Prediction of Autoignition Temperature for Diverse Organic Compounds using In Silico Methods and Molecular Descriptors

Georgia Melagraki[a] , Antreas Afantitis[b], and Olga Igglessi - Markopoulou[b]

[a]*Hellenic Naval Academy, Pireaus, Greece*
[b]*School of Chemical Engineering, National Technical University of Athens, Athens, Greece*

Abstract. A large dataset of 446 structurally diverse organic compounds has been used to develop mathematical models that relate the structures within this heterogeneous group of compounds to their autoignition temperature (AIT) values. For the development of such Quantitative Structure Property (QSPR) models, the molecular structure of each compound was represented by calculated molecular descriptors which encode their topological, electronic and geometric features. Correlation Feature Selection method combined with Best First evaluator was used to select the most significant descriptors that were used as inputs for the development of several models. Different modeling methodologies such as kNN, SVM and MLR were then applied and the ability of the new models to predict AIT was assessed and compared to available experimental data. The accuracy and robustness of the produced models was based on validation principles as described by the Organisation for Economic Cooperation and Development (OECD). The kNN model was proven to be the most accurate model. Moreover the applicability domain of the models based on similarity measurements has been defined to indicate reliable predictions. For structures that fall within the domain of applicability, the proposed models can be used to predict AIT values based solely on their structure.

Keywords: chemoinformatics, QSPR, molecular descriptors, autoignition temperature

INTRODUCTION

Reliable and accurate data of physicochemical properties are always required and even considered to be absolutely necessary before making a decision and investment to formulate, synthesize, scale-up, test and manufacture a new material for use in both military and civilian applications. The knowledge of autoignition temperature of toxic, volatile, explosive and radioactive compounds is essential in risk assessment calculation. The autoignition temperature (AIT) is defined as the lowest temperature at which a substance in air will ignite in the absence of a spark or flame [1] and as proposed in literature, the AIT of a compound is strongly depended upon its structure. Since AIT gives an indication of the temperature at which a material will spontaneously burst into flames when exposed to the atmosphere, it is an important fire performance parameter in process design and operational procedures. In many common situations, such as the manufacture, handling, transport, and storage of combustible materials, the AIT has been widely used to characterize the hazard potential of chemicals.

A valuable tool for developing quantitative relationships between structural characteristics of a compound and its properties is Quantitative Structure Property Relationships (QSPR). Multiple QSPRs can be developed for predicting a material's physical/chemical properties and environmental effects. The molecular structures of the compounds are represented by calculated numerical descriptors which encode their topological, electronic, and geometric features. The availability of large numbers of theoretical descriptors that provide diverse sources of chemical information are useful to better understand relationships between molecular structure and experimental evidence, also taking advantage of more and more powerful methods, computational algorithms, and fast computers. Many different physicochemical properties have been studied in the context of QSPR [2-4]. Different approaches to predict AIT values of organic compounds have been previously reported in literature [5-8]. Among the various methods proposed, QSPR modeling has also been used in order to derive models that predict AIT values from the molecular structure and therefore have the potential to provide information on hazards of chemicals, while reducing time and cost required.

Models reported in literature have been developed using different modeling workflows and although they have been shown to accurately predict AIT values they have several limitations such as the following: (i) the models are targeting a narrow structural space by i.e. focusing in a limited number of similar compounds and therefore cannot be used for generalization purposes, (ii) the models lack internal and external validation that would propose significant accuracy and robustness or they have not been properly validated following well established validation criteria, (iii) the models lack a well defined applicability domain that indicates which predictions could be considered reliable and (iv) the models rely on parameters which are not easily available for every compound or can not be accurately measured.

In this work, a quantitative structure−property relationship study is performed to develop mathematical models that relate the structures of a large heterogeneous group of organic compounds to their autoignition temperature values. The molecular structures of the compounds are represented by calculated numerical descriptors which encode their topological, electronic, and geometric features. These descriptors can be easily calculated solely from the molecular structure. Selected descriptors are used to develop several models that predict AIT based on various modelling techniques, such as k Nearest Neighbor (kNN) Support Vector Machines (SVMs), and Multiple Linear Regression (MLR). The ability of the new models to predict AIT is assessed and compared to available experimental data. All models are internally and externally validated paying special attention to the principles of model validation for accepting QSAR models as described by the Organisation for Economic Cooperation and Development (OECD). Moreover the applicability domain of the models has been defined to indicate reliable predictions.

MATERIALS AND METHODS

Data Set

The dataset that was used for this study initially includes a diverse set of 446 organic compounds and their corresponding experimental values for autoignition temperature (AIT) [8]. The dataset includes hydrocarbons, halogenated compounds, alcohols, ethers, esters, aldehydes, ketones, carboxylic acid, amines, amides, nitriles, nitro compounds and compounds with multiple functional groups. The observed AIT values for these compounds were found in the range of 170 to 680^{o}C. A few representative organic compounds included in the dataset are shown in Figure 1.

Each compound was represented by a 2D structure using MarvinSketch provided by ChemAxon [9]. For each compound in the dataset, different descriptors could be calculated to account for different structural characteristics. In this work almost 780 descriptors were calculated using Mold2 [10]. After removing useless descriptors, i.e. descriptors with no variation, by using the unsupervised attribute filter provided by Weka [11] in total 500 physicochemical constants, topological and structural descriptors were finally considered as possible input candidates to the model. Before the calculation of the descriptors, the structures were fully optimized using PM6 method in MOPAC2007 suite which, as proposed in literature, offers a good balance between computational speed and accuracy [12].

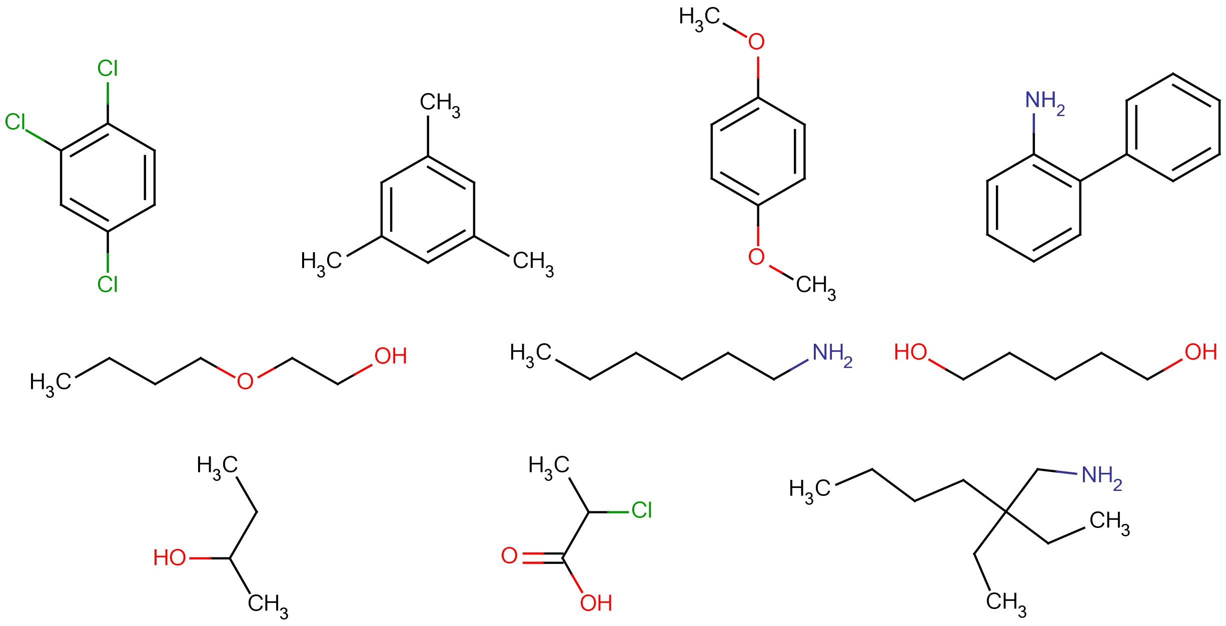

FIGURE 1. Organic compounds included in the dataset

Modeling Methodology

Variable Selection

Before running the modeling methodology the most significant attributes among the 500 available were preselected by using Correlation – based feature subset selection (CfsSubset) variable selection and BestFirst evaluator which are included in Weka [11]. CfsSubset algorithm evaluated the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that were highly correlated with the class while having low inter-correlation were preferred. The attribute selection mode was set to 10 fold cross validation.

Machine Learning Method

For preprocessing, cleansing, attribute selection, modeling and validation of our data we have created a KNIME [13] workflow suitable to run step by step all the aforementioned tasks simultaneously for each of the described modeling methodologies. KNIME is a very popular modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models. KNIME is a very powerful tool for data analysis which also integrates all analysis modules of the well known Weka data mining.

A great variety of machine learning methods have been applied in QSAR studies [14-16] and the best approach for a specific problem needs to be explored. In this work we have used KNIME platform in order to simultaneously run and compare three different modeling methodologies and explore which of the available methods (or combination) best suites our data.

We have considered the following machine learning methods available in KNIME for performing regression to our available dataset: k Nearest Neighbor (kNN), Support Vector Machines (SVM) and Multiple Linear Regression (MLR). The kNN and SVM modeling methodologies used are briefly described below:

k-Nearest neighbors (kNN) algorithm [17] is a method for classifying objects based on closest training examples in the feature space and belongs to instance-based (or lazy) learning. Based on the kNN algorithm an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (where k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. In this work we have used automatic selection of the optimal k value based on the internal cross-validation procedure. Euclidean distance was used with all descriptors and contributions of neighbors are weighted by the inverse of distance.

Support Vector Machines (SVM) was proposed in 1963 by Vapnik et al., [18] and was shown as an effective tool for solving classification and regression problems. For a given regression problem the goal of SVM is to find the optimal hyperplane for which the distance to all data points is minimum. A detailed presentation of the theory behind the SVM technique can be found in several books and tutorials [19]. For this work we have used Support Vector Machine regression (SVMreg) methodology using the RBF kernel with the complexity parameter c equal to 1000. The training data were normalized and as the learning algorithm RegSMOimproved was chosen [20].

Model Validation

The internal performance, as represented by goodness-of-fit and robustness, and the predictivity of a model, as determined by external validation, needs to be evaluated. The produced models were validated using external validation and cross validation methods [21]. External validation was applied, by randomly splitting the dataset into training and validation set in a proportion of 70:30. The 134 compounds that constituted the test set were not involved by any means in the training procedure. The following statistical criteria were used to assess the robustness, reliability and predictive activity of the model: the coefficient of determination between experimental values and model predictions (R^2), Mean Absolute Error (MAE) and Root Mean Square Error (RMS). These measurements are used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. Coefficient of determination accounts for the percentage of variation of the dependent value that is explained by the descriptors. Regarding cross validation, both 10-fold cross validation and Leave–One–Out (LOO) cross validation methods were applied. Cross-validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). In particular, the leave-one-out (LOO) and the 10-fold (10fCV) cross-validation procedures were utilized in this study, which produce a number of models, by deleting one or several objects, respectively, from the training set.

Applicability Domain

The need to define an applicability domain expresses the fact that QSPRs are models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions.

In this work similarity measurements were used to define the domain of applicability of the models based on the Euclidean distances among all training compounds and the test compounds [22]. The distance of a test compound to its nearest neighbor in the training set was compared to the predefined applicability domain (APD) threshold. The prediction was considered unreliable when the distance was higher than APD. APD was calculated as follows:

$$APD = \langle d \rangle + Z\sigma$$

Calculation of <d> and σ was perfomed as follows: First, the average of Euclidean distances between all pairs of training compounds was calculated. Next, the set of distances that were lower than the average was formulated. <d> and σ were finally calculated as the average and standard deviation of all distances included in this set. Z was an empirical cutoff value and for this work, it was chosen equal to 0.5.

RESULTS AND DISCUSSION

The original dataset of 446 diverse organic compounds was randomly partitioned into training and validation set consisting of 312 and 134 compounds respectively. The training set was used to develop the QSPR models as described below whereas the test set was not involved by any means in the model development. For each compound 777 descriptors were calculated using Mold2 software which account for the topological, geometric and structural characteristics of compounds. As some of the descriptors do not have any discrimination power (i.e. they have no variation) a filter was applied for their removal. In total 500 descriptors remained to be used as possible inputs during the QSPR model development.

The CfsSubset variable selection with BestFirst evaluator method was then applied on the training data to select the most significant, among the 500 available descriptors. Nine descriptors were selected as the most important to describe the relationship between structural characteristics of compounds and AIT. The selected descriptors are mean atomic van der Waals Carbon-scale (D144), average valence vertex connectivity order-4 Index (D222), average valence vertex connectivity order-5 Index (D223), structure centric index (D252), Moran topological structure autocorrelation length-1 weighted by atomic Sanderson electronegativities (D495), Moran topological structure autocorrelation length-3 weighted by atomic Sanderson Electronegativities (D496), number of Csp2 (D606), number of group CH2RX (D719), number of group =CHR (D729) and number of group R~CR~R (D738).

The chemical meaning of the descriptors used in the development of each model is briefly discussed below [23, 24]. The combination of these descriptors have several advantages such as unique representation of the compound and high discriminating power.

To account for steric effects in molecule interactions, the weighted information indices by volume have been selected by the algorithm (Descriptor D144). These molecular descriptors are calculated in the same way as the indices of neighborhood symmetry using the atomic van der Waals volumes to get the probabilities of the equivalence classes. In other words, the van der Waals volumes of the atoms belonging to each equivalent class are summed up to give a molecule subvolume then divided by the total molecule volume. The effective van der Waals volume of an atom is defined as the van der Waals volume of the atom minus half the sphere overlapping of the atom due to covalent bonding of the adjacent atoms in the molecule.

Descriptors, D222 and D223 account for the topological characteristics of compounds and more specifically for the connectivity between atoms within the molecule. In general topological indices such as D222 and D223 are based on the two-dimensional representation of the molecule and give information about the atomic composition of a compound, the presence and character of chemical bonds and the connectivity between atoms. D222 is the average valence vertex connectivity order-4 Index and D223 is the average valence vertex connectivity order-5 Index.

Centric indices (D252) are molecular descriptors that quantify the degree of compactness of molecules based on the recognition of the graph center centric indices. These descriptors quantify the degree of compactness of molecules by distinguishing between molecular structures organized differently with respect to their centers. Based on the recognition of the graph center, these indices are mainly defined by the information theory concepts applied to a partition of the graph vertices made according to their positions relative to the center.

Descriptors D495 and D497 encode information related to Sanderson Electronegativities combined with Moran topological index of spatial autocorrelation. Moran coefficient is related to atomic properties, the number of atoms and the topological distance between specific atoms. Electronegativity is a property of the state of the system; electrons tend to flow from a region of low electronegativity to a region of high electronegativity. With the formation of a molecule, electronegativities of the constituent atoms or fragments equalize, all becoming equal to the electronegativity of the final state of the molecule.

Descriptors D606, D719, D729 and D738 are indicators that account for the presence or absence of a specific atom or structural group. More specifically D606 is the number of secondary carbon (sp2) that are included in the compound. Descriptors D719, D729 and D738 are the number of the following groups respectively CH2RX, =CHR, R~CR~R which might be present in the compound.

The aforementioned descriptors have different weights that influence the increase or decrease of AIT values among different compounds. Based on the previous discussion and the positive or negative influence of each descriptor, new derivatives with desired properties can be designed.

We have used a KNIME workflow in order to compare different methodologies and explore which of the available methods best suites our data. As described above three different methodologies have been used, kNN, SVM and MLR.

By applying on our training data, k Nearest Neighbors (kNN) methodology with automatic selection of the optimal k value based on the internal cross-validation procedure, a k value of 4 was selected. Euclidean distance was used with all nine descriptors and contributions of neighbors weighted by the inverse of distance.

A comparison between the three modeling methodologies has been carried out (Table 1). All methodologies were employed to describe the relation between AIT and the selected descriptors. The same descriptors and training set have been used for the all three methodologies. Validation of the models was performed using the techniques mentioned in the previous section. The corresponding statistics for the three methodologies are presented in Table 1, illustrating the accuracy, significance and robustness of the produced models. As it can be derived from the Table amongst the different models the kNN model is the one that outperforms all other alternatives in external validation prediction. As can be seen from Table 1, kNN methodology results in a mean average error (MAE) as low as the experimental error of AIT determination, which is around ±30^{o}C for both training and test sets. We can conclude that the selected descriptors selected by CfsSubset and BestFirst algorithm can encode the structural features of the compounds related to AIT. Based on the produced statistics SVM model is the second best model followed by MLR.

TABLE 1. Statistical Results

	kNN	SVMreg	MLR
R^2_{train}	0.9978	0.6854	0.6519
MAE_{train}	2.9304	43.5838	47.4394
RMS_{train}	5.1559	58.8778	61.742
R^2_{LOO}	0.6170	0.6159	0.6354
MAE_{LOO}	48.8984	50.407	48.6178
RMS_{LOO}	65.4826	65.2544	63.383
R^2_{10fCV}	0.6093	0.6197	0.6277
MAE_{10fCV}	50.9991	49.5706	48.8874
RMS_{10fCV}	65.8018	64.9972	63.9038
R^2_{test}	0.8653	0.7881	0.7487
MAE_{test}	31.0896	39.9065	41.8626
RMS_{test}	38.9867	49.2699	53.1411

Figures 2 and 3 present a plot of experimental versus predicted values of AIT for compounds in the test set for the best two models, kNN and SVM model respectively.

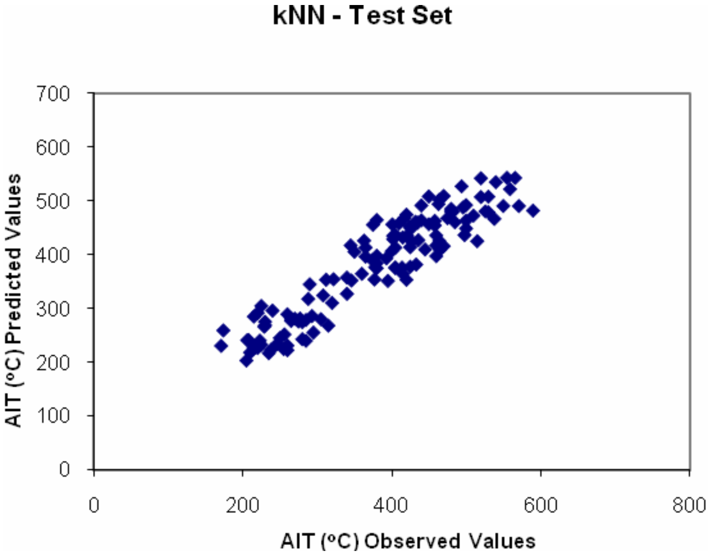

FIGURE 2. kNN model: Experimental vs Predicted AIT values for the Training Set.

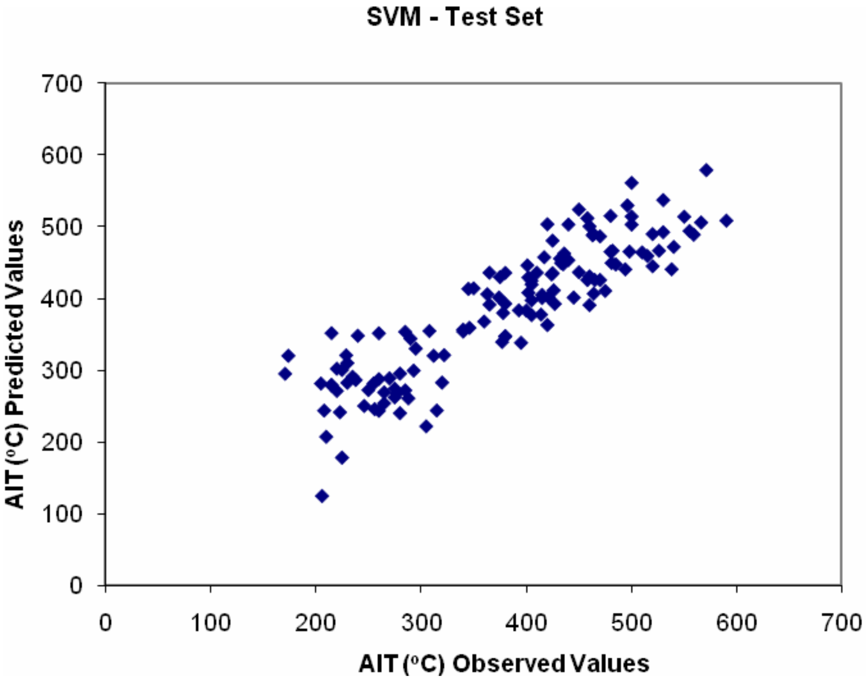

FIGURE 3. SVM model: Experimental vs Predicted AIT values for the Test Set.

The applicability domain was defined for all compounds that constituted the training set for AIT models as described in the Materials and Methods section. The applicability domain limit value was equal to 7.6995 and all compounds in the test set had values in the range of 0-4.0002. Since all validation compounds fell inside the domain of applicability, all model predictions for the external test set were considered reliable.

The proposed method, due to the high predictive ability and the fact that it requires information related only to the 2D structure of a compound, could be a useful aid to the costly and time consuming experiments for determining the AIT. The method can also be used to screen existing databases or virtual chemical structures to identify organic compounds with desired properties. In this case, the applicability domain will serve as a valuable tool to filter out “dissimilar” chemical structures.

CONCLUSIONS

In this paper we have successfully built and compared several models for the prediction of AIT based on a large dataset of 446 diverse organic compounds. All models were fully validated following OECD model validation guidance. The most accurate and reliable model was proven to be the kNN model. The molecular descriptors used in QSAR encode information about the structure, branching, electronic effects, chains and rings of the modules and thus implicitly account for cooperative effects between functional groups. Applicability domain was defined to identify the reliable predictions. The developed models can predict AIT values for compounds that fall within the domain of applicability and guide the design of novel molecules by prioritizing compounds with desired characteristics.

ACKNOWLEDGMENTS

G.M. would like to acknowledge support from ChemAxon that provided scientific software license at no cost.

REFERENCES

1. ASTM International, ASTM Standard Test Method E659-78, The American Society for Testing and Materials, West Conshohocken, PA. 2000.
2. X. Yu, X. Wang, H. Wang, X. Li, J. Gao, *QSAR Comb. Sci.* **2**, 151–61 (2005).
3. A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi – Markopoulou *QSAR Comb. Sci.* **27**, 432 – 436 (2008).
4. G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi – Markopoulou *J Mol Model* **13**, 55-64 (2007).
5. T.A. Albahri, *Chem. Eng. Sci.* **58**, 3629-3641 (2003).
6. J. Tetteh, E. Metcalfe, S. Howells, *Chemometric. Intell. Lab. Syst.* **32**, 177-191 (1996).
7. B.E. Mitchell, P.C. Jurs *J. Chem. Inf. Comput. Sci.* **37**, 538-547 (1997)
8. Y. Pan, J. Jiang, R. Wang, H. Cao, Y. Cui, *J Hazardous Materials*, **164**, 1242-1249 (2009).
9. ChemAxon Suite of programs, www.chemaxon.com
10. H. Hong, Q. Xie, W. Ge, F. Qian, F. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong, Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model* **48**,1337–1344 (2008).
11. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **11**, 10-18 (2009).
12. J.J.P. Stewart J Mol Model **14**, 499–535 (2008).
13. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization, edited by C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker, GfKl: Springer, 2007, pp. 319-326.
14. J.L. Melville, E.K. Burke, J.D. Hirst, Machine Learning in Virtual Screening. *Comb Chem High T Scr* **12**, 332-343 (2009).
15. G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, O. Igglessi – Markopoulou and G. Kollias *Molecular Diversity* **13**, 301-311 (2009).
16. G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopoulos and O. Igglessi – Markopoulou Bioorganic & Medicinal Chemistry **15**, 7237-7147 (2007).
17. Franco-Lopez H, Ek AR, Bauer ME. "Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method". *Remote Sensing of Environment* **77**, 251-274 (2001).
18. C. Cortes, V. Vapnik, *Mach Learning* **20**, 273-297 (1995).
19. C. J. C. Burges. *Data Mining and Knowledge Discovery* **2**,121–167 (1998).
20. S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, Improvements to the SMO Algorithm for SVM Regression. *IEEE T Neural Networ* **11**, 1188 – 1193 (2000).
21. OECD Principles for the validation, for regulatory purposes of (Quantitative) Structure Activity Relationship Models (www.oecd.org)
22. S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, A. Tropsha A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J Chem Inf Model* **46**, 1984-1995 (2006).
23. R. Todeschini, V. Consonni, R. Mannhold R. In: Kubinyi H, Timmerman H (Series Ed.) Handbook of molecular descriptors. Wiley-VCH, Weinheim, 2000.
24. J. Devillers, A.T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR. The Netherlands: Gordon and Breach Science Publishers, 1999.