CHEMICAL ENGINEERING TRANSACTIONS
VOL. 116, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Bruno Fabiano, Valerio Cozzani

ISBN 979-12-81206-18-2; **ISSN** 2283-9216
DOI: 10.3303/CET25116007

Can AI Revolutionize QSPR Models for the Chemical Mixtures Hazards?

Guillaume Fayet*, Nour Helou, Patricia Rotureau

Ineris, Parc Technologique Alata, BP2, 60550 Verneuil-en-Halatte
guillaume.fayet@ineris.fr

The physical hazards of chemical mixtures are typically characterized using experimental tools that could benefit to be prioritized by using predictive methods. Indeed, experimental tests can be costly, complex, time-consuming, and potentially dangerous for the operator. In the last decades, particularly with the implementation of the REACH regulation, predictive methods such as QSAR/QSPR (Quantitative Structure-Activity/Property Relationships) have been encouraged and utilized as rapid and economical alternatives to experimental testing for determining (eco)toxicological and physical hazards of chemical substances.
Initially designed for pure compounds, adaptations of the QSPR approach were proposed to predict the properties of mixtures even if their development, in particular for physical hazards, is still an emerging field. Indeed, existing QSPR models still present some limitations to complement mixing rules and experimental approaches, and there is a need for new and more reliable models to extend applicability and improve prediction accuracy. A possible orientation could be using advanced machine learning approaches, taking advantage of scientific progress in artificial intelligence beyond classical multilinear regressions. More complex non-linear approaches (such as neural networks or random forests) have recently been used with the hope of better accounting for mixture complexity in QSPR models for mixtures.
This research aims to investigate if integrating advanced AI analytical methods can enhance the performance and applicability of QSPR models for predicting the physical hazards of chemical mixtures. To this end, applications of different machine learning methods were tested to evidence the advantages and limits of these Advanced AI algorithms compared to the more classical MLR approach when developing models for the flammability of liquid mixtures.

1. Introduction

Dealing with the physical hazards of chemical mixtures is inherently complex not only due to the need for multiple tests to account for concentration effects but also regarding the effects of interactions between substances within the mixtures, possibly complex and non-linear. Indeed, the complexity of chemical mixtures arises from the myriad interactions between different components, which can significantly alter their physical properties. So, the experimental campaigns needed to investigate the hazard profiles of mixtures completely can be costly and time-consuming. Among existing predictive methods, Quantitative Structure-Property Relationships (QSPR) revealed as efficient and economical alternatives to traditional experimental testing for assessing the (eco)toxicological and physical hazards of chemical substances even in regulatory frameworks like the European REACH regulation (Nieto-Draghi et al., 2015). Nevertheless, these models were initially developed for pure compounds and their adaptation to predict the properties of mixtures remains an emerging field (Fayet and Rotureau, 2023). Current models face limitations in accurately predicting physical hazards, necessitating the development of more reliable and applicable models. An area of progress for these models is to better account for the complexity of these systems in the QSPR models.
In the last decade, Artificial Intelligence (AI) methods have gained significant interest and popularity due to their ability to solve complex problems and make accurate predictions. AI techniques are notably transforming various domains within chemistry, including organic synthesis, drug discovery, and analytical techniques. For instance, machine learning models can predict reaction outcomes, optimize synthetic routes, and identify potential drug

Paper Received: 13 December 2024; Revised: 1 February 2025; Accepted: 17 April 2025
Please cite this article as: Fayet G., Helou N., Rotureau P., 2025, Can Ai Revolutionize Qspr Models for the Chemical Mixtures Hazards?, Chemical Engineering Transactions, 116, 37-42 DOI:10.3303/CET25116007

candidates with high precision (Joshi, 2023). These advancements are not only accelerating research but also improving the efficiency and accuracy of chemical experiments (Aal E Ali et al., 2024; Lin and Mo, 2024).
In this context, this study challenges the integration of advanced AI methods in QSPR models, particularly for predicting the flammability of liquid mixtures. Considering that we successfully developed QSPR models based on multilinear regression for the flash point of chemical mixtures in previous works (Fayet and Rotureau, 2019; Gaudin et al., 2015), different algorithms (random forest, support vector regressions, neural networks) are here applied to derive new models for the prediction of this safety property. By comparing existing models with the new ones, issued from various machine learning approaches, this research aims to evidence the possible improvements in prediction accuracy and model applicability for mixtures offered by advanced AI techniques compared to traditional multilinear regression methods, as well as their limits.

2. QSPR models to predict the physical hazards of mixtures

The development of QSPR models for mixtures is still challenging, in particular when focusing on physical hazards. In a recent review (Fayet and Rotureau, 2023), we only identified 23 articles (all published after 2013) that were related to a few physical hazard properties (mainly flammability). Indeed, experimental data related to these properties are more likely available than other physical hazards due to required safety data linked to the development of new fuels for example. However, these data (about 1500 data for the largest dataset used) are, in general, not very diversified in terms of included chemical compounds (a limited number of different pure compounds) and mostly dedicated to simple mixtures (binary or ternary).
The most commonly used approach to encode the concentration effects in these models consisted of the calculation of mixture descriptors obtained from molecular descriptors of the different constituents and their respective concentrations in the mixture according to different formulas like those proposed in previous work (Gaudin et al., 2015) (see Table 1).

Table 1: List of formulas used to calculate mixture descriptors

ID	Formula	ID	Formula
mol_sum	$D = x_1 d_1 + x_2 d_2$		
fmol_diff	$D = \lvert x_1 d_1 - x_2 d_2 \rvert$		
sqr_fmol	$D = x_1^2 d_1 + x_2^2 d_2$		
root_fmol	$D = \sqrt{x_1} d_1 + \sqrt{x_2} d_2$		
sqr_fmol_sum	$D = (x_1 d_1 + x_2 d_2)^2$		
norm_cont	$D = \sqrt{(x_1 d_1)^2 + (x_2 d_2)^2}$		
mol_dev	$D = (1 - \Delta x)\Delta d$		
sqr_mol_dev	$D = (1 - \Delta x^2)\Delta d$		
mol_dev_sqr	$D = (1 - \Delta x)^2 \Delta d$		

Most of the models identified in literature were developed based on multilinear regressions. Only few explored, more complex non-linear approaches (like neural networks or random forest). These methods could be interesting to take into account the complexity of mixtures, but their potential would be to confirm. Attempts to develop new QSPR models based on advanced IA are described in the following.

3. Application of advanced AI methods to predict the flash point of mixtures

3.1 Data preparation and exploration

The data set used for this study was the same as the one used in the last model developed by Ineris (Fayet and Rotureau, 2019) to avoid the possible influence of additional data on the new model compared to the MLR model proposed in 2019. This database contains 650 data points for 60 binary liquid mixtures (involving 47 different pure compounds). They have been identified and selected from a wide literature compilation (18 different publications). All flash points were measured in closed cup apparatus according to ASTM D56, ASTM D93 or ASTM D3828 standards.
The preliminary data exploration exhibits information that can impact the development of the model in the following. At first, two compounds are highly represented in the dataset (as shown in Figure 1): methanol and n-octane, which are associated with respectively 12% and 10% of the data points. Looking more globally at the

chemical families involved in the mixtures, it appeared that alcohol-hydrocarbon mixtures are particularly represented (37% of the data points). This suggests their importance in studies related to the flammability risks of liquid mixtures, and predictive models should be particularly robust and accurate for these substances. However, the imbalance between the compounds/mixtures could lead to lower predictive performances for other compounds and to generalization issues in the models.

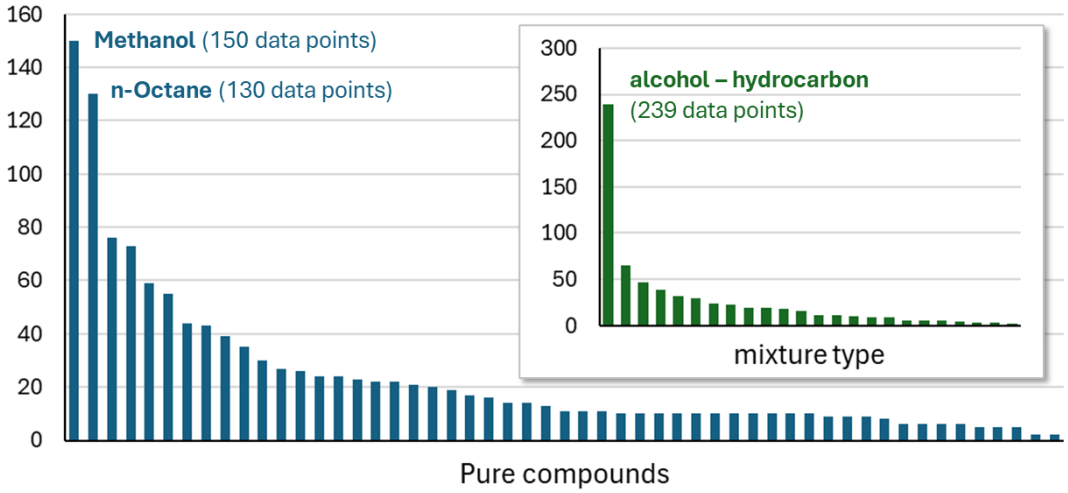

Figure 1: Distribution of data points in the dataset by pure compounds and by type of mixtures (in terms of chemical families)

We also analysed the property distribution in the database (Figure 2). One may notice that only few data are above 60°C (only 9%) indicating that almost all the tested mixtures were flammable. So, the developed models might be more relevant to compare the flash point among flammable mixtures than to identify non-flammable mixtures from flammable ones.

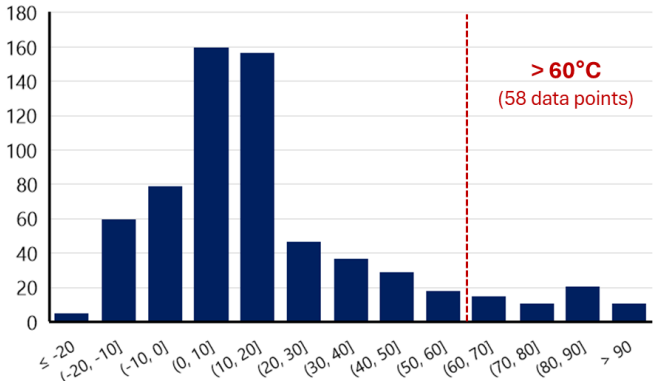

Figure 2: Distribution of flash point (in °C) data in the dataset

At last, it may be noticed that different shapes of flash point profiles are observed. Indeed, as illustrated in Figure 3, some mixtures present quasi-ideal behaviours (like in the case of 1-propanol / 1-pentanol mixture), whereas others are strongly non-ideal. For instance, the propanol / n-octane mixture presents a minimum flash point behaviour, with the flash point of the mixture lower than the ones of the pure compounds. Another example of non-ideal profile is the n-decane / acetone mixture. Acetone is highly flammable, and a large concentration of n-decane (less flammable) is needed to increase the flash point, but this increase is then fast between 0.9 and 1 in mole fraction of n-decane.

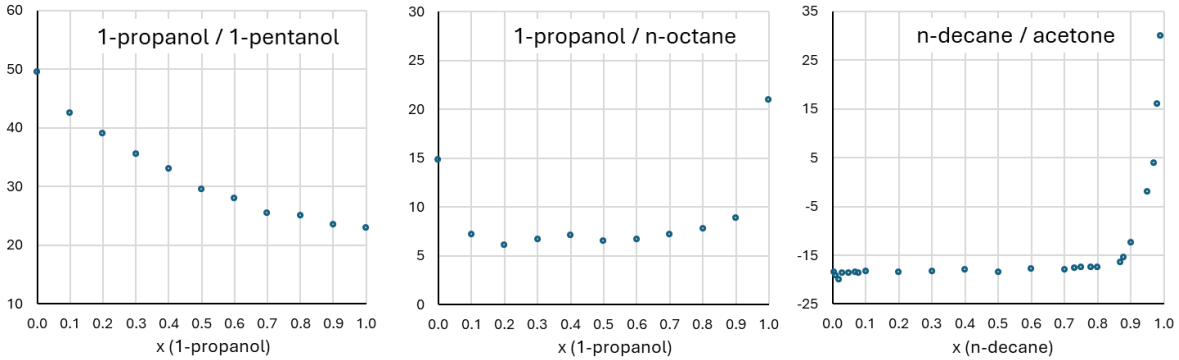

Figure 3: Examples of flash point profiles: 1-propanol/1-pentanol (Hristova and Damgaliev, 2013), 1-propanol/ n-octane (Noorollahy et al., 2010) and n-decane/acetone (Liaw et al., 2008)

3.2 Descriptor calculations

Once again, no change was made to the descriptors used for the development of models that included quantum-chemical descriptors to focus the investigation on the impact of machine learning methods on the final models. The structures of the 47 pure compounds in the database were calculated using density functional theory (DFT) with the Gaussian09 package at B3LYP/6-31+G(d,p) level, which is commonly used for organic compounds. Based on these molecular structures, over 300 molecular descriptors were then computed using Codessa software. These descriptors fall into different categories: (i) constitutional descriptors, which relate to the number and occurrence of specific molecular features, such as the presence and number of specific bonds or chemical groups; (ii) topological descriptors, calculated from the connectivity table of molecules; (iii) geometric descriptors, related to the three-dimensional (3D) structure, such as distances, angles in the molecules, or molecular volumes; and (iv) quantum chemical descriptors, which provide information on binding, energetic, electronic, and thermodynamic properties. In addition to those calculated by Codessa, some descriptors were determined manually, such as the presence and counts of functional groups identified in the dataset, like alcohol functions. This was also the case for quantum chemical descriptors, such as conceptual DFT descriptors, as done in previous works (Fayet et al., 2009).

To account for the trends followed by the property with the concentration of its constituents, the series of mixture formulas proposed in previous work (Fayet and Rotureau, 2019) was used. These so-called mixture descriptors D are based on the molecular descriptors d_i and the molar fraction x_i of each component in the mixture, as summarized in Table 1.

In the case of the flash point of mixtures, their profile with concentration does not follow a simple dilution effect. Indeed, it is influenced by different factors, such as the affinity between compounds and their respective vapour pressures. Therefore, both linear and non-linear formulas were considered, as presented in Table 1.

3.3 Development and validation of the models

Different machine learning algorithms were tested (using a Python environment) to develop predictive models of the target property from the mixture descriptors: Support Vector Regression (SVR), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Artificial Neural Networks (ANN). Models were first developed on the training set. Descriptor selection was done using *Recursive Feature Elimination* (RFE) coupled with RF training to estimate the importance of each descriptor. The less important descriptors were eliminated until they reached the best cross-validation performances for the model. Model quality was evaluated based on the correlation within the training set (coefficient of determination R^2 and mean absolute error MAE) complemented by a cross-validation procedure to assess the models' robustness.

An external validation was conducted to assess the predictive performance of models. The classical method used for the validation of models for pure compounds was adapted to better account for the specificity of mixtures. Indeed, in a database of pure compounds, each data sample is independent, but in the case of mixtures, several data points are related to the same pure compounds. Muratov et al. proposed three external validation strategies applicable to QSPR models applied to mixtures (Muratov et al., 2012):

- **Points-out**: In this strategy, data points are distributed between the training set and the validation set without any consideration of the fact that two points concern the same compounds (in different ratios). So, each mixture of two compounds is present simultaneously in both the training and validation sets; only the ratio of the compounds in the mixture is different. This method is the simplest to apply and

reflects the capacity of the model to predict the property for mixtures already present in the dataset but with new concentration ratios. This strategy is used in the large majority of the published models.

- **Mixture-out**: All data points corresponding to a mixture (i.e. composed of the same constituents but in different ratios) are placed in the same dataset (training or validation). So, each mixture is present either in the training set or in the validation set, but never in both. More rarely used, this method should give a higher prediction error values than the points-out strategy, but it allows an evaluation of the prediction capacity of the model for new mixtures.
- **Compounds-out**: In this last strategy, pure compounds and their mixtures are placed in the same dataset. Each mixture of the validation set contains at least one compound absent from the training set. This is the most rigorous validation method for mixtures, that allow to assess the capacity of the model to predict the property of mixtures containing new chemical compounds. This strategy is the more difficult to be implemented in practice. So, it was only rarely used in existing literature.

In this study (as in our previous models), the compounds-out scheme was favored. Partition was even done, here, by considering a balanced representation of the chemical families in each set.

4. Results

The results obtained using the different algorithms are presented in Table 2.

Table 2: Performances of the developed models using different algorithms for the flash point of mixtures

	SVR	RF	XGBoost	ANN
R^2_{train}	0.90	0.99	0.98	0.99
R^2_{valid}	0.88	0.91	0.90	0.87

The performances of the models developed using different machine learning methods are finally almost the same, with a significant decrease between the training and the validation. This could be due to the intrinsic complexity of these algorithms very powerful for big data problems, but that might be subject to overparameterization for small datasets.The main limitation of these models remains the size and diversity of the data sets available for the training and validation of the models. In particular, the data associated to highly non-ideal profiles represent a real training challenge for the models. For instance, when eliminating all mixtures presenting a minimum flash point, better models were obtained with the ANN algorithm (R^2_{train} = 0.98 and R^2_{valid} = 0.94), as illustrated in Figure 4.

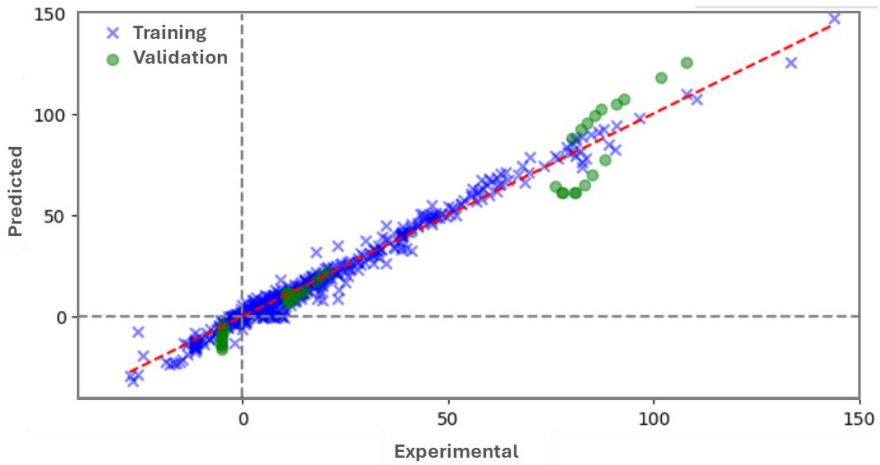

Figure 4: Performance of the ANN model developed after exclusion of the Minimum Flash Point Mixtures

The detailed performance metrics for this ANN models are provided in Table 3 and compared to the ones of our previous MLR model. If the training statistics are better for the ANN model, the predictive capabilities are the same for the two models in the validation set (in terms of MAE), despite the fact that the more complex mixtures were excluded from the development of the ANN model. Moreover, with a smaller difference in the MAE between training and validation observed for the MLR models, a lower risk of overparameterization is expected. So, the simpler MLR model remains finally more favorable on this case study.

Table 3: Performances of the best ANN model compared to previous MLR model (Fayet and Rotureau, 2019)

	ANN (no MinFP)	MLR (All data)
MAE_{train} (°C)	3.0	7.6
R^2_{train}	0.98	0.82
MAE_{valid} (°C)	7.6	7.3
R^2_{valid}	0.94	0.89

5. Conclusions

If the QSPR approach was initially designed for pure compounds, QSPR models for the prediction of physical hazards of mixtures recently appeared. These models are mainly developed based on multilinear regressions. This study aimed at investigating the potential of more complex machine learning approaches.

Despite the complexity of the phenomena associated with flammability properties, advanced machine learning tools didn't highlight better performances than MLR models (as shown in Table 3). Indeed, the amount of available data is relatively limited and associated with a small diversity of mixtures since only a few pure compounds are represented among the available mixture data. In this context, advanced machine learning methods, particularly powerful in the case of big data problems, revealed more likely subject to overparameterization issues than MLR regression due to their native complexity. Nevertheless, they could reveal their full potential when extended dataset would become available. So, efforts to expand the available datasets should be a main axis of progress in the field.

At the time being, mixing rules based on phenomenological thermodynamic principles might still be preferred, of course when available and applicable, like in the case of flash points. QSPR models could help to fulfil their limitations, notably in two situations. First, QSPR models for pure compounds can be used to access the data on pure compounds required in the mixing rules. Secondly, QSPR models for mixtures could be developed and used when no mixing rule is available, even if their use can be recommended rather for screening purposes and for first estimations before conducting experimental tests.

References

Aal E Ali R.S., Meng J., Khan M.E.I., Jiang X., 2024, Machine learning advancements in organic synthesis: A focused exploration of artificial intelligence applications in chemistry, Artificial Intelligence Chemistry, 2, 100049.

Fayet G., Joubert L., Rotureau P., Adamo C., 2009, On the use of descriptors arising from the conceptual density functional theory for the prediction of chemicals explosibility, Chemical Physics Letter, 467, 407-411.

Fayet G., Rotureau P., 2019, New QSPR Models to Predict the Flammability of Binary Liquid Mixtures, Molecular Informatics, 38, 1800122.

Fayet G., Rotureau P., 2023, QSPR models to predict the physical hazards of mixtures: a state of art, SAR and QSAR in Environmental Research, 34, 745-764.

Gaudin T., Rotureau P., Fayet G., 2015, Mixture Descriptors toward the Development of Quantitative Structure-Property Relationship Models for the Flash Points of Organic Mixtures, Industrial & Engineering Chemistry Research, 54, 6596-6604.

Hristova M., Damgaliev D., 2013, Flash point of organic binary mixtures containing alcohols: experiment and prediction, Central European Journal of Chemistry, 11, 388-393.

Joshi P.B., 2023, Navigating with chemometrics and machine learning in chemistry, Artificial Intelligence Review, 56, 9089-9114.

Liaw H.-J., Chen C.-T., Gerbaud V., 2008, Flash-point prediction for binary partially miscible aqueous–organic mixtures, Chemical Engineering Science, 63, 4543-4554.

Lin J., Mo F., 2024, Empowering research in chemistry and materials science through intelligent algorithms, Artificial Intelligence Chemistry, 2, 100035.

Muratov E.N., Varlamova E.V., Artemenko A.G., Polishchuk P.G., Kuz'min V.E., 2012, Existing and Developing Approaches for QSAR Analysis of Mixtures, Molecular Informatics, 31, 202-221.

Nieto-Draghi C., Fayet G., Creton B., Rozanska X., Rotureau P., De Hemptinne J.-C., Ungerer P., Rousseau B., Adamo C., 2015, A General Guidebook for the Theoretical Prediction of Physico-Chemical Properties of Chemicals for Regulatory Purposes, Chemical Reviews, 115, 13093-13164.

Noorollahy M., Moghadam A.Z., Ghasrodashti A.A., 2010, Calculation of mixture equilibrium binary interaction parameters using closed cup flash point measurements, Chemical Engineering Research and Design, 88, 81-86.