction

Predictive toxicology is concerned with the development of models that are able to predict the toxicity of chemicals [1]. Toxicological experimental studies and development of molecular biology allow the interpretation of chemical actions on biological systems [2]. Various bio-assays and chemical descriptors are combined with computational approaches to understand, explain or predict the potential toxicity of chemicals [3], [4]. Methods that are used to generalize experimental data in order to design or optimize new biologically active compounds that are more effective, selective, less toxic, or satisfy various toxicological criteria, are called Quantitative Structure-Activity Relationship (QSAR) models.

There are several chemical compound representations and thousands of available chemical descriptors [5], [6]. The chemical descriptors represent physical and chemical properties that are calculated from the molecular representation of a chemical compound. QSAR models correlate the chemical structure and properties with its biological, chemical or environmental activity [7]. These models can be represented as regression models which deal with continuous data, and classifiers to assess an activity of a given chemical compound in terms of toxicity.

Several statistical and data mining techniques are used to build predictive models. Recently, many QSAR modelling tools have been developed using: Partial Least Squares Regression (PLS) [8], Decision Trees [9], K-Nearest Neighbors (KNN) [10], Support Vector Machines (SVM) [11], Artificial Neural Networks (ANN) [1], [12] and Random Forest [13], [14] to name a few methods. Such a large number of available techniques allow to capture various relevant chemical properties according to the observed chemical activity. Then ensemble techniques can be applied in order to provide more accurate predictions [15], [16].

Fig. 1. The model development process.



Generally, the QSAR model development process is divided into three steps: data preparation, analysis and validation (see Figure 1). Several rules and "best practices" for high quality predictive models development have been recently discussed [17]. Data preparation requires proper data curation that leads to increased quality of collected information and this has further impact on model development in terms of model predictivity and reliability. Many techniques are used in model generation including: feature selection, cross-validation, and bootstrapping. These techniques can capture various structure-activity relationships and result in a collection of models that may differ on subsets of chemical descriptors, applicability domain [7], [18], [19] and model performance.

Model selection has been studied in the context of model validation [20]. According to the Organisation for Economic Co-operation and Development (OECD) Principles for QSAR Model Validation [21], a model should be statistically significant and robust, have its application boundaries defined and be validated by an external dataset [22], [23], [24]. Various methods for internal and external model validation and conditions when the model is considered highly predictive were established. These conditions support model selection from a collection of newly generated models and are widely applied in automated model development frameworks [25] such as OpenTox [26] or InkSpot [27].

Consequently, a large number of highly predictive models has become publicly available in recent years. Due to current requirements in the protection of human and environmental health, many institutions have become interested in making

use of existing toxicity information. They are furthermore focused on toxicity data and model integration. Such a collection can be used in *in silico* modelling to detect toxic effects in early stages of the product development. This will speed up the discovery process and it will reduce the volume of animal testing. Therefore, data integration, data quality assessment and model aggregation have currently become the main challenges in predictive toxicology [28].

A good example of a large collection of QSAR models is the JRC QSAR Models Database [29]. This database includes reports of model generation and validation according to the OECD standards. Each submitted model is reviewed and then published in the model database. The main goal of a such database is to provide the sufficient documentation of QSAR models used in *in silico* modelling to the regulatory bodies like the European Commission of Registration, Evaluation and Authorization of Chemicals (REACH) [30]. Models that are stored in the database can be reused to predict toxicity of new chemical compounds. Unfortunately, this involves a manual process of model selection. A potential user is required to make a comparison of models applicability domain and models predictivity for a given endpoint, before decides to use one of them. Additionally, models are represented by equations or links to authors sources or papers. In this case, users have to implement model by themselves or contact authors to get an access to the running model source. This makes model identification for new chemicals a difficult task.

Based on the above observation, we noticed that there is a lack of automated model mining techniques. There is also a need to develop an interoperable framework for model governance that can become a powerful tool for intelligent model management. This will support model quality control, data and model integration, model comparison and model selection. Our research addresses this gap. In this paper, we introduce a novel concept of automated model selection framework, that allows an effective model identification for a new unseen chemical compound. We also demonstrate that the new classification methods, used for model selection, improve the quality of the test set, or unseen chemical compound prediction.

The paper is organized as follows: Section II defines the concept of automated model selection. The partitioning model that splits the chemical space into m disjoint groups, where m is a number of available predictive models, is introduced. In Section III, the Double Min-Score algorithm is described. It is a classification algorithm that allows to assign a given chemical compound to one of m models. In Section IV the experimental results and model accuracy are discussed. Section V includes conclusions and the further work.

II. Model Selection

A *chemical space* X is defined as a set of pairs $x = (x^d, x^f)$, where $x^d \in \mathbb{R}^{K_1}$ represents descriptors, $x^f \in \{0,1\}^{K_2}$ is a fingerprint, and $K_1 + K_2$ is the dimension of the chemical space. Descriptors represent various topological, geometrical, physio-chemical properties of a chemical compound. A fingerprint is a binary vector whose coordinates define the presence or absence of predefined structural fragments within a molecule [6]. A fingerprint is also a one dimensional representation of the molecular descriptor and it is widely used for the chemical similarity search in large databases [31]. It is also worth noting that a fingerprint is not a unique chemical compound representation because it encodes only a fragment of a molecule. There can be two different molecules having the same fingerprint representation.

A *predictive model* M is a mapping $X \rightarrow Y$, where $Y \subset \mathbb{R}$ is the output space. The output space Y might, for example, represent a particular biological, physical or chemical activity of a chemical compound.

A *model applicability domain* is defined as the activity and the chemical space in which a model makes reliable predictions [7], [18], [19]. The applicability domain determines the boundary of the chemicals space where models are reliable and it also supports the controlled extrapolation of models into the entire chemical space. This fact ensures that the QSAR model can be used for chemicals which fall into its applicability domain and at the same time it does not guarantee the model's high predictivity. Applying models for chemicals from outside of their applicability domains increases the likelihood of inaccurate prediction.

Predictive models are generated using subsets of the chemical space, for example, a particular group of chemicals. Consequently, a large number of models can be defined by the same, overlapping or disjoint applicability domains. Additionally, the model performance is described by the predictive squared coefficient correlation q^2 [22], [23]. This correlation represents the relationship between the trained model and the model that uses the average of the chemical activity as a predictor for the validation dataset. It is easy to notice that the size and content of such datasets may differ for various models. Thus, the value of q^2 is not a sufficient condition in model comparison. Several model performance matrices were analysed in the context of model validation and model selection [32]. They are applied in automated model development where models are validated by the same dataset. In a case where two models come from different sources, model comparison becomes challenging (the mean of the activity of the validation datasets may differ). This requires predictive models to be validated across the entire chemical space, which is very difficult when the list of available chemicals and assays may be limited.

QSARs represent predictive models derived from various data mining or statistical techniques and currently they are being applied in many domains (e.g risk assessment, toxicity prediction, and regulatory decisions [33]). European Commission REACH legislation [30] allows the registration of chemicals which were tested using, inter alia, virtual screening tools. Therefore, there is a need to develop a framework for data and model integration, automated model quality control, and intelligent model management and mining. In this section we present a concept of an automated

model selection framework that supports model validation and model selection. An input data is represented by the pairs:

$$(x_i, y_i) \in X \times Y, \qquad i = 1, \ldots, n,$$

where x_i is an element of the chemical space and y_i is the measured activity of that element. There is also a set of m predictive models $\mathcal{M} = \{M_1, \ldots, M_m\}$ associated with the activity Y. We assume that these models were generated using various statistical or data mining techniques and they have different applicability domains and performances. To select the most predictive model from the model family $\mathcal{M}$ for a new chemical compound x, we define a novel *partitioning model* that splits the chemical space into disjoint groups and allows for unambiguous model selection.

Definition 2.1: A *partitioning model* $\hat{M}$ is a mapping $X \to Y$ given by the following formula:

$$\hat{M}(x) = \begin{cases} M_1(x), & x \in D_1, \\ M_2(x), & x \in D_2, \\ \vdots, & \vdots, \\ M_m(x), & x \in D_m, \end{cases}$$

where

- $D_1, \ldots, D_m \subseteq X$ are disjoint,
- $\bigcup_{i=1}^{m} D_i = X$.

A computer representation of the partitioning model is a demanding task due to the size of the chemical space – one has to store the sets $D_1, \ldots, D_m$. The partitioning model aims at dividing the chemical space in such a way that every element $x \in X$ is assigned to the model, from the set of available models, with the highest predictive power. This task is clearly infeasible as the set X is large whereas available information is limited. Therefore, we concentrate on approximate solutions to build the partitioning model.

III. Double Min-Score Algorithm

The construction of the partitioning model is a similarity-based classification problem, that assigns a given chemical compound to the most predictive model. The similarity-based classifier estimates the class label of a test item using similarities between the test item and a set of labelled training items [34]. While most learning methods derive a set of classification rules from training data, in our method the classification is obtained by applying a pre-defined classification function on a given dataset. This function is a combination of the chemical compounds similarity and model performance. According to the Definition 2.1, the partitioning model splits the chemical space in groups in order to maximize the similarity of their chemical compounds and to minimize the error of a model associated with this group. We will call such a defined group - *a model group*. It is easy to notice that this is a bi-criteria problem and the solutions have to represent a trade-off between optimality of these criteria (the so-called Pareto points[35]). Pareto optimality is a multicriteria optimisation problem. In this paper, we simplify this problem to the one-criteria problem based on the chemical compound similarity hypothesis [36] which states that similar compounds have similar properties. We define a mapping between the chemical space and a set of model indexes using the Double Min-Score (DMS) algorithm (see Algorithm 1). In this section two classification rules are introduced to predict which model is the most predictive for a chemical compound x.

Algorithm 1 Double Min-Score Algorithm

Input: A dataset T, a family of models $\mathcal{M}_T$ and a new data x.
Output: The most predictive model M.
Step 1: Calculate the error $e_{i,j}$ for every model M_j and every item x_i in dataset T.
Step 2: Split the dataset T into m disjoint model groups.
Step 3: Calculate the nearest neighbourhood of x.
Step 4: Select the model M_j assigned with the nearest neighbour of x.

Let's consider a dataset T of pairs $(x_i, y_i) \in X \times Y$, where $i = 1, \ldots, n$, and the family of predictive models $\mathcal{M}_T$. In Step 1 of the DMS algorithm presented above, the error $e_{i,j}$ of the model M_j for the i-th data item is defined as follows:

$$e_{i,j} = |y_i - M_j(x_i)|, \tag{1}$$

where $i = 1, \ldots, n$ and $M_j \in \mathcal{M}_T$ for $j = 1, \ldots, m$.

In the next step we define a mapping of the chemical space into a set of model indexes $D : X \leftarrow \{1, 2, \ldots, m\}$. First, we define a mapping D on the dataset T in such a way that for each $x_i \in T$:

$$\begin{aligned} D(x_i) \quad = \quad & \min\{j \in \{1, 2, \ldots, m\} : e_{i,j} = \\ & \min\{e_{i,l} : l = 1, \ldots, m\}\} \end{aligned} \tag{2}$$

In this step, a class (a model index) is defined for elements in the dataset T. In Step 2 of Algorithm 1 we divide a dataset into m disjoint sets. According to formula (2), each data item $x_i \in T$ is assigned to the model that has the minimal error defined by formula (1) over all available models. In a case where more than one model has the same predictive error, the model with the lowest index is chosen.

In the next step, the mapping D is extended to the whole chemical space X in the following way: for $x \in X$

$$\begin{aligned} D(x) \quad = \quad & \min\{D(x_i) : \rho(x^f, x_i^f) = \\ & \min\{\rho(x^f, x_k^f) : k = 1, \ldots, n\}, i = 1, \ldots, n\} \end{aligned} \tag{3}$$

where $D(x_i)$ is defined by formula (2) and ρ is the fingerprint-based similarity coefficient (widely used in chemical similarity searching [37]). The DMS algorithm uses only the molecular similarity of chemicals and does not require knowledge of the model applicability domain. In this stage (Steps 3-4, Algorithm 1) the nearest neighbourhood of x is calculated. Then, the element x is assigned to the model group of its nearest neighbour x_i according to formula (3). The selected model can be applied on x to predict its activity y.

It is worth noting that the automated model selection framework can also be used for the applicability domain estimation. The partitioning model groups chemicals according to the model performance, and then ranges for model descriptors can be easily obtained from the chemical space X. In the next section we present the experimental results. We show that the partitioning model defined in Section II leads to increased prediction accuracy (in terms of ensemble modelling).

IV. Results

The partitioning model can be used in terms of existing model selection in predictive toxicology. To show how the proposed model increases the prediction accuracy, the following experiment was carried out. A dataset of 1129 chemicals was obtained from [38]. This dataset consists of a compilation of toxicity data for the unicellular ciliated protozoa Tetrahymena pyriformis collected from results of the assay described by the Schultz group in [39]. The measure of toxicity is 50% growth inhibition concentration (IGC50). This measure is calculated by linear regression of the percent control-normalized absorbency and toxicant concentration in mg/lm [40]. Additionally, two QSAR regression models were obtained from [38]. The first, non polar narcosis QSAR [41], was trained on 87 chemicals identified as non polar narcotics with $q^2 = 0.95$. The linear regression model was defined as follows:

$$\log(1/IGC50) = 0.83 \log P - 2.07,$$

where $\log P$ is the octanol-water partition coefficient. The second, polar narcosis QSAR model [42] for Tetrahymena pyriformis, was trained on 138 polar narcotics chemicals with $q^2 = 0.75$ and defined as follows:

$$\log(1/IGC50) = 0.62 \log P - 1.00$$

These models were used to build the partitioning model, without having knowledge of their applicability domain. We can notice that these models were trained on a small subset of the chemical space, therefore, the extrapolation of QSAR models on the entire dataset decreased model accuracies.

Let consider a new group of chemicals for which we would like to predict the toxicity for Tetrahymena, using one of the existing models developed on the available dataset. In the next section we present a simulation of toxicity forecasting for the randomly excluded chemicals from the original dataset.

A. Model performance

The input dataset was split randomly into training and validation datasets. The training set contains 831 chemical compounds, whereas the test set contains remaining 298 chemicals. $\log P$ was calculated from SMILES [5] using the CDK library [43]. Non polar narcosis and polar narcosis models were applied for the training dataset, and errors were calculated according to formula (1). The Double Min-Score algorithm was used to assign chemicals with the most predictive model. The fingerprint-based similarity ρ used in formula (3) was calculated using the Tonimoto coefficient [37]:

$$\rho(x_1^f, x_2^f) = \frac{a}{a + b - c}$$

where a is a number of bit set in x_1^f and x_2^f, b is a number of bit set only in x_1^f and c is a number of bits set in x_2^f. This is the most common measure used in chemical compound similarity search.

TABLE I

Model performance metrics.

Model Name	q^2	MAE	RSME
PART	**0.6670022**	**0.3821944**	**0.5404472**
NPN	0.3283711	0.5674606	0.7675334
PN	0.6073535	0.4762509	0.5868587

TABLE II

Confusion matrix for the DMS algorithm

obs/pred	Non polar Narcosis	Polar narcosis
Non polar narcosis	86	31
Polar narcosis	48	133

To demonstrate how the partitioning model leads to an increased prediction accuracy - the validation dataset was used. As was mentioned above, the chemicals were chosen randomly without the knowledge of the model applicability domain. For a given chemical compound the DMS algorithm was used to select the most predictive model. The polar narcosis (PN), non polar narcosis (NPN) and partitioning model (PART) performances were analysed by comparing the predictive squared coefficient correlation q^2, minimum absolute error *MAE* and root mean squared error *RMSE* [22], [23], [44]. Table I illustrates the model performance metrics. In literature, a model is considered predictive when $q^2 > 0.5$. We can observe that q^2 for NPN is below this threshold. The partitioning model can be compared with the polar narcosis model while they were validated on the same dataset. In this case, the PART model has better prediction performance than the PN model but the difference is small.

To validate the DMS algorithm an *oracle model* was introduced. The oracle model maps the chemicals space into the set of model indexes according to formula (2). The confusion matrix is presented in Table II. The number of correctly classified items was 219 (73.49%), whereas the number of wrongly classified chemicals is 79 (26.51%).

The following statistics were calculated for the DMS algorithm:

- Sensitivity=$\frac{tp}{(tp+fn)} = 64\%$
- Specificity =$\frac{tn}{(tn+fp)} = 81\%$
- Accuracy= $\frac{tp+tn}{(tp+fn+fp+tn)} = 74\%$

where specificity and sensitivity define model robustness and the accuracy demonstrates its predictivity.

In the next step we restricted the validation dataset by removing items for which the oracle model returns the absolute error greater than 0.5. The remaining set of 190 chemicals

TABLE III

MODEL PERFORMANCE METRICS II.

Model Name	q^2	MAE	RSME
PART	**0.8414306**	**0.265396**	**0.3446489**
NPN	0.6399403	0.3961324	0.5193432
PN	0.6876759	0.4031394	0.4836931

TABLE IV

CONFUSION MATRIX FOR THE DMS ALGORITHM II

obs/pred	Non polar Narcosis	Polar narcosis
Non polar narcosis	74	15
Polar narcosis	35	66

was tested and the following statistics were obtained (see Table III). We can notice, that prediction performances increased for all models and there is a significant difference for the partitioning model performance compared with the results in Table I. In this case, q^2 for the partitioning model is 0.84 - much higher than for the PN and NPN models.

Analysis of the DMS algorithm shows that the number of correctly classified chemicals is 140 (73.68%) and the number of incorrect classification is 50 (26.32%) (see Table IV). The following statistics were calculated:

- Sensitivity=$\frac{tp}{(tp+fn)} = 67\%$
- Specificity =$\frac{tn}{(tn+fp)} = 81\%$
- Accuracy= $\frac{tp+tn}{(tp+fn+fp+tn)} = 74\%$

In this case the sensitivity measure is higher than this presented in the previous experiment, whereas the accuracy remains the same. It is related to the applicability domains of the NPN and PN models. The number of chemicals which were used to trained the QSAR models states 20% of entire dataset. The validation dataset contained only 5% of them. The experiment showed that model selection allows an extrapolation of the applicability domain into a larger group of chemicals on which the partitioning model gives the reliable prediction. We can compare the accuracy of the partitioning model with the consensus model presented in [16]. The consensus model was build as an average of 15 models developed by various institutions. The accuracy of the consensus model obtained in the validation process was 0.85 and MAE=0.29 for chemicals from the models applicability domains and 0.67 and MAE=0.39 for chemicals for which models did not give reliable predictions. Our approach gave similar results, what proves that the probability of the selected model reliability and predictivity for the new unseen chemicals, is high. This is very important aspect in the research phase of a new chemical development, because allows early elimination of unsuccessful chemicals, reducing the number and cost of the further *in vitro - in vivo* testing.

The proposed DMS algorithm is the first approach towards automated model selection in QSAR model collections. Although the similarity measure is based on a chemical fingerprint, the algorithm has a relatively high accuracy of the model identification. This is an important outcome because fingerprints do not uniquely identify chemical molecules (two different chemicals can be represent by the same fingerprint). Moreover, the DMS algorithm as a similarity-based classifier is over-fitting-resistant: the classification rules are pre-defined and not derived from data. The Double-Mean-Score algorithm was implemented in the R programming environment. The QSAR models were stored in a model database as programming objects (.rda files), that allowed an easy retrieval and application of models for new chemicals.

Although the DMS algorithm concept is based on a naive similarity-based approach, these first experimental results demonstrate that the usage of proper model selection techniques could lead to better prediction performance and speed up the process of virtual screening. We also intend to study more efficient multicriteria optimisation methods in order to improve model selection accuracy. This experiment also shows that toxicity data and model integration combined within an automated model mining framework can become a powerful tool in order to make models reusable sources of information.

V. CONCLUSIONS

A novel concept of automated model selection in predictive toxicology was introduced in this paper. The authors noticed that there is not only a large amount of publicly available toxicity information but also an increasing number of good quality models. These models can be further reused in *in silico* modelling to speed up the process of high-throughput screening. Thus, there is a need to develop a framework for intelligent model management and mining.

In this paper, the authors introduced the concept of the partitioning model in terms of model selection. The authors showed that this problem is a bi-criteria classification problem. The main idea is to split the chemical space into disjoint model groups. Each group is assigned with a particular predictive model in order to maximize the similarity of chemicals and to minimize the model error within a group. To solve this problem, the Double Min-Score algorithm was proposed. The authors assumed that model performance is equal for similar chemicals according to the similarity hypothesis and they reduced the problem to a single classification problem. A new item is classified to the particular group based on the nearest neighbourhood. Additionally, the authors demonstrated that this classification technique for model selection improves the quality of the test set, or unseen chemical compound prediction. Moreover, the authors aim to continue this research to provide more efficient methods for an automated model mining framework.

ACKNOWLEDGMENTS

This work is partially supported by BBSRC and Syngenta through the Industrial CASE Partnership Grant BB/H530854/1 ”Data Mining Applications in Product Safety”. The authors would like to thank Prof. Mark Cronin and Dr. Steven Enoch from the School of Pharmacy and Chemistry, Liverpool John Moores University (LJMU), Liverpool, UK, who facilitated the access to the Tetrahymena

pyriformis dataset and QSAR models, for their time and advice.

REFERENCES

[1] C. Helma, Ed., *Predictive Toxicology*. Taylor & Francis Group, 2005.

[2] M. D. Waters and J. M. Fostel, "Toxicogenomics and systems toxicology: aims and prospects," *Nature Review Genetics*, vol. 5, no. 12, pp. 936–948, December 2004.

[3] R. Kavlock. (2003) A framework for computational toxicology research in ORD. [Online]. Available: http://www/epa.gov/comtox/comptox_framework.html

[4] R. Judson, "Public databases supporting computational toxicology," *Journal of Toxicology and Environmental Health, Part B*, vol. 13, no. 2, pp. 218–231, 2010.

[5] J. Gasteiger, Ed., *Handbook of Chemoinformatics: From Data to Knowledge*. John Wiley and Sons Inc, 2003.

[6] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*. Wiley-VCH, 2000.

[7] T. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. Cronin, P. Gramatica, J. Jaworska, S. Kahn, G. Klopman, C. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D. Stanton, J. van de Sandt, W. Tong, G. Veith, and C. Yang, "Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. the report and recommendations of ECVAM workshop 52." *ATLA. Alternatives to laboratory animals*, vol. 33, no. 2, pp. 155–73, 2005.

[8] S. Wold, M. Sjstrm, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109 – 130, 2001.

[9] A. Rusinko, M. W. Farmen, C. G. Lambert, P. L. Brown, and S. S. Young, "Analysis of a large structure/biological activity data set using recursive partitioning," *Journal of Chemical Information and Computer Sciences*, vol. 39, pp. 1017–1026, 1999.

[10] G. W. Kauffman and P. C. Jurs, " QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 6, pp. 1553–1560, 2001.

[11] R. N. Jorissen and M. K. Gilson, "Virtual screening of molecular databases using a support vector machine," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 549–561, 2005.

[12] K. Balakin and S. Ekins, *Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery - Wiley Series on Technologies for the Pharmaceutical Industry*. Wiley-Blackwell (an imprint of John Wiley and Sons Ltd), 2010.

[13] L. Breiman and E. Schapire, "Random forests," in *Machine Learning*, 2001, pp. 5–32.

[14] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[15] L. Kuncheva, *Combining pattern classifiers: Methods and Algorithms. Wiley,*. Wiley, 2004.

[16] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatical, T. Öberg, P. Dao, A. Cherkasov, and I. Tetko, "Combinatorial QSAR modeling of chemical toxicants tested against tetrahymena pyriformis." *Journal of Chemical Information and Modeling*, vol. 48, no. 4, pp. 766–784, 2008.

[17] A. Tropsha, "Best Practices for QSAR Model Development, Validation, and Exploitation," *Molecular Informatics*, vol. 29, no. 6-7, pp. 476–488, 2010.

[18] J. Jaworska, M. Comber, C. Auer, and C. Leeuwen, "Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints." *Environ Health Perspectives*, vol. 111, pp. 1358–1360, 2003.

[19] J. Jaworska, N. Nikolova-Jelizkova, and T. Aldenerg, " QSAR applicability domain estimation by projection of the training set in descriptor space: A review." *ATLA. Alternatives to laboratory animals*, vol. 33, pp. 445–459, 2005.

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[21] OECD. OECD principles for the validation , for regulatory purposes, of QSAR models. [Online]. Available: http://www.oecd.org/dataoecd/33/37/37849783.pdf

[22] A. Golbraikh and A. Tropsha, "Beware of q^2," *Journal of Molecular Graphics and Modeling*, vol. 20, pp. 269–276, 2002.

[23] A. Tropsha, P. Gramatica, and V. Gombar, "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models." *QSAR and Combinatorial Science*, vol. 22, pp. 69–77, 2003.

[24] P. Gramatica, "Principles of QSAR models validation: internal and external," *QSAR and Combinatorial Science*, vol. 26, pp. 694–7012, 2007.

[25] J. Cartmell, S. Enoch, D. Krstajic, and D. Leahy, "Automated QSPR through competitive workflow," *Journal of Computer-Aided Molecular Design*, vol. 19, pp. 821–833, 2005, 10.1007/s10822-005-9029-8.

[26] Opentox. [Online]. Available: http://www.opentox.org

[27] Inkspot. [Online]. Available: http://www.inkspotscience.com/

[28] B. Hardy, N. Douglas, C. Helma, M. Rautenberg, N. Jeliazkova, V. Jeliazkov, I. Nikolova, R. Benigni, O. Tcheremenskaia, S. Kramer, T. Girschick, F. Buchwald, J. Wicker, A. Karwath, M. Gutlein, A. Maunz, H. Sarimveis, G. Melagraki, A. Afantitis, P. Sopasakis, D. Gallagher, V. Poroikov, D. Filimonov, A. Zakharov, A. Lagunin, T. Gloriozova, S. Novikov, N. Skvortsova, D. Druzhilovsky, S. Chawla, I. Ghosh, S. Ray, H. Patel, and S. Escher, "Collaborative development of predictive toxicology applications," *Journal of Cheminformatics*, vol. 2, no. 1, p. 7, 2010.

[29] Jrc qsar model reporting format (qmrf). [Online]. Available: http://qsardb.jrc.ec.europa.eu/qmrf/

[30] Reach. [Online]. Available: http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm

[31] D. R. Flower, "On the properties of bit string-based measures of chemical similarity," *Journal of Chemical Information and Computer Sciences*, vol. 38, no. 3, pp. 379–386, 1998.

[32] Openqsar. [Online]. Available: http://www.openqsar.com/aboutModelMetrics.jsp

[33] W. Tong, H. Hong, Q. Xie, L. Shi, H. Fang, and R. Perkins, "Assessing qsar limitations - a regulatory perspective," *Current Computer - Aided Drug Design*, vol. 1, pp. 195–205(11), April 2005.

[34] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based classification: Concepts and algorithms," *J. Mach. Learn. Res.*, vol. 10, pp. 747–776, June 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1577069.1577096

[35] M. Ehrgott, *Multicriteria Optimization*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[36] M. Jahnson and G. Maggiora, *Concept of application of Molecular Similarity*. John Willey & Sons, 1990.

[37] P. Willet, J. Berdnard, and G. Downs, "Chemical similarity searching." *Journal of Chemical Information and Computer Sciences*, vol. 38, pp. 983–996, 1998.

[38] Inchemicotox. [Online]. Available: http://www.inchemicotox.org/results/

[39] T. W. Schultz, "Tetratox: Tetrahymena pyriformis population growth impairment endpointa surrogate for fish lethality," *Toxicology Methods*, vol. 7, pp. 289–309(21), 1 December 1997.

[40] Tetratox. [Online]. Available: http://www.vet.utk.edu/TETRATOX

[41] C. M. Ellison, M. T. D. Cronin, J. C. Madden, and T. W. Schultz, "Definition of the structural domain of the baseline non-polar narcosis model for tetrahymena pyriformis," *SAR and QSAR in Environmental Research*, vol. 19, October 2008.

[42] S. Enoch, M. Cronin, T. Schultz, and J. Madden, "An evaluation of global qsar models for the prediction of the toxicity of phenols to tetrahymena pyriformis," *Chemosphere*, vol. 71, no. 7, pp. 1225 – 1232, 2008.

[43] rcdk. [Online]. Available: http://cran.r-project.org/web/packages/rcdk/index.html

[44] D. J. Hand, P. Smyth, and H. Mannila, *Principles of data mining*. Cambridge, MA, USA: MIT Press, 2001.