



**HAL**  
open science

## Autoignition temperature: comprehensive data analysis and predictive models

I.I. Baskin, S. Lozano, M. Durot, G. Marcou, Dragos Horvath, A. Varnek

► **To cite this version:**

I.I. Baskin, S. Lozano, M. Durot, G. Marcou, Dragos Horvath, et al.. Autoignition temperature: comprehensive data analysis and predictive models. SAR and QSAR in Environmental Research, 2020, 31 (8), pp.597-613. 10.1080/1062936X.2020.1785933 . hal-02950552

**HAL Id: hal-02950552**

**<https://hal.science/hal-02950552>**

Submitted on 13 Nov 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## **Autoignition Temperature: comprehensive data analysis and predictive models**

Igor I. Baskin<sup>1,2,3</sup>, Sylvain Lozano<sup>4</sup>, Maxime Durot<sup>4</sup>, Gilles Marcou<sup>1</sup>, Dragos Horvath<sup>1</sup>, Alexandre Varnek<sup>1\*</sup>

<sup>1</sup> University of Strasbourg, Laboratory of Chemoinformatics, UMR 7140 CNRS/UniStra, 4, rue Blaise Pascal, 67000 Strasbourg, France

<sup>2</sup> Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya str. 18, 420008 Kazan, Russia

<sup>3</sup> Faculty of Physics, M.V. Lomonosov Moscow State University, Leninskie Gory, 119991 Moscow, Russia

<sup>4</sup> Total, Centre de Recherche de Solaize, Chemin du Canal – BP 22, 69360 Solaize, France

### **Abstract**

Here, we report a new predictive model for autoignition temperature (AIT), an important physical parameter widely used to assess potential safety hazards of combustible materials. Available structure-AIT data extracted from different sources was critically analyzed. Support Vector Regression (SVR) models on different data subsets were built in order to identify a reliable compound set on which a realistic models could be built. This led to selection of the dataset containing 875 compounds annotated with AIT values. The thereupon based SVR model performs reasonably well in cross-validation with the the determination coefficient  $R^2 = 0.77$  and mean absolute error MAE= 37.8°C. External validation on 20 industrial compounds missing in the training set confirmed its good predictive power ( $R^2 = 0.87$  and MAE= 29°C).

**Keywords:** Quantitative Structure-Property Prediction (QSPR), autoignition temperature, support vector regression, fragment descriptors, generative topographic mapping

## 1. Introduction

The autoignition temperature (AIT) is one of the key physicochemical properties, along with the flash point, the fire point, the adiabatic flame temperatures and the flammability limits, all related to combustion. It is defined as the lowest temperature at which a chemical substance spontaneously ignites in the air at the atmospheric pressure in the absence of sparks or flames. Autoignition occurs when the rate of the heat production resulting from the exothermic oxidation reaction with the oxygen in the air exceeds the rate of heat dissipation. Knowledge of the AIT is essential for defining risk-free handling procedures of combustible materials. It is also important for the design of fuels for internal combustion engines, because spontaneous ignition (engine knock) leads to the reduction in power and efficiency.

The AIT of a liquid is typically measured using a flask placed in a temperature-controlled oven according to the ASTM E659 standard procedure (<https://www.astm.org/Standards/E659.htm>). The measurement results largely depend on various experimental conditions, some easier to control than others: flask volume and shape, the material used for flask construction, air pressure and, therefore, the height above sea level, weather conditions, oxygen concentration in the air, its humidity and the amount of moisture absorbed, characteristics of the dust in the air, impurities in the chemical compound, injection rate and uniformity of sample dispersion, etc (1-3). Moreover, since the flame inside the flask is detected by visual inspection, the measured AIT values depend also on human error. As a result, a compound's AIT value may experiment-dependently vary by up to several hundred of degrees centigrade (4, 5). According to some estimates, the average measurement error of AIT is around  $\pm 30^{\circ}\text{C}$  (6, 7).

Early attempts to establish relationships between chemical structure of hydrocarbons and AIT were made in the 1950s -1960s. It has been found that high branching inhibits spontaneous ignition and hence raises AIT. AIT was related to the number of methylene groups and the relative location of branching points in the molecules (2, 8, 9). Other important structural features impacting the AIT of hydrocarbons are the presence of cycles, steric strain, aromaticity and unsaturation in molecules (2). All these factors can be attributed to the easiness of free radical formation, which indicates the free radical mechanism of spontaneous ignition (2, 8, 10). It was also shown that the transition from hydrocarbons to aldehydes dramatically reduces AIT, whereas the transition to ketones leads to the increase of AIT (6).

The first QSPR (Quantitative Structure-Property Relationships) models for AIT were built on small series of hydrocarbons, alcohols, and esters in the pioneering work of Egolf and Jurs (3). Application of the Multiple Linear Regression (MLR) method in conjunction with topological, geometrical and electronic descriptors led to small standard errors 12-16 $^{\circ}\text{C}$  for training sets. Two different mechanisms, the low-temperature and the high-temperature ones, leading to different structure-property relationships were suggested. This methodology was further extended in the study by Mitchell and Jurs (11) who considered additional subsets of nitrogen and sulfur containing compounds and applied both the MLR and the Artificial Neural Networks (ANN) methods. Predictive performance of ANN models on external test sets was found similar to typical experimental errors of AIT measurements. Nonetheless an attempt of modeling on relatively large and diverse data set on 300 organic compounds composed of the afore-mentioned subsets led to a poor MLR with the root mean square error (RMSE) of 58.5 $^{\circ}\text{C}$  assessed on an external test set of

27 compounds. Moreover, an MLR model built on a subset of 223 heteroatom-containing compounds led to RMSE = 61.8°C, which was much worse compared to the models built on small congeneric sets of compounds.

Suzuki built a MLR model for a diverse set of 100 monofunctional organic compounds, including hydrocarbons, alcohols, phenols, ethers, aldehydes, ketones, acids, amines, esters, and halogenated compounds, using a set of 17 calculated molecular descriptors and 4 known intrinsic molecular properties (flash point, boiling point, critical temperature and pressure), which was validated on an external set of 150 testing compounds with MAE = 33.9°C (6). Using the same dataset, Tetteh et al. built an ANN model with only 6 descriptors (7, 12) which provided with the good statistics for the external test set (MAE = 30°C). Kim et al. (13) used the Genetic Functional Approximation – Multiple Linear Regression (GFA-MLR) method to build a model on a dataset of 200 organic compounds encoded by 72 molecular descriptors, which performed on a test set of 43 compounds with a MAE = 29°C. In 2003, Albahri (14) applied the Structural Group Contribution method to prepare a predictive model on a set of some 500 hydrocarbons, with MAE = 28°C on the training set. Applying the ANN to the same dataset, Albahri and George (15) reduced MAE for the training set to 17.8°C and for a small external test set of 20 compounds obtained MAE = 16.7°C. In 2009, Chen et al. applied the GCM method based on polynomial regression to build a model using a training set containing 400 organic compounds with heteroatoms (5). The average prediction error estimated for this model using a test set containing 83 compounds was reported to be 70°C.

In 2008, Pan et al. used ANN in combination with atom-type electrotopological-state indices to build a model for 118 hydrocarbons, validated on a test set containing 42 compounds with RMSE = 31.1°C (16). The same authors also applied the SVM (Support Vector Machines) method in combination with the descriptors previously used by Tetteh et al. (7, 12) to build a QSPR model on a training set containing 52 diverse organic compounds; its validation on a test set containing 90 compounds led to RMSE = 31°C, which is considerably better than the results obtained for the same data using ANN and MLR (17). This methodology was further extended to larger datasets (356/90 compounds for training / test sets, respectively), using various types of molecular descriptors and Genetic Algorithm (GA) – Partial Least Squares (PLS) methods for descriptor selection (18). Statistical parameters obtained for a test set (MAE= 28.9°C and RMSE= 36.9°C) lie within the range of experimental errors.

In 2011, Lazzus built a model for AIT using a combination of SGC with ANN trained using the PSO (Particle Swarm Optimization) algorithm on a dataset of 250 compounds, and the average absolute error estimated on a test set containing 93 compounds was reported to be 10.5°C (with  $R^2=0.99$ ), while the same measure for ANN trained using the standard backpropagation algorithm appeared to be 45.3°C (19). Such a huge difference in the quality estimates of the models built using two neural network training algorithms, as well as a surprisingly low prediction error (three times lower than the average measurement error), suggests the presence of strong overfitting (overtraining for ANNs). The possibility of the emergence of overfitting in this case follows from the fact that the number of fitting parameters of the neural network (42 descriptors, one hidden layer consisting of 4 neurons) is 177, which is only slightly less than the number of training examples (250 in this case). On the other hand, it is known that, to prevent overfitting, the number of adaptable parameters should be, when using optimization methods that can find a global minimum and in the absence of regularization, several times smaller. For example, the classical “rule of 5” (called also “Topliss law”) widely accepted in chemoinformatics states that for building

good QSPR models there should be at least 5 data points per adjustable variable (20, 21). Since this condition can only be satisfied when using very small neural networks, global-minimum searching algorithms, such as PSO and GA, are almost never used to train ANNs instead of backpropagation, because they always cause very strong overfitting, although they are sometimes used to optimize the architecture of neural networks and to form optimal sets of descriptors. Unfortunately, the global minimum of the error function in not very small ANNs always corresponds to the retrained model. As for the backpropagation algorithm (which has always been the standard method of training ANNs), due to the use of gradual gradient optimization of the error function in combination with regularization methods such as “early stop”, the dynamically growing “effective number of adaptable parameters” always turns out to be rather small, which prevents retraining even when using large ANNs. As for the very low reported prediction error on the test set, it should be taken into account that it was formed manually in a special way leading to low prediction errors on it. In particular, it was ensured that all fragments of the structures of the prediction set were well represented in the training set. Thus, all the compounds that are problematic for predicting fall into the training, and not into the prediction set. Although this approach allows one to get very low errors on a fixed test set, however, it gives a highly optimistically biased estimate of the predictive ability of the model.

Gharagheizi also published in 2011 a study on application of the SGC with ANN trained using the standard Levenberg-Marquardt algorithm based on backpropagation using a training set consisting of 821 compounds to obtain a QSPR model well performing on the test set of 102 compounds (RMSE = 15° and R<sup>2</sup>=0.984) (22), which is also much lower than the average measurement error. In 2012, Bagheri et al. published MLR and ANN models based on only 3 molecular descriptors selected using the PSO algorithm for predicting AIT of organic sulfur chemicals with the root mean square prediction errors being 17.6°C and 14.7°C, respectively (23), and it is not clear from the article whether these values concern the training or the external prediction. Keshavarz et al. published in 2013 a very simple linear model involving only two constitutional parameters and two specially designed factors capable of predicting the AIT of hydrocarbons with the average accuracy (root mean square error estimated on a test set consisting of 26 compounds) of 27°C (24). Borhani et al. built in 2016 MLR and ANN models based on 3 descriptors selected using the GA (Genetic Algorithm) on a training set with 650 organic compounds, and the average absolute prediction error estimated using a test set with 163 compounds were reported to be 36.4°C and 37.6°C, respectively (25). However, we could not reproduce the aforementioned linear model built on 3 descriptors using the data published in this article. Finally, Dashti et al. in recent study built QSPR models using the GA-ANFIS (Adaptive Neuro Fuzzy Inference System), PSO-ANFIS, DE-ANFIS and GP (Genetic Programming) methods on a training set containing 356 compounds and using a test set with 90 compounds to assess the predictive performance which was shown to be within the average experimental error (4).

Thus, there exists an extensive literature reporting QSPR models for AIT. However, predictive performance of these models was estimated in most cases using a fixed test set of a small size. Since statistical parameters characterizing predictive ability of models largely depends on the test set composition, the real predictive performance of the reported in the literature models is still questionable. Particularly serious doubts concern the models for which published estimates of prediction errors are considerably lower than experimental measurement errors (19). These shortcomings unfortunately concern many of models published in the last decade.

In this article, we analyze the consistency of available data on AIT and reproducibility of different QSPR models for it. Then we rebuild QSPR models for AIT using as large and diverse as possible database formed by merging the data taken from different sources with a special focus assessment of the reliability of predictions.

## 2. *Methods and Materials*

### 2.1. *ISIDA Fragmental Descriptors*

Fragment descriptors are counts of fragments (subgraphs) of various types and sizes occurring in every molecule (26, 27). ISIDA Fragment Descriptors (28, 29) can be generated for several user-defined fragment types: sequences of atoms and bonds, augmented atoms (circular fragments), Carhart atom pairs, atom triplets, etc. A compound-fragment count matrix results by associating each fragment to a column, each molecule to a row, with each cell containing the (integer) number of times the column-associated fragment appears in the row-associated molecule. After adding the property value as explained variable (“Y column”), machine learning algorithms use such matrices to propose mathematical models approximating the property as a function of (some or all) descriptors. Structure-property relationships of any complexity can be approximated using different kinds of fragment descriptors (30-33). In this paper, they are used in conjunction with Support Vector Regression and Generative Topographic Mapping approaches.

### 2.2. *Support Vector Regression*

Support Vector Regression (SVR), is a modern robust machine learning regression method based on the  $\epsilon$ -tube concept: it tries to provide predictions with absolute errors under a certain threshold value  $\epsilon$  (34). Non-linearity is introduced into the method by means of special functions – *kernels*. Construction of SVM models always requires optimization of *hyperparameters* – numbers defining model characteristics, like the  $\epsilon$ -value, kernel type, cost parameter, etc.

In this work, all SVR models were built using the LIBSVM software (35), while the hyperparameters were optimized using the genetic algorithm implemented in the GAconfig program (36), which simultaneously selects the best suited particular types of ISIDA fragment descriptors by considering the “descriptor type” as an additional hyperparameter to be tuned.

As such evolutionary simulations result in a population of near-optimal SVR models, their *consensus* (averaging of individual model predictions  $\pm$  standard deviation) was taken as the final predicted value of AIT. This has two advantages. First, prediction error decreases due to mutual elimination of prediction noise, and, second, the standard of predictions is an estimator of the trustworthiness of predictions. In this study, different models result from different splits of datasets.

### 2.3. *Evaluation of Model Predictive Performance*

Validation of models is performed by splitting initial dataset into *training* and *test* sets, so the models are constructed with data from the training sets, while the test sets are used for making predictions and evaluating the predictive performance. Different splits can be produced in  $n$ -fold cross-validation loops, in which each compound is moved to a test set exactly one time. Cross-validated determination coefficient,  $R^2$ , root mean square error, RMSE, mean absolute error, MAE, and average absolute relative error, AARE, characterize the predictive performance quantitatively:

$$R^2 = 1 - \frac{\sum_1^n (y_{pred,i} - y_{exp,i})^2}{\sum_1^n (y_{pred,i} - \bar{y}_{exp,i})^2}, \text{ where } \bar{y}_{exp,i} = \frac{\sum_1^n y_{exp,i}}{n}$$

$$RMSE = \sqrt{\frac{\sum_1^n (y_{pred,i} - y_{exp,i})^2}{n}}$$

$$MAE = \frac{\sum_1^n |y_{pred,i} - y_{exp,i}|}{n}$$

$$AARE = \frac{1}{n} \sum_1^n \frac{|y_{pred,i} - y_{exp,i}|}{y_{exp,i}}$$

where  $n$  is the number of compounds,  $y_{pred,i}$  value of property  $y$  for  $i$ -th compound measured in experiment,  $y_{exp,i}$  – the value of property  $y$  for  $i$ -th compound predicted by the model. Notice that only AARE depends explicitly on the temperature scale (Celsius or Kelvin) used for AIT measuring.

Cross-validation loops were repeated 12 times after random shuffling of the order of chemical compounds in the dataset in order to eliminate the dependence of statistical characteristics on the order of compounds in data sets. The above indices can then be taken as an average over the reshuffling attempts.

#### 2.4. Generative Topographic Mapping

Generative Topographic Mapping (GTM) is a dimensionality reduction method, a probabilistic counterpart of self-organizing maps (37). In chemoinformatics, GTM can be used for mapping chemical space and building classification and regression predictive models (38-41). In this case, any descriptor (pattern) matrix computed for a set of chemical compounds defines a data cloud, with each row (chemical compound) corresponding to a data point in it. GTM inserts a *manifold* (looks like a flexible sheet of paper hovering in the data cloud), projects the data points onto the manifold, which after unfolding produces a map on a 2D latent space. So, each chemical compound is projected from the descriptor space to a single point on the map. The GTM latent space serves as support for map visualization: it can be grey scaled to indicate local data density, displaying *density landscapes*. It can be “colored” according to the mean property values for chemical compounds mapped to it to produce *property landscapes* in which density is used to modulate color intensity. Property landscapes are predictive – an external compound projected on such landscape can have its predicted property estimated by the local map “color” (mean landscape property). Therefore, predictive GTM landscapes may also be quantitatively cross-validated, as stated in the previous paragraph.

### 3. Results and Discussion

#### 3.1. Models from Individual Literature Sources

At the first stage, we have rebuilt QSPR models using the data taken from the articles published no later than 10 years ago. For all of them, strong statistical parameters for QSPR models were reported. In addition, we used AIT data extracted from the DETHERM, database (<https://dechema.de/en/detherm.html>). The data sources used here are listed in *Table 1*. For all 7 data sources, SVR models were built on the basis of ISIDA fragment descriptors using the GAconfig software, as discussed above. The statistical characteristics of the models are given in *Table 2*. The corresponding scatter plots are given in Supplementary Information.

Table 1. Data sources used in modeling AIT

Data source	Classes	# data points	test set size	Data source
Bagheri2012	organic sulphur compounds	45		reference (23)
Keshavarz2013	hydro-carbons	270		reference (24)
Pan2009	Diverse	446		reference (18)
Chen2009	Diverse	476		reference (5)
Lazzus2011	Diverse	93 <sup>c</sup>		reference (19)
DETERM	Diverse	522		DETERM database
Borhani2016	Diverse	806		reference (25)

<sup>c</sup>Only test data are provided in the publication

Table 2. Statistical characteristics of AIT prediction models built using data from individual sources

Data source	Statistics from original publication (assessed for test set) <sup>a</sup>				Statistics for rebuilt models (assessed in cross-validation) <sup>b</sup>			
	R <sup>2</sup>	RMSE	MAE	AARE	R <sup>2</sup>	RMSE	MAE for	AARE (in %)
Bagheri2012					0.856	23.0	17.2	3.2
Keshavarz2013	0.930	27			0.841	41.7	29.3	4.6
Pan2009	0.878	36.9	28.9		0.809	46.0	34.6	5.3
Chen2009	0.536	70		11.0	0.738	56.9	39.7	6.2
Lazzus2011	0.9899		10.5	1.6	0.649 <sup>c</sup>	75.1 <sup>c</sup>	54.8 <sup>c</sup>	9.1 <sup>c</sup>
DETERM					0.517	84.1	57.5	9.2
Borhani2016	0.7567 <sup>d</sup>	45.4 <sup>d</sup>	37.6 <sup>d</sup>	6.8 <sup>d</sup>	0.080	95.1	78.1	12.7

<sup>a</sup> Obtained on the external test set. <sup>b</sup> Obtained in cross-validation. <sup>c</sup> The model was built using only the test data, because the training data are not provided in the original publication. <sup>d</sup> We couldn't reproduce the published model

One can conclude from the results presented in Table 2 that the data taken from the sources Bagheri2012 (18), Keshavarz2013, Pan2009, Chen2009 and Lazzus2011 are rather well suited for QSPR modeling. On the other hand, we failed to build predictive models with the DETERM and Borhani2016 data sources. The latter result is in sharp contradiction with that reported in ref. (25) for the model built on the same data. However, we failed to reproduce the linear model built on 3 descriptors published in ref. (25) using the same dataset, descriptors and training/test splitting of data sets (we obtained R<sup>2</sup>=0.16 and RMSE=91.1 degrees for the training set and R<sup>2</sup>=-0.162, RMSE=105.5 degrees for the test set).

### 3.2. Models from Merged Data

First, all extracted datasets were merged into one big dataset. There is a large overlap between the individual datasets, so most compound are contained in several datasets. The standard deviation of the AIT values for compounds appearing in several datasets is 42.3 degrees. However, the

quality of the model built using the above-discussed methodology on this set appeared to be very bad, presumably because of low data quality. It can be supposed that the reason for bad data modeliability is bad data quality because the well and poorly modeled data sets in this study contain highly overlapping sets of chemical compounds described by the same descriptors. Therefore, we decided to prepare a smaller size dataset aggregating only good quality data points. The merging process began with the dataset supporting the model with the best predictive ability in terms of RMSE, and at each step a dataset was added that provides the best predictive ability of the model based on the merged data. The median value of the auto-ignition temperature were taken in the cases when several values are reported for the same compound. For each of the merged databases (m1,...,m6) QSPR models were built using the GAconfig software, like in the previously described case of individual data source. The statistical characteristics of the resulting models are presented in Table 3, while the corresponding scatter plots are given in Supplementary Material. We have chosen the **m4** merged dataset containing 875 chemicals as an optimal one for AIT modeling. In this case, the RMSE values for the test sets in cross-validation is 54.4°C, the MAE value is 37.8°C, AARE=6%, while the corresponding values of R<sup>2</sup> is 0.77. The scatter plot for test set prediction in cross-validation is given in Figure 1, while the most significant outliers are presented in Table 4.

The **m4** dataset is visualized on the generative topographic map shown on Figure 4. One can see that particular chemotypes populate distinct areas on the map: hydrocarbons, aromatic compounds, ethers, esters, carboxylic acids, organohalogen compounds, amines, heterocyclic compounds, etc. Rows of compounds formed by representatives of individual classes are clearly visible on the map.

Table 3. Statistical characteristics <sup>a</sup> of QSPR models built on different merged data

Merge	Composition	# data points	R <sup>2</sup>	RMSE	MAE	AARE %
m1	Bagheri2012 + Keshawarz2013	314	0.827	43.1	28.6	4.7
m2	m1 + Pan2009	679	0.831	44.6	31.8	5.0
m3	m2 + Chen2009	851	0.777	53.6	36.6	5.8
m4	m3 + Lazzus2011	875	0.770	54.4	37.8	6.0
m5	m4 + DETHERM	973	0.668	67.6	43.6	7.0
m6	m5 + Borhani2016	1235	0.465	81.8	58.0	9.4

<sup>a</sup> obtained in cross-validation

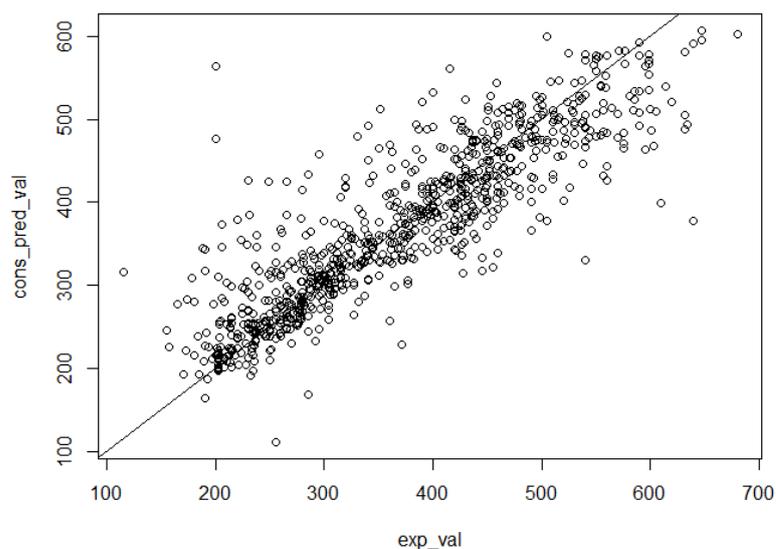


Figure 1. Scatter plot for the selected model with merged data

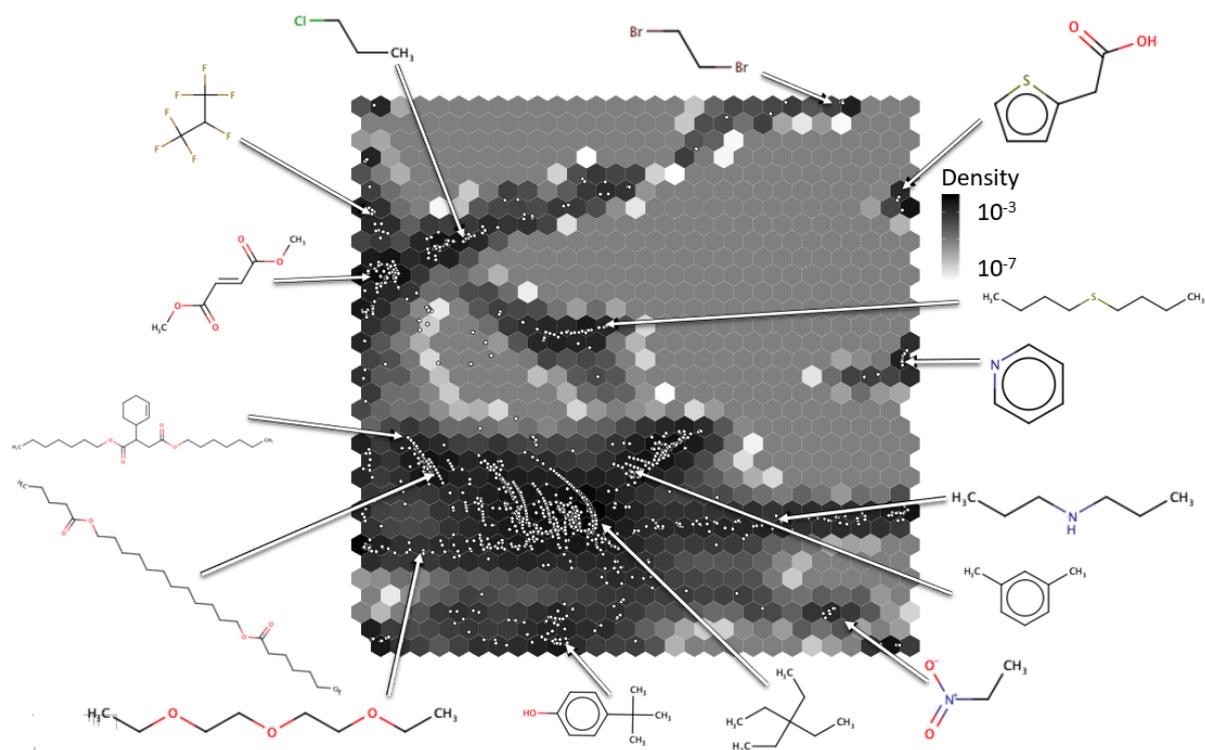
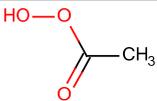
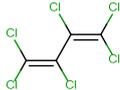
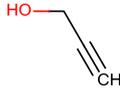


Figure 2. Generative topographic map for *m4* dataset. The arrows show some zones populated by compounds of particular chemotype

Table 4. Top outliers detected for *m4* model in cross-validation

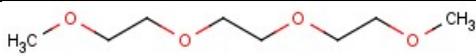
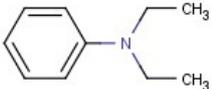
ID	Structure	Experimental value	Predicted Value	Predicted - experimental

780		200	564.1	364.1
289		200	476.5	276.5
210		640	377.8	-262.2
767		610	399.4	-210.6
220		540	330.8	-209.2
187		115	316.2	201.2

### 3.3. Assessment of Data Consistency

Internal data compatibility was checked by comparing the AIT values reported in different data sources for the same compound. Thorough analysis of the data has revealed big problems with data compatibility. For 4 compounds, the difference between the maximum and minimum reported AIT values exceed 400 degrees (see Table 5), for 11 compounds it is between 300 and 400 degrees (see Table 6), for 55 compounds – between 200 and 300 degrees, and for 155 compounds – between 100 and 200 degrees. It should be noted that it is often not clear what criteria should be used to select the correct values, and how to automate this process. In some cases, such big errors can be explained by erroneous transformation between Celsius and Fahrenheit temperature scales. For example, the values of 332°C and 630°C are reported for N,N-diethylaniline in Table 6 (CAS reg. num. 91-66-7), while 332°C exactly corresponds to 630°F. Sometimes, Kelvin temperature scale also enters the game. For example, 283°C, 550°C and 560°C are reported for 1,3,5-trimethylbenzene (CAS reg. num. 108-67-8) in different sources, while 283°C is equal to 556.15 K.

**Table 5.** Compounds for which experimentally measured AsIT values differ more than 400 degrees

CAS RN	Structure	Range	AITs
112-49-2		439.0	191.0, 367.0, 630.0
91-66-7		425.0	205.0, 332.0, 332.0, 630.0

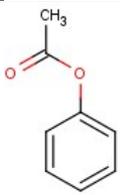
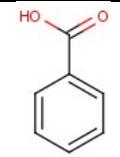
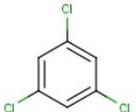
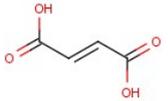
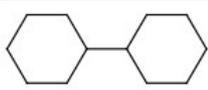
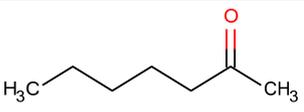
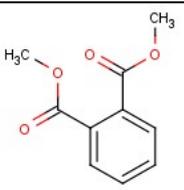
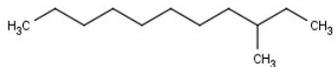
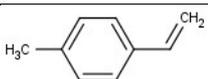
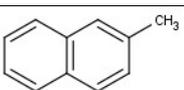
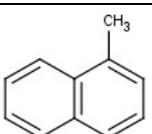
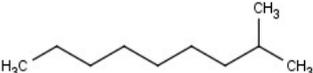
122-79-2		414.0	171.0, 585.0
65-85-0		402.0	172.0, 532.0, 567.0, 574.0

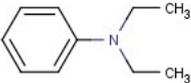
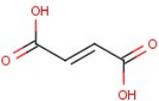
Table 6. Compounds with the range of reported AIT values between 300 and 400 degrees

CAS RN	Structure	Range	AITs
108-79-3		370.0	207.0, 577.0
110-17-8		365.0	375.0, 740.0
75-07-0		355.0	130.0, 130.0, 175.0, 175.0, 175.0, 176.0, 485.0
92-51-3		341.0	244.0, 244.0, 245.0, 585.0
110-43-0		331.0	202.0, 393.0, 393.0, 393.0, 393.0, 533.0, 533.0, 533.0, 533.0
131-11-3		331.0	225.0, 460.0, 490.0, 555.0, 556.0, 556.0
1002-43-3		328.0	207.0, 535.0
622-97-9		318.0	257.0, 538.0, 575.0, 575.0
91-57-6		317	212.0, 488.0, 529.0, 529.0
90-12-0		316.0	214.0, 529.0, 529.0, 529.0, 529.0, 530.0

871-83-0		315.0	214.0, 214.0, 214.0, 529.0
----------	---	-------	----------------------------

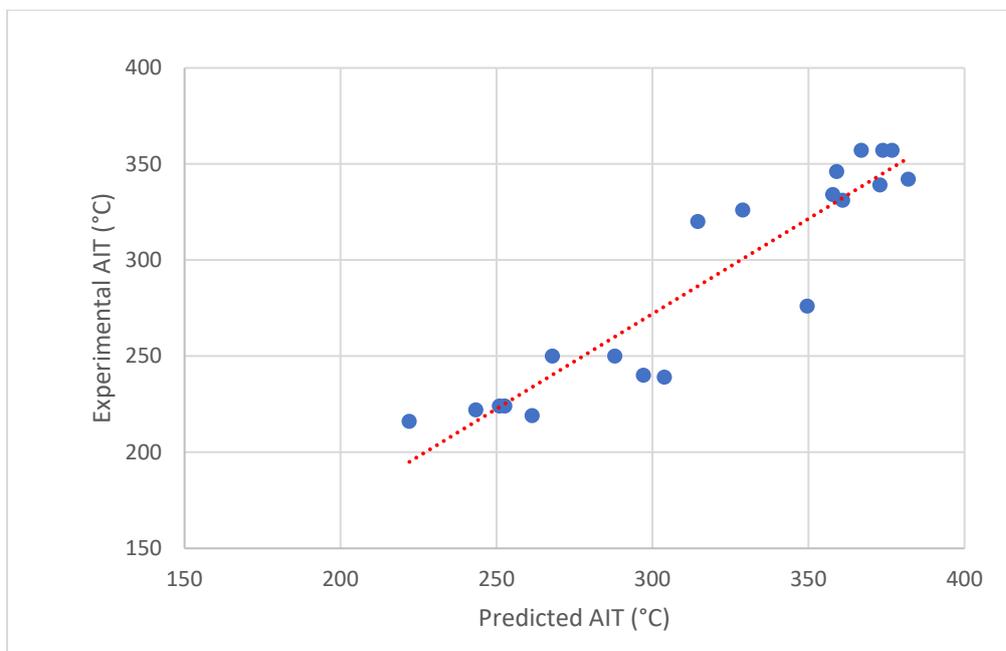
We have also discovered that the confusion between different AIT values can also be observed in Safety Data Sheets published by commercial suppliers of chemicals. Four cases in which AIT reported for the same chemical compound in the data sheets provided by different suppliers of chemicals are given in Table 7. So, one can see that for N,N-diethylaniline the difference in AIT reported in the Safety Data Sheets provided by different commercial suppliers of chemicals is 245 °C, for N-butylchloride – 228 °C, for 1,4-dioxane – 176 °C, while for fumaric acid – 341 °C. A sad consequence of such confusion is that easily flammable chemical compounds can be registered as being completely safe, which, in turn can lead to serious accidents in chemical laboratories or plants.

Table 7. AIT values extracted from different data safety sheets

Structure	Product name	Supplier	AIT, °C	Difference
	N,N-Diethylaniline	BorsodChem MCHZ, s.r.o.	385	245
		Acros Organics N.V.	630	
	N-Butylchloride	Central Drug House (P) Ltd	245	228
		Scientific & Chemical Supplies Limited	473	
	1,4-Dioxane	Fisher Scientific	180	176
		Thermo Fisher Scientific	356	
	Fumaric acid	Central Drug House (P) Ltd	399	341
		InterAtlas Chemical Inc.	740	

### 3.4 External Predictions

The AIT model m4 has been applied to an external data set of 39 molecules (which are not present in the training set) for which the experimental values were measured by TOTAL teams. 19 chemicals were predicted out of applicability domain. **Erreur ! Source du renvoi introuvable.** shows the scatter plot representing experimental AIT versus predicted ones for the 20 other chemicals. The correlation is quite good with statistics slightly better than the ones presented in Table 3 ( $r^2=0.87$ , RMSE=34°C, AAE=28°C), confirming the predictive ability of the m4 model.



**Figure 3.** Correlation between experimental and predicted AITs for the external dataset

#### 4. Conclusions

In this paper, we have performed detailed analysis of available AIT experimental data for organic compounds and critically reviewed reported in literature QSPR models for AIT prediction. Internal data compatibility was checked by comparing the AIT values reported in different data sources for the same compound, and several serious problems were detected. In particular, from the total number of 1235 compounds analyzed, the difference between the maximum and minimum AIT values is over 400 degrees for 4 compounds, between 300 and 400 degrees - for 11 compounds, between 200 and 300 degrees - for 55 compounds, and between 100 and 200 degrees - for 155 compounds. A possible reason for this might be both a strong dependence of the results measured AIT values on uncontrolled variations in external conditions, and human errors when working with data. We have also demonstrated in this paper that sharply different AIT values for the same compound can be found not only in journal papers and database records but even in the Safety Data Sheets published by commercial suppliers of chemicals. The potential danger of such incompatibility is that spontaneously ignitable chemical compounds can be registered to be completely safe, and this can lead to serious accidents in chemical laboratories and plants.

Although there is an extensive literature concerning building QSPR models for AIT, their predictive performance was either not estimated at all or estimated using a small manually prepared test set. As a result, all assessments of the models performance were biased towards overestimation of the quality of prediction. This explains why the prediction errors on the test set are reported in several publications to be lower on average than the experimental measurement errors and the errors on the corresponding training set.

In this study, a SVM regression model was built on a dataset containing 875 compounds formed by merging the sets of compounds published in the papers dealing with building QSPR models for AIT. For this model, the RMSE of prediction assessed on the test sets formed in repeated cross-validation is 54.4°C, the MAE value is 37.8°C, while the corresponding value of  $R^2$  is 0.77. The

MAE value obtained in this work is not far from the average experimental error estimated as 30°C. An additional verification of the predictive ability of this model on external data has confirmed its good predictive ability.

**Supplementary Information** contains an information about optimized SVR parameters and type of ISIDA fragmental descriptors used in the *m4* model and the scatter plots AIT (pred) vs AIT (exp) for different data sets mentioned in Table 2.

## References

1. Frank CE, Blackham AU. Reaction Processes Leading to the Spontaneous Ignition of Hydrocarbons. *Industrial & Engineering Chemistry*. 1954;46(1):212-7.
2. Swarts DE, Orchin M. Spontaneous Ignition Temperature Of Hydrocarbons. *Industrial & Engineering Chemistry*. 1957;49(3):432-6.
3. Egolf LM, Jurs PC. Estimation of autoignition temperatures of hydrocarbons, alcohols, and esters from molecular structure. *Industrial & Engineering Chemistry Research*. 1992;31(7):1798-807.
4. Dashti A, Jokar M, Amirkhani F, Mohammadi AH. Quantitative structure property relationship schemes for estimation of autoignition temperatures of organic compounds. *Journal of Molecular Liquids*. 2020;300:111797.
5. Chen C-C, Liaw H-J, Kuo Y-Y. Prediction of autoignition temperatures of organic compounds by the structural group contribution approach. *Journal of Hazardous Materials*. 2009;162(2):746-62.
6. Suzuki T. Quantitative structure—property relationships for auto-ignition temperatures of organic compounds. *Fire and Materials*. 1994;18:81-8.
7. Tetteh J, Metcalfe E, Howells SL. Optimisation of radial basis and backpropagation neural networks for modelling auto-ignition temperature by quantitative-structure property relationships. *Chemom Intell Lab Syst*. 1996;32(2):177-91.
8. Frank CE, Blackham AU. Spontaneous Ignition of Organic Compounds. *Industrial & Engineering Chemistry*. 1952;44(4):862-7.
9. Affens WA, Johnson JE, Carhart HW. Effect of Chemical Structure on Spontaneous Ignition of Hydrocarbons. *Journal of Chemical & Engineering Data*. 1961;6(4):613-9.
10. Morley C. A Fundamentally Based Correlation Between Alkane Structure and Octane Number. *Combustion Science and Technology*. 1987;55(4-6):115-23.
11. Mitchell BE, Jurs PC. Prediction of Autoignition Temperatures of Organic Compounds from Molecular Structure. *J Chem Inf Comput Sci*. 1997;37(3):538-47.
12. Tetteh J, Howells S, Metcalfe E, Suzuki T. Optimisation of radial basis function neural networks using biharmonic spline interpolation. *Chemom Intell Lab Syst*. 1998;41(1):17-29.
13. Kim YS, Lee SK, Kim JH, Kim JS, Tai No K. Prediction of autoignition temperatures (AITs) for hydrocarbons and compounds containing heteroatoms by the quantitative structure—property relationship. *Journal of the Chemical Society, Perkin Transactions 2*. 2002(12):2087-92.
14. Albahri T. Flammability characteristics of pure hydrocarbons. *Chemical Engineering Science*. 2003;58:3629-41.
15. Albahri TA, George RS. Artificial Neural Network Investigation of the Structural Group Contribution Method for Predicting Pure Components Auto Ignition Temperature. *Industrial & Engineering Chemistry Research*. 2003;42(22):5708-14.
16. Pan Y, Jiang J, Wang R, Cao H, Zhao J. Prediction of auto-ignition temperatures of hydrocarbons by neural network based on atom-type electrotopological-state indices. *Journal of Hazardous Materials*. 2008;157(2):510-7.
17. Pan Y, Jiang J, Wang R, Cao H. Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds. *Chemom Intell Lab Syst*. 2008;92(2):169-78.
18. Pan Y, Jiang J, Wang R, Cao H, Cui Y. Predicting the auto-ignition temperatures of organic compounds from molecular structure using support vector machine. *Journal of Hazardous Materials*. 2009;164(2):1242-9.
19. Lazzús JA. Autoignition Temperature Prediction Using an Artificial Neural Network with Particle Swarm Optimization. *International Journal of Thermophysics*. 2011;32(5):957.
20. Hansch C, Kim KH, Sarma RH. Structure-activity relation in benzamides inhibiting alcohol dehydrogenase. *J Am Chem Soc*. 1973;95(19):6447-9.
21. Topliss JG, Costello RJ. Chance correlations in structure-activity studies using multiple regression analysis. *J Med Chem*. 1972;15(10):1066-8.

22. Gharagheizi F. An accurate model for prediction of autoignition temperature of pure compounds. *Journal of Hazardous Materials*. 2011;189(1):211-21.
23. Bagheri M, Borhani TNG, Zahedi G. Estimation of flash point and autoignition temperature of organic sulfur chemicals. *Energy Conversion and Management*. 2012;58:185-96.
24. Keshavarz MH, Gharagheizi F, Ghanbarzadeh M. A simple correlation for prediction of autoignition temperature of various classes of hydrocarbons. *Journal of the Iranian Chemical Society*. 2013;10(3):545-57.
25. Borhani TNG, Afzali A, Bagheri M. QSPR estimation of the auto-ignition temperature for pure hydrocarbons. *Process Safety and Environmental Protection*. 2016;103:115-25.
26. Baskin I, Varnek A. Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. In: Varnek A, Tropsha A, editors. *Chemoinformatics Approaches to Virtual Screening* Cambridge: RSC Publisher; 2008. p. 1-43.
27. Baskin I, Varnek A. Building a chemical space based on fragment descriptors. *Comb Chem High T Scr*. 2008;11(8):661-8.
28. Varnek A, Fourches D, Hoonakker F, Solov'ev V. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J Comput-Aided Mol Des*. 2005;19(9):693-703.
29. Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, et al. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr Comput-Aided Drug Des*. 2008;4(3):191-8.
30. Baskin II, Skvortsova MI, Stankevich IV, Zefirov NS. On the Basis of Invariants of Labeled Molecular Graphs. *J Chem Inf Comput Sci*. 1995;35(3):527-31.
31. Artemenko NV, Baskin II, Palyulin VA, Zefirov NS. Prediction of physical properties of organic compounds using artificial neural networks within the substructure approach. *Dokl Chem*. 2001;381(1):317-20.
32. Artemenko NV, Baskin II, Palyulin VA, Zefirov NS. Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. *Russ Chem Bull*. 2003;52(1):20-9.
33. Zhokhova NI, Baskin II, Palyulin VA, Zefirov AN, Zefirov NS. Fragmental descriptors with labeled atoms and their application in QSAR/QSPR studies. *Dokl Chem*. 2007;417(2):282-4.
34. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: Mozer MC, Jordan JI, Petsche JI, editors. *Advances in Neural Information Processing Systems*: MIT Press; 1997. p. 155-61.
35. Chang C-C, Lin C-J. LIBSVM : a library for support vector machines. *ACM Trans Intel Syst Technol*. 2001;2(3):27:1-:.
36. Horvath D, Brown JB, Marcou G, Varnek A. An Evolutionary Optimizer of libsvm Models. *Challenges*. 2014;5(2):450-72.
37. Bishop CM, Svensén M, Williams CKI. GTM: The Generative Topographic Mapping. *Neural Comput*. 1998;10(1):215-34.
38. Kireeva N, Baskin II, Gaspar HA, Horvath D, Marcou G, Varnek A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol Inf*. 2012;31(3-4):301-12.
39. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A. GTM-Based QSAR Models and Their Applicability Domains. *Mol Inf*. 2015;34(6-7):348-56.
40. Gaspar HA, Marcou G, Horvath D, Arault A, Lozano S, Vayer P, et al. Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J Chem Inf Mod*. 2013;53(12):3318-25.
41. Gaspar HA, Sidorov P, Horvath D, Baskin II, Marcou G, Varnek A. Generative Topographic Mapping Approach to Chemical Space Analysis. *Frontiers in Molecular Design and Chemical Information Science - Herman Skolnik Award Symposium 2015: Jürgen Bajorath*. ACS Symposium Series. 1222: American Chemical Society; 2016. p. 211-41.