



**HAL**  
open science

## QSPR models to predict the physical hazards of mixtures: a state of art

G. Fayet, P. Rotureau

► **To cite this version:**

G. Fayet, P. Rotureau. QSPR models to predict the physical hazards of mixtures: a state of art. SAR and QSAR in Environmental Research, 2023, 34 (9), pp.745-764. 10.1080/1062936X.2023.2253150 . ineris-04277160

**HAL Id: ineris-04277160**

**<https://ineris.hal.science/ineris-04277160v1>**

Submitted on 5 Dec 2023

**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.

# QSPR models to predict the physical hazards of mixtures: a state of art

Guillaume Fayet<sup>a,\*</sup>, Patricia Rotureau<sup>a</sup>

<sup>a</sup>Ineris, Parc Technologique Alata, 60550 Verneuil-en-Halatte, France

\*Corresponding author: guillaume.fayet@ineris.fr; tel: +33(0)344618126

## Abstract

Physical hazards of chemical mixtures, associated for example with their fire or explosion risks, are generally characterized using experimental tools. These tests can be expensive, complex, long to perform and even dangerous for operators. Therefore, for several years and especially with the implementation of the REACH regulation, predictive methods like quantitative structure-property relationships have been encouraged as alternatives tests to determine (eco)toxicological but also physical hazards of chemical substances.

Initially, these approaches were intended for pure products, by considering a molecular similarity principle. However, additional to those for pure products, QSPR models for mixtures recently appeared and represent an increasing field of research.

This study proposes a state of the art of existing QSPR models specifically dedicated to the prediction of the physical hazards of mixtures. Identified models have been analyzed on the key elements of model development (experimental data and fields of application, descriptors used, development and validation methods). It draws up an overview of the potential and limitations of current models as well as areas of progress towards enlarged deployment as a complement to experimental characterizations, for example in the search for safer substances (according to safety-by-design concepts).

**Keywords:** Quantitative Structure-Property Relationships, physical hazards, flammability, explosivity, mixtures.

## 1. Introduction

Human beings and the environment are confronted with mixtures of chemical substances/products, in constant evolution. These are manufactured industrial products containing different constituents mixed intentionally (e.g. pesticides, cosmetics, fuels, additives) or accidentally. For example, the fuel sector presents important industrial challenges with the continuous search for improved performance (combustion properties) of fuels mixtures from various sources (petroleum, renewable biomass) [1]. The addition of alcohols, ethers and oxygenates in alternative fuels impacts their properties (e.g. ignition) [2].

The characterization of the physical hazards of chemical substances is required to apprehend the risk of fire or explosion, for instance, in the perspective of their safe handling, storage and transportation. But, beyond the intrinsic hazards of each constituent, other parameters have to be considered in their evaluation (for classification purpose or in process safety analyzes) [3-5]: the concentration of each component in the mixture, their respective physical properties (e.g. volatilities), miscibility (in liquid mixtures) and heterogeneity (of solid mixtures), molecular interactions between constituents in the mixture (like in non-ideal liquid mixtures). These factors can strongly increase or decrease the value of the property of the mixture and can even reveal synergistic or antagonistic effects.

For this reason, knowledge of the hazardous properties of pure compounds can be not sufficient (and conservative) for risk assessment purpose. Indeed, some mixtures may be more dangerous than their constituents when used as pure compounds. This is notably the case of flammable liquid mixtures presenting a so-called Minimum Flash-Point Behavior [6, 7]. Such mixtures present an important positive deviation from ideality such a way that the flash point reaches a minimum value (in the concentration profile), lower than that of the pure products. For instance, as shown in Figure 1, the flash points measured by Liaw et al. [6] on octane-ethanol mixtures reached a minimum of about 5°C at 0.5 in mole fraction whereas pure octane and ethanol presented flash points of 15°C and 13°C, respectively. In such case, the flash point of the most flammable compound is not a conservative value and its use for risk assessment purpose would lead to underestimate the flammable potential of the mixture.

Similarly, changes in mixture compounds should imply new hazards evaluations to check if the modification of composition can increase for example the flammability of the mixture. Such question is asked for new multi-component substances but also in risk assessment studies. For instance, the evaporation of an inert component (during a process or during long term storage) can increase the flammability of the mixture. [8]

To investigate physical hazards of mixtures and the impact of the mixture composition, the required characterization tests can be costly, complex, long to perform and even dangerous for the operator. It can be inconceivable to experimentally characterize the hazards for too large numbers of substances, even more when concerning mixtures, which represent large part of industrial substances. For this reason, computational methods represent valuable tools to complete the experimental approach and fill data gaps, prioritize experimental campaigns or provide a large amount of data for screening applications in industry for example.

Some mixing rules exist to estimate some properties of mixtures, like Le Chatelier's law [9] for the prediction of flammability limits or thermodynamic-based methods for the flash point (e.g. the

methods of Gmehling & Rasmussen [10] and Liaw [11]). These last are notably cited as possible estimation methods, for the classification of flammable liquids (under some conditions) in the recommendations for the Transport of Dangerous Goods [12] and in the Globally Harmonized System of classification and labelling of chemicals (GHS) [13]. But they require knowledge of the values of the target property for the different constituents and sometimes additional thermodynamic parameters (like activity coefficients). Moreover, such mixing rules are only available for few properties related to physical hazards.

For several years and especially with the implementation of the REACH regulation (Registration, Evaluation, Authorisation and Restriction of Chemicals) [14], QSAR/QSPR (quantitative structure-activity/property relationships) have been encouraged and used as fast and economical alternatives to tests to determine the (eco)toxicological hazards of chemical substances but also physical hazards such as explosivity and flammability. Recommendations, based on five principles, were even proposed by the OECD (Organisation for Economic Cooperation and Development) in 2004 for the validation of QSAR/QSPR models in order to facilitate their use in a regulatory framework [15]. These methods are also of particular interest in the optimization of experimental campaigns for the definition of formulations in a safer-by-design approach.

But these methods are based on a similarity principle according to which molecules with similar structures present similar properties. So, the QSPR approaches as well as their development and validation principles were originally designed for the development of predictive models of pure product properties. The prediction of the mixtures properties presents additional scientific and methodological challenges to take into account the specificities of mixtures. In 2012, Muratov [16] analyzed existing methods toward QSAR/QSPR predictions for mixtures, highlighting the use of mixture descriptors of non-additive type to account for interaction effects between constituents. Recommendations were also given on methods for examining and using the databases, but also for the appropriate validation of these models.

Recently, reviews of existing QSAR models capable of predicting toxicological properties of mixtures was carried out by Ineris (Mombelli [17]) and by Belfield et al. [18], identifying and examining in detail respectively 54 and 40 relevant scientific papers (mainly acute toxicity oriented and more associated to environmental impacts (than human health effects) for pesticides, pharmaceuticals, industrial products, priority pollutants). The limitations of the existing models were exposed, and recommendations were proposed to improve the relevance and quality of predictions. At first, they pointed out the emerging state of these developments. For instance, no computer tool (available online) already implements such models. If the performances of most models were proposed to be comparable to those of pure compound models, this could be to confirm in the future when more models and data will be available. The authors in particular encouraged to better taking into account the potential interaction effects between constituents in the definition of mixture descriptors and considering more realistic exposure scenarios (towards more complex environmental mixtures than the theoretical mixture of different chemicals). To achieve these goals, the gathering of larger databases including mixtures associated with realistic exposure scenarios is recommended. A particular attention should be required on their reliability, storage, and accessibility. Moreover, information issued from other modeling approaches (like Adverse Outcome Pathways and read-across techniques) could help to incorporate more information on interaction effects into the models. At last, when dealing with more complex mixtures, sophisticated machine learning approaches could improve

the model performances even if they are in general less transparent and interpretable than linear regressions (in the perspective of regulatory applications).

Considering physical hazards (associated with combustion, flammability, explosivity risks), the availability of QSPR models increased in the last two decades [19-22]. But they are still mostly limited to few specific hazardous physico-chemical properties (such as flash point or flammable limits), for which experimental data are the more likely available. Moreover, they were up to now mainly limited to pure compounds. These QSPR models for pure compounds can be associated to mixing rules, as demonstrated for the prediction of flash point of binary mixtures [23]. Such approach demonstrated good results, predictions being mostly limited by the applicability of the mixing rules and the reliability of the predictions for the flash points of each constituent. First QSPR models directly dedicated to the physical hazards of mixtures were only published recently, for the flash point some years ago (Saldana [24] and Gaudin [25]).

For this reason, this study proposes to review existing literature associated to the development of QSPR models dedicated to the prediction of physical hazards of mixtures to draw up an overview of the potential and limitations of current models and to identify possible perspectives to help the deployment of these new methodological approaches that are complementary to experimental characterization.

## **2. Scope of investigation and literature collection**

In this study, chemical mixtures of pure compounds are particularly addressed, i.e. pure compounds that have been put together, mixed or brought into contact. The specific case of impurities is not considered. Moreover, the properties of mixtures of substances are addressed but not the possible reactivity observed when substances come into contact (chemical incompatibilities).

This review is focused on publications available in the scientific literature accessible in June 2022 via the Web of Science platform [26]. This review targets in particular hazardous physico-chemical hazards used for the classification of substances in the 16 physical hazard classes of the GHS (e.g. explosive, flammable liquids or gases, oxidizing substances) or in the annex VII of the REACH regulation (e.g. self-ignition temperature).

A search with too simple and generic keywords brings up a large number of publications that are not related to predictive models such as QSPR and/or physical hazards. For example, a search with the keywords "Prediction" and "Mixture" brings up more than 39,000 references, but none of which are relevant among the first 50 proposed (ranked by relevance by the Web of Science platform).

Therefore, refined queries were needed to identify as much as possible all publications related to QSPR models dedicated to physical hazards while eliminating irrelevant articles. The first one searches all references for which the titles, abstracts or keywords contain both the word "mixture", a term related to predictive methods ("QSPR" or "structure-property") and a term related to the targeted properties ("flammability" or "explosibility" or "flash point" or "self-ignition" or "auto-ignition" or "combustion" or "explosiv\*" or "detonation" or "deflagration" or "decomposition heat" or "heat of decomposition" or "decomposition temperature" or "impact sensitivity" or "friction sensitivity" or "combustion heat" or "heat of combustion"). The second is focused only on the title of the references by adding "predict\*" among the terms related to predictive methods.

The combination of these two queries brings out 139 references among which some publications still do not concern QSPR models or aren't dedicated to the targeted properties (such as the octane number of fuels [27] or the diffusion coefficient of organic compounds [28]). Articles only mentioning the issue of prediction of mixture properties without proposing predictive models or only quoting existing models are also excluded at this stage.

Finally, 23 publications corresponding to the proposition of QSPR models for physical hazards of mixtures were selected (listed in Table 1). These articles reflect an emerging but active field of research since they were all published between 2013 and 2022.

In some publications, several models (investigating for instance the potential of different methods) are proposed for the same property. In such cases, all the proposed models are not considered independently but only one single model is considered in the statistical analyses proposed in the following (whether the authors have mentioned a preferred model or not).

### **3. Results and discussion**

This study doesn't propose a catalogue of models with their details presented individually. Such information is available in the original publications. It rather proposes a general state of art of the availability of these models, of the methods used and of their limits in the perspective of future developments towards new and/or better models to predict the physical hazards of mixtures.

#### **3.1. Endpoints and target chemical families**

For mixtures, even more than for pure products, the availability of sufficient quality experimental data is the main limitation to the access of QSPR models. Indeed, the reference experimental data used for the development and validation of the models are of great importance since their reliability will influence the performance of the model by the propagation effect of their uncertainty. Moreover, the data included in the training set define the domain of applicability of the model. Thus, in order to obtain efficient QSPR models, a sufficiently large number of reliable experimental data sets is needed, on the one hand, to present a sufficient variability of properties and structures, and, on the other hand, to allow a split into two robust datasets, one for training and another for the external validation.

It is therefore not surprising that the identified models focus on some particular properties and on the same families of substances for which databases are available. In all the publications listed, the data used for the development and validation of the models are derived from the compilation of experimental results collected in the scientific literature. Indeed, the few databases containing information on the physical hazards of substances concern (mainly) pure products, like the OECD eChemPortal database [29], or the German ChemSafe database [30].

As shown on Figure 2, all the identified models concerned the flammability of organic liquids, with the exception of the model of He et al. (2021) [31] which focuses on the decomposition temperature of ionic liquids. It should be noted that the property predicted by this last model is not the temperature of onset of decomposition (used for example as a pre-selection criterion for the classification of organic peroxides, self-reactive materials or explosive materials) but the temperature at 5% decomposition ( $T_{d,5\%}$ ). The flammable properties of interest are the most commonly used data for assessing the

flammability of liquids and gases, namely the flash point (FP), lower and upper flammability limits (LFL/UFL), and auto-ignition temperature (AIT).

Although mixtures in industry can contain numerous constituents, the studied flammable liquid mixtures are relatively simple binary and/or ternary organic mixtures. Among the models developed for flash point of binary mixtures, two studies tested their potential on ternary mixtures (Fayet (2019) [32] and Torabian (2019) [33]) and showed similar performance to that observed for binary mixtures.

If the datasets used for the development of models contain sometimes large number of data, they remain relatively modest in terms of the chemical diversity of mixtures. As shown in Supporting Information (Table S1), most of these models are relatively global, targeting organic compounds rather than focused on a specific family. They are constituted around a limited number of classical organic compounds (hydrocarbons, alcohol, ketones, acids) used for example in the composition of fuels for which an evaluation of flammability is necessary in an industrial context, explaining the availability of experimental data.

For example, the largest database encountered was used by Jiao et al. (2020) [34] to develop their model for the flash point. It contained 1458 data but for binary and ternary mixtures from (only) 47 different pure products. Moreover, this dataset was chemically unbalanced with 3 overrepresented compounds (methanol, octane, and ethanol) found in 36%, 27%, and 26% of the data, respectively (as shown in Figure 3).

In the case of models for ternary mixtures, the number of data may sometimes seem particularly large, but they generally concern only few different mixtures (based on the same compounds, only concentrations changing). For example, the model of Toropova (2020) [35] specifically developed for ternary mixtures is based on 808 flash point data but concerned only 8 different mixtures.

Two exceptions on this point are the models developed by Wang et al. (2018) [36] and by Aljaman et al. (2022) [37] for the flash point. In these studies, a database was composed of flash points values for both pure compounds (548 and 474 data, respectively) and mixtures (753 binary and 79 ternary for Wang et al.; 279 binary, 26 ternary, 6 quaternary and 8 quinary for Aljaman (2022)). This approach could appear as a way to the lack of chemical diversity of constituents in the databases of mixture properties. But the number of data related to pure compounds is significantly larger than those for mixtures. So, the ability of the final model to account for concentration and interaction effects between constituents might be to check, in particular for those only represented as pure products in the database.

### **3.2. Molecular and mixture descriptors**

In a QSPR model for pure compounds, the molecular structure is represented by a series of molecular descriptors. Many descriptors have been proposed and are available with different levels of complexity. For instance, 1D or constitutional descriptors are based on the knowledge of the sole composition of the molecule (ex: number of atoms or specific groups). 2D descriptors consider the topological structure, i.e. the way in which the atoms are connected to each other (thus distinguishing isomers for example). Finally, 3D descriptors can encode the spatial arrangement of the molecule (geometrical descriptors) and information, resulting from quantum chemistry calculations, on molecular electronic properties and on the reactivity of molecules (quantum-chemical descriptors).

These descriptors can finally consider either the molecules in their entirety (integral descriptors) or only some parts of them (fragment descriptors).

In the case of a mixture model, beyond the molecular structures of each chemical component, other parameters can influence the property and in particular the concentration of each constituent and interactions between them in the mixture, which can result in synergistic or antagonistic effects.

To account for these factors, most of the models identified in this review introduced Mixture Descriptors  $D_m$  calculated by combining the molecular descriptors  $d$  for each component via a mixing formula according to their respective concentrations. An additivity of the contribution of each constituent is in general considered using a simple linear weighting with respect to the molar fraction  $x$  (or more rarely the volume fraction), according to the following equation:

$$D_m = \sum x_i d_i \quad (1)$$

Other mixture formulas have also been proposed to consider the non-linearity of concentration effects encountered in mixtures. Some of them aim at taking directly into account a non-linear dependence of the property with respect to the molar fraction, while others reflect a deviation from the linear behavior with respect to the concentration. For example, 12 formulas (in Table 2) have been tested for the search of predictive models of the flash point of binary mixtures by Gaudin et al. [25]. Two main types of formulas were proposed. Some are intended to be directly correlated (linearly as in Eq. 1, or not) with the property (direct combinations) whereas the others (based on the difference in mole fractions between the components of the mixture) might relate to the deviation of the property from the linear contribution of each component with their respective concentrations. Moreover, they have been defined to fit with some imposed constraints. For instance, they must be independent of the order of the two components of the binary mixtures (e.g. 0.1 ethanol in 0.9 octane is the same as 0.9 octane in 0.1 ethanol) and the mixture descriptors for pure compounds (with a mole fraction of 1) should be 1 for direct formulas and 0 for deviation formulas.

Each formula was investigated independently in multilinear regressions, and the best models were obtained for direct formula, either when considering the whole database or when analyzing independently the results on the different types of mixtures involved in it (e.g alcohol/alcohol, hydrocarbon/alcohol, alcohol/ester...). Moreover, formulas based on the difference in mole fractions are limited to the case of binary mixtures, since they cannot be generalized to more complex mixtures. Towards application to ternary or more complex mixtures, formulas based directly on the mole fraction (or concentration) of each component might be favored.

Different mixing formulas were finally combined in the last model developed by Fayet et al. in 2019 [32] with increased performances compared to the models based on a single type of mixing rule ( $MAE_{ext} = 7.3^\circ\text{C}$  vs.  $10.3^\circ\text{C}$ ).

$$FP (^{\circ}\text{C}) = 20.3 + 28.6 \sum x_i T_i^E + 24.6 \sum \sqrt{x_i} \chi_i + 59.6 \sum \sqrt{x_i} V_{YZ,i} - 315.0 \sum \sqrt{x_i} Q_{H,max,i} - 107.6 \sum \sqrt{x_i} V_{H,min,i} + 2.0 (\sum \sqrt{x_i} \mu)^2 \quad (2)$$

For the same property (flash point of binary mixtures), Torobian [33] investigated particularly three direct formulas (linear combination, quadratic mean value, or cubic mean value) for the development



of multilinear models. They obtained the best results with the simple linear combination. So, they focused their final neural network model on this formula.

Various types of formulas (including those proposed by Gaudin et al. [25]) were also investigated for the lower flammability limits by Pan (2019) [38] and for the auto-ignition temperature by Jin (2020) [39].

For the LFL, Pan et al. [38] achieved to similar conclusions than Gaudin et al. for the flash point. Direct formula appeared as the more relevant (in MLR focused on a single mixing formula). The best model was finally obtained using the norm of the molar contribution formula (Eq. 3).

$$D_m = \sqrt{\sum(x_i d_i)^2} \quad (3)$$

Jin et al. (2020) [39] also identified direct-type formulas as the most favorable ones for their models for the AIT of binary mixtures. In particular, the simple linear weighting (Eq. 1) gave the best results based on multilinear regressions and was then used in the final support vector machine model.

In this work, various integral molecular descriptors (constitutional, topological, geometrical and quantum-chemical) were employed. These classical descriptors characterize isolated molecules, although some may reflect their potential to create intermolecular interactions. For example, in this model,  $Q_{H,max}$  and  $V_{H,min}$  respectively designate the maximum charge and minimum valence for a hydrogen atom and reflect the potential of the products involved in the mixture to form hydrogen bonds with other constituents.

This use of mixing formulas to derive mixture descriptors has also been applied on group contributions or fragment descriptors. For example, in their group contribution approach for the prediction of the TAI of binary mixtures, Ye et al [40] introduced a formulation taking into account the number of occurrences of contribution groups in each molecule and their respective volume fractions.

$$TAI = a + bX + cX^2 + dX^3 + eX^4 \quad (4)$$

$$\text{with } X = v_1 \sum_i n_{1i} f_i + v_2 \sum_j n_{2j} f_j \quad (5)$$

where  $n_1$  and  $n_2$  are the numbers of occurrences of a group in each of the two substances in the binary mixture,  $f_i$  is the associated contribution and  $v_i$  is the volume fraction of each constituent.

Similarly, Aljaman et al [37] considered the mass concentration associated with different groups present in the molecule ( $CH_3$ ,  $CH_2$ ,  $CH=CH_2$ ,  $CHO$ ,  $COO$ , etc.) for the prediction of flash point of oxygenated fuels.

Most of the identified fragment-models use the Simplex Representation of Molecular Structure (SiRMS) approach [41]. This latter considers the decomposition of molecules into fragments of 2 to 6 atoms called simplexes. An originality of the application of this approach to the case of (binary) mixtures is the introduction of so-called unbound simplexes in which some atoms constituting the simplex are not bound one to each other. They can belong to both molecules of the binary mixture and thus characterize specific intermolecular interactions in the mixture. Two different mixture formulas are then applied: Eq. 6 for (intramolecular) bonded simplexes  $d_i$  and Eq. 7 for (intermolecular) unbound simplexes  $d_{1+2}$ .

$$D_m = x_1 d_1 + x_2 d_2 \quad (6)$$

$$D_m = \sum x_i d_{i+2} \quad (7)$$

This approach was used by Shen et al. (2019) [42] for the auto-ignition temperature and then by Yao et al. (2020) [43] and Cao et al. (2020) [44] for the flash point, in all cases for binary organic liquid mixtures.

He et al [31] used another fragment-based approach, by using ISIDA descriptors, to predict the decomposition temperature of ionic liquid mixtures. To account for the specificity of ionic liquids (consisting of an anion and a cation), some fragments were proposed to characterize explicitly the anion-cation interaction within an ionic liquid, but no fragment was proposed to encode possible interaction between two ionic liquids. As for the SiRMS approach, mixing rules are used to include concentration effects.

Finally, in contrast to the previously mentioned works, Toropova et al [35, 45] do not apply any mixing formula in their quasi-SMILES approach. quasi-SMILES are defined as extensions of the SMILES code [46] for molecules by adding other symbols for encoding additional information that is specific to the systems under study. In the case of mixtures, the quasi-SMILES code represents the entire mixture and takes the following form:

$$[SMILES\#1][\%X1][SMILES\#2][\%X2] \quad (\text{for a binary mixture}) \quad (8)$$

$$[SMILES\#1][\%X1][SMILES\#2][\%X2][SMILES\#3][\%X3] \quad (\text{for a ternary mixture}) \quad (9)$$

where  $[SMILES\#i]$  is the SMILES code of component  $i$  and  $[\%Xi]$  is a string representing its molar fraction in the mixture.

It may be noticed that, in a quasi-SMILES code, the mole fraction is not considered as a continuous value but through a character string assigned to a given concentration range. Moreover, the concentration is considered in the final model in the same way as the fragments identified in the structure of each constituent (in their SMILES code), thus either alone or in combination with a particular molecular fragment.

### 3.3. Methods for model development

The analysis of existing QSPR models for mixtures does not highlight any particularity in the algorithms used when compared to the models dedicated to pure products. As shown in Figure 4, the models identified are mostly based on multilinear regressions (MLR, for Multi-Linear Regression). Few models are based on more complex machine learning approaches such as Artificial Neural Networks (ANN), Support Vector Machines (SVM) or other non-linear methods (non-linear regressions, k-Nearest-Neighbors (k-NN), Random Forest (RF), Bootstrap Tree (BT)). If these non-linear approaches could maybe help to take into account the complexity of mixtures, no definitive demonstration can be highlighted among the existing models.

Some studies compare different approaches and seem to show some potential of nonlinear methods to consider the complexity of phenomena associated with mixture properties. For example, Jiao et al (2020) [47] employed 5 different methods (MLR, k-NN, RF, BT and SVM) for the lower flammability limit of binary hydrocarbon mixtures and found the best performance for the model based on the

Random Forest approach. Nevertheless, considering the limited size of the database used (i.e. 54 data), the performances of the different approaches revealed close one to each other (as shown in Table 3).

### 3.4. Model Validation

To evaluate the performance of a QSPR model, different validation methods are employed. If the quality of the predictions within the training set allows to estimate the goodness of fit of the model, other internal and external validation methods are recommended, notably in the OECD validation principles for regulatory purpose [15]: cross-validation and/or bootstrapping to evaluate the robustness of the model, Y-Randomization to verify that the model was not obtained by chance and external validation on data different from those used for its training (validation set) to evaluate its predictive capacity.

Because the models dedicated to the physical hazards of mixtures are recent, they have all undergone validation beyond the simple correlation between experimental and predicted data from their training set. Only one study (Jiao et al. (2016) [48]) did not include external validation, for which the models were only validated by cross-validation.

Concerning internal validations, 78% of the studies used cross-validations while other internal validations by Bootstrapping or Y-randomization is more rarely used, as shown in Figure 5. No specific adaptation of internal validation methods was proposed until very recently and the work of Chatterjee et al. (2022) [49]. So, the cross validations used in the reviewed models were based on a classical approach (the same than for pure compounds).

Concerning external validation, Muratov et al. [16], yet in 2012, distinguished three types of external validation based on how the validation set differs from the training set:

- **Points-out:** The partition between training and validation sets considers each experimental data individually without considering the fact that they can be associated to the same mixtures and constituents. A mixture can therefore be present in both the training set and the validation set (with different proportions between its constituents). This method mainly evaluates the ability of the models to predict the properties of existing mixtures in new proportions.
- **Mixture-out:** In this type of partition, all data points corresponding to mixtures composed of the same constituents are grouped in the same set (training or validation). The expected error for such an external validation is higher than for the Points-out strategy but it will evaluate the ability of the model to predict the properties of new mixtures beyond the concentration effect alone.
- **Compounds-out:** The data in the validation set only concern mixtures for which at least one of the constituents is absent from the training set. This is the most rigorous method of external validation in QSAR/QSPR modeling of mixtures. If the expected prediction error for this strategy is the largest, the models validated with such a partition will have demonstrated their ability to predict the property under study for mixtures consisting of one or more new products.

Most studied performed external validations based on a Points-out partition (70%), fewer models being validated on Compounds-out partitions (35%) (in some case in association to the other types of partitions). Points-Out external validation is the simplest to implement but it is not, a priori, the most

rigorous way to evaluate the predictive power of the model since the validation set is not strictly independent of the training set, some mixtures consisting of the same products and differing only slightly in concentration can be found in the both the training and validation sets.

However, as highlighted earlier in this study, the available databases present a rather limited chemical diversity. This fact impacts the goodness of fit of the model but also its reliability of external validations, especially in the framework of Mixtures-out or Compounds-out partitions. Indeed, in these cases, the validation set is even more restricted in terms of structural diversity. If a Points-out validation set is not strictly (chemically) independent from the training set, it has in such cases the advantage of offering an external validation on more chemically diverse samples than with Mixtures or Compounds-out partitions. So, none of these different partitions seems to be ideal with the databases up to now used in the case of physical hazards of mixtures. Besides, Wang et al. (2019) [50] applied the three types of partitions and finally recommended to retain the Points-out model because of the better performances observed in external validation (with an  $R^2_{\text{ext}}$  of 0.965 against 0.879 and 0.923 respectively for the Mixtures and Compounds-out partitions).

Finally, all the identified work proposed final models meeting generally expected performances for a valid model with correlations (largely) higher than 0.6 in cross-validation [51] and predictive powers higher than 0.7 in external validation [52]. Moreover, if larger and more diversified databases would be necessary to have more robust evaluations of the predictive power of the models, the quality of prediction observed in these works are close to the performances observed for pure compounds. For instance, in the case of the flash point, the models dedicated to mixtures present average absolute errors in prediction from 3.5K to 10.3K in their external validations. These performances are similar to those observed by Rowley et al. [53] in their evaluation of a series of models on 1062 pure organic compounds, as shown in Figure 6.

### 3.5. Applicability domain

QSPR models are empirical models whose validity is limited to an applicability domain (AD) delimited by the physico-chemical or structural space related to the data for which it has been trained. As already explained in section 3.1, the models have been developed on databases of mixtures involving a relatively limited number of compounds even if they cover various types of organic compounds commonly found for instance in the petroleum industry (alkanes, alcohols, ketones...), as shown in table S1. An exception is the model of He [31] focused on ionic liquids.

65% of the reviewed studies include an evaluation of the applicability domain of the models, using a Williams plot analysis, in the large majority of the cases (13/15). Williams plots are the graphical representation of the leverage approach [54] which consists in calculating the "leverage" associated with each sample with respect to the global distribution of the training set. No particular adaptation was performed for the case of mixtures, except that it was based on mixture descriptors rather than molecular descriptors. It doesn't consider the fact that data can concern same mixtures only differing slightly in concentration.

The only works using another approach are the models of Toropova et al [35, 45] that analyzed the distribution of fragments identified in the quasi-SMILES. This approach was the same as the one they used for pure compound models [55].

In both approaches, all predictions (for each data point) are considered independently without considering that they can relate to the same mixtures and pure compounds. It could be interesting to think about the possibility of considering the specific character of the mixtures by considering not only the chemical diversity of the molecules involved but also the effects of concentrations and intermolecular interactions.

#### 4. Conclusions

If the QSPR approach was initially designed for pure compounds, QSPR models for the prediction of physical hazards of mixtures recently appeared. Mixing rules based on phenomenological thermodynamic principles might still be preferred when available and applicable, but QSPR models can help to fulfill their limitations by using QSPR models for pure compounds to access the data on pure compounds required in the mixing rules. QSPR models for mixtures could even be developed and used when no mixing rule is available.

The development of such models is still an emerging field with only 23 recent publications (identified in the scientific literature from 2013 to nowadays) proposing such QSPR models. These models focus on few properties (mainly flammability), like in the case of pure products, for which data existed in relatively large numbers (up to about 1500 data). But these data are in general not very varied in terms of chemical diversity since they are often limited to few different pure products. Moreover, they are mostly dedicated to binary or ternary mixtures.

To encode the specificities of mixtures (concentration and chemical interactions between constituents), the most commonly used approach is based on the calculation of mixture descriptors from molecular descriptors of the different constituents and their respective concentrations in the mixture. If only few models introduced descriptors characterizing explicitly intermolecular interactions between two constituents (via non-bonded fragments), some classical integral molecular descriptors can translate the potential of individual constituents to favor interactions with other constituents in the mixture.

Beyond that, these models are developed from the same algorithms as the ones used for pure products models, in general by multilinear regressions. More complex nonlinear approaches (like neural networks or random forest) could be interesting to take into account the complexity of mixtures, but their potential would be to confirm on larger datasets.

Due to the recent character of this field of research (all works were published after 2013), all these models were validated according to classical recommendations for valid models (including external validation and with definition of a domain of applicability) and their performances are similar to those obtained for pure compounds. Nevertheless, the robustness of their validations remains impacted by the lack of chemical diversity of the available experimental databases.

Existing QSPR models for the physical hazards of mixtures already complement the experimental approach. Nevertheless, the development of new and more reliable QSPR models could be awaited in the future to extend their applicability and improve prediction accuracies.

To achieve this goal, the most critical needs are experimental databases. Indeed, new data are needed to extend the field of application of the predictive models which are focus on few properties and on classical compounds. New databases are the first steps towards the development of models for other

properties (beyond the flammability of liquids and gases) and over larger ranges of substances (including some emerging substances like biobased solvents or emerging refrigerants). The gathering of additional datasets might also allow to access better models to fulfill the limits observed in the existing models that are generally focused on a limited number of constituents. Such datasets would allow to strengthen the assessment of the predictive power of models by using validation sets that are both independent of the training sets (by favoring a compounds-out approach) while having a diversity of substances chemically representative of the mixtures targeted by the model.

Beyond these primary needs, methodological progress can be expected. At first, the availability of larger databases could open better perspectives for non-linear methods like neural networks that are more and more widely used in artificial intelligence approaches.

Moreover, the methods used for the validation of models and predictions are still the same than for pure compounds, expected the first recommendations proposed for the evaluation of the predictive power of the models. Reflections would deserve to be engaged on the determination of the applicability domain of the models, to take into account the specificities of mixtures. Moreover, the recommendations for the validation of predictions (which depend not only on the model but also on the substance and the way in which the prediction is going to be used) could also be analyzed in order to verify whether specific recommendations should be proposed for the case of mixtures.

## References

- [1] A. Krutof, and K. Hawboldt, *Blends of pyrolysis oil, petroleum, and other bio-based fuels: A review*, *Renew. Sustain. Energy Rev.* 59 (2016), pp. 406-419.
- [2] O.I. Awad, R. Mamat, T.K. Ibrahim, A.T. Hammid, I.M. Yusri, M.A. Hamidi, A.M. Humada, and A.F. Yusop, *Overview of the oxygenated fuels in spark ignition engine: Environmental and performance*, *Renew. Sustain. Energy Rev.* 91 (2018), pp. 394-408.
- [3] L. Centrella, M. Portarapillo, G. Luciani, R. Sanchirico, and A. Di Benedetto, *Synergistic behavior of flammable dust mixtures: A novel classification*, *J. Hazard. Mater.* 397 (2020), p. 122784.
- [4] M.G. Zabetakis, *Flammability characteristics of combustible gases and vapors - Bulletin 627*, Bureau of Mines, United States, 1964.
- [5] O. Dufaud, L. Perrin, D. Bideau, and A. Laurent, *When solids meet solids: A glimpse into dust mixture explosions*, *J. Loss Prevent. Proc. Ind.* 25 (2012), pp. 853-861.
- [6] H.J. Liaw, T.P. Lee, J.S. Tsai, W.H. Hsiao, M.H. Chen, and T.T. Hsu, *Binary liquid solutions exhibiting minimum flash-point behavior*, *J. Loss Prevent. Proc. Ind.* 16 (2003), pp. 173-186.
- [7] H.-J. Liaw, and H.-Y. Chen, *Study of Two Different Types of Minimum Flash-Point Behavior for Ternary Mixtures*, *Ind. Eng. Chem. Res.* 52 (2013), pp. 7579-7585.
- [8] T. Pénelon, V. Debuy, B. Truchot, C. Wagner, L. Donnat, and J.-F. Lechaudel, *Pool evaporation: Experimental tests at medium-scale with gasoline*, *J. Loss Prevent. Proc. Ind.* 65 (2020), p. 104072.
- [9] H. Le Chatelier, *Estimation of firedamp by flammability limits*, *Annals of Mines* 19 (1891), pp. 388-395.
- [10] J. Gmehling, and P. Rasmussen, *Flash points of flammable liquid mixtures using UNIFAC*, *Ind. Eng. Chem. Fund.* 21 (1982), pp. 186-188.
- [11] H.-J. Liaw, and Y.-Y. Chiu, *A general model for predicting the flash point of miscible mixtures*, *J. Hazard. Mater.* 137 (2006), pp. 38-46.
- [12] *Manual of Tests and Criteria, 7th Revised Edition*, United Nations, New York and Geneva, 2019.
- [13] *Globally harmonized system of classification and labelling of chemicals (GHS), ST/SG/AC.10/30/Rev.8*, United Nations, New York and Geneva, 2019.
- [14] *Regulation (EC) N° 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)*, European Commission, 2006.
- [15] *OECD principles for the Validation, for Regulatory Purpose, of (Q)SAR Models*, Organisation for Economic Co-operation and Development (OECD), 2004.
- [16] E.N. Muratov, E.V. Varlamova, A.G. Artemenko, P.G. Polishchuk, and V.E. Kuz'min, *Existing and Developing Approaches for QSAR Analysis of Mixtures*, *Mol. Inform.* 31 (2012), pp. 202-221.
- [17] E. Mombelli, *État de l'art sur la modélisation QSAR de la toxicité des mélanges chimiques*, Ineris, 2022.
- [18] S.J. Belfield, J.W. Firman, S.J. Enoch, J.C. Madden, K. Erik Tollefsen, and M.T.D. Cronin, *A review of quantitative structure-activity relationship modelling approaches to predict the toxicity of mixtures*, *Comput. Tox.* 25 (2023), p. 100251.
- [19] Z. Jiao, H.U. Escobar-Hernandez, T. Parker, and Q. Wang, *Review of recent developments of quantitative structure-property relationship models on fire and explosion-related properties*, *Process Saf. Environ. Protect.* 129 (2019), pp. 280-290.
- [20] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. De Hemptinne, P. Ungerer, B. Rousseau, and C. Adamo, *A General Guidebook for the Theoretical Prediction of Physico-Chemical Properties of Chemicals for Regulatory Purposes*, *Chem. Rev.* 115 (2015), pp. 13093-13164.
- [21] G. Fayet, and P. Rotureau, *Chemoinformatics for the Safety of Energetic and Reactive Materials at Ineris*, *Mol. Inform.* 41 (2022), p. 2000190.

- [22] F.A. Quintero, S.J. Patel, F. Munoz, and M.S. Mannan, *Review of Existing QSAR/QSPR Models Developed for Properties Used in Hazardous Chemicals Classification System*, Ind. Eng. Chem. Res. 51 (2012), pp. 16101-16115.
- [23] T. Gaudin, P. Rotureau, and G. Fayet, *Combining mixing rules with QSPR models for pure chemicals to predict the flash points of binary organic liquid mixtures*, Fire Saf. J. 74 (2014), pp. 61-70.
- [24] D.A. Saldana, L. Starck, P. Mouglin, B. Rousseau, and B. Creton, *Prediction of Flash Points for Fuel Mixtures Using Machine Learning and a Novel Equation*, Energy Fuels 27 (2013), pp. 3811-3820.
- [25] T. Gaudin, P. Rotureau, and G. Fayet, *Mixture Descriptors toward the Development of Quantitative Structure-Property Relationship Models for the Flash Points of Organic Mixtures*, Ind. Eng. Chem. Res. 54 (2015), pp. 6596-6604.
- [26] *Web of Science*, Clarivate <https://www.webofscience.com>.
- [27] G. Cai, Z. Liu, L. Zhang, Q. Shi, S. Zhao, and C. Xu, *Systematic performance evaluation of gasoline molecules based on quantitative structure-property relationship models*, Chem. Eng. Sci. 229 (2021), p. 116077.
- [28] S.A. Mirkhani, F. Gharagheizi, and M. Sattari, *A QSPR model for prediction of diffusion coefficient of non-electrolyte organic compounds in air at ambient condition*, Chemosphere 86 (2012), pp. 959-966.
- [29] *EChemPortal*, OECD, <https://www.echemportal.org/echemportal/>.
- [30] *Chemsafe - Database for safety characteristics in explosion protection*, BAM, Dechema, PTB, <https://www.chemsafe.ptb.de/>.
- [31] H. He, Y. Pan, J. Meng, Y. Li, J. Zhong, W. Duan, and J. Jiang, *Predicting Thermal Decomposition Temperature of Binary Imidazolium Ionic Liquid Mixtures from Molecular Structures*, ACS Omega 6 (2021), pp. 13116-13123.
- [32] G. Fayet, and P. Rotureau, *New QSPR Models to Predict the Flammability of Binary Liquid Mixtures*, Mol. Inform. 38 (2019), p. 1800122.
- [33] E. Torabian, and M. Amin Sobati, *New structure-based models for the prediction of flash point of multi-component organic mixtures*, Thermochim. Acta 672 (2019), pp. 162-172.
- [34] Z. Jiao, C. Ji, S. Yuan, Z. Zhang, and Q. Wang, *Development of machine learning based prediction models for hazardous properties of chemical mixtures*, J. Loss Prevent. Proc. Ind. 67 (2020), p. 104226.
- [35] A.P. Toropova, A.A. Toropov, D. Leszczynska, and J. Leszczynski, *The index of ideality of correlation: models of the flash points of ternary mixtures*, New J. Chem. 44 (2020), pp. 4858-4868.
- [36] Y. Wang, F. Yan, Q. Jia, and Q. Wang, *Distributive structure-properties relationship for flash point of multiple components mixture*, Fluid Phase Equil. 474 (2018), pp. 1-5.
- [37] B. Aljaman, U. Ahmed, U. Zahid, V.M. Reddy, S.M. Sarathy, and A.G. Abdul Jameel, *A comprehensive neural network model for predicting flash point of oxygenated fuels using a functional group approach*, Fuel 317 (2022), p. 123428.
- [38] Y. Pan, X. Ji, L. Ding, and J. Jiang, *Prediction of Lower Flammability Limits for Binary Hydrocarbon Gases by Quantitative Structure—Property Relationship Approach*, Molecules 24 (2019), p. 748.
- [39] Y. Jin, J. Jiang, Y. Pan, and L. Ni, *Prediction of the auto-ignition temperature of binary liquid mixtures based on the quantitative structure—property relationship approach*, J. Therm. Anal. Calori. 140 (2020), pp. 397-409.
- [40] L.-T. Ye, Y. Pan, and J.-C. Jiang, *Experimental Determination and Calculation of Auto- Ignition Temperature of Binary Flammable Liquid Mixtures*, Acta Petro. Sin. (Petroleum Proc. Section) 31 (2015), pp. 753-759.
- [41] V. Kuz'min, A. Artemenko, L. Ognichenko, A. Hromov, A. Kosinskaya, S. Stelmakh, Z.L. Sessions, and E.N. Muratov, *Simplex representation of molecular structure as universal QSAR/QSPR tool*, Struct. Chem. 32 (2021), pp. 1365-1392.



- [42] S. Shen, Y. Pan, X. Ji, Y. Ni, and J. Jiang, *Prediction of the Auto-Ignition Temperatures of Binary Miscible Liquid Mixtures from Molecular Structures*, *Int. J. Mol. Sci.* 20 (2019), p. 2084.
- [43] J. Yao, R. Qi, Y. Pan, H. He, Y. Fan, J. Jiang, and J. Jiang, *Prediction of the flash points of binary biodiesel mixtures from molecular structures*, *J. Loss Prevent. Proc. Ind.* 65 (2020), p. 104137.
- [44] W. Cao, Y. Pan, Y. Liu, and J. Jiang, *A novel method for predicting the flash points of binary mixtures from molecular structures*, *Safety Science* 126 (2020), p. 104680.
- [45] A.P. Toropova, A.A. Toropov, E. Carnesecchi, E. Benfenati, and J.L. Dorne, *The index of ideality of correlation: models for flammability of binary liquid mixtures*, *Chemical Papers* 74 (2020), pp. 601-609.
- [46] D. Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*, *J. Chem. Inf. Comput. Sci.* 28 (1988), pp. 31-36.
- [47] Z. Jiao, S. Yuan, Z. Zhang, and Q. Wang, *Machine learning prediction of hydrocarbon mixture lower flammability limits using quantitative structure-property relationship models*, *Process Saf. Prog.* 39 (2020), p. e12103.
- [48] L. Jiao, X. Zhang, Y. Qin, X. Wang, and H. Li, *QSPR study on the flash point of organic binary mixtures by using electrotopological state index*, *Chemometr. Intell. Lab.* 156 (2016), pp. 211-216.
- [49] M. Chatterjee, and K. Roy, *Application of cross-validation strategies to avoid overestimation of performance of 2D-QSAR models for the prediction of aquatic toxicity of chemical mixtures*, *SAR QSAR Environ. Res.* 33 (2022), pp. 463-484.
- [50] B. Wang, K. Xu, and Q. Wang, *Prediction of upper flammability limits for fuel mixtures using quantitative structure-property relationship models*, *Chem. Eng. Comm.* 206 (2019), pp. 247-253.
- [51] N. Chirico, and P. Gramatica, *Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient*, *J. Chem. Inf. Model.* 51 (2011), pp. 2320-2335.
- [52] N. Chirico, and P. Gramatica, *Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection*, *J. Chem. Inf. Model.* 52 (2012), pp. 2044-2058.
- [53] J.R. Rowley, R.L. Rowley, and W.V. Wilding, *Estimation of the Flash Point of Pure Organic Chemicals from Structural Contributions*, *Process Saf. Prog.* 29 (2010), pp. 353-358.
- [54] A.C. Atkinson, *Plots, Transformations and regression - An introduction to graphical methods of diagnostic regression analysis*, Oxford Science Publications, 1985.
- [55] A.P. Toropova, A.A. Toropov, M. Marzo, S.E. Escher, J.L. Dorne, N. Georgiadis, and E. Benfenati, *The application of new HARD-descriptor available from the CORAL software to building up NOAEL models*, *Food and Chemical Toxicology* 112 (2018), pp. 544-550.
- [56] B. Wang, H. Park, K. Xu, and Q. Wang, *Prediction of lower flammability limits of blended gases based on quantitative structure-property relationship*, *J. Therm. Anal. Calori.* 132 (2018), pp. 1125-1130.
- [57] S. Shen, X. Ji, Y. Pan, R. Qi, and J. Jiang, *A new method for predicting the upper flammability limits of fuel mixtures*, *J. Loss Prevent. Proc. Ind.* 64 (2020), p. 104074.
- [58] L. Zhou, B. Wang, J. Jiang, G. Reniers, and L. Liu, *A mathematical method for predicting flammability limits of gas mixtures*, *Process Saf. Environ. Protect.* 136 (2020), pp. 280-287.
- [59] Y. Ni, Y. Pan, J. Jiang, Y. Liu, and C.-M. Shu, *Predicting both lower and upper flammability limits for fuel mixtures from molecular structures with same descriptors*, *Process Saf. Environ. Protect.* 155 (2021), pp. 177-183.

**Table 1** - Summary of the compiled studies on the development of models to predict the physical hazards of mixtures

Author (year)	Endpoints	Mixture type	Chemical type	Ref.
Saldana (2013)	flash point	binary and ternary	organic compounds	[24]
Gaudin (2015)	flash point	binary	organic compounds	[25]
Ye (2015)	auto-ignition temperature	binary	organic compounds	[40]
Jiao (2016)	flash point	binary	organic compounds	[48]
Wang (2018)a	lower flammability limit	binary and ternary	organic compounds	[56]
Wang (2018)b	flash point	binary and ternary*	organic compounds	[36]
Fayet (2019)	flash point	binary <sup>#</sup>	organic compounds	[32]
Pan (2019)	lower flammability limit	binary	organic compounds	[38]
Shen (2019)	auto-ignition temperature	binary	organic compounds	[42]
Torabian (2019)	flash point	binary <sup>#</sup>	organic compounds	[33]
Wang (2019)	upper flammability limit	binary and ternary	organic compounds	[50]
Cao (2020)	flash point	binary	organic compounds	[44]
Jiao (2020)a	flash point auto-ignition temperature lower flammability limit upper flammability limit	binary and ternary	organic compounds	[34]
Jiao (2020)b	lower flammability limit	binary	organic compounds	[47]
Jin (2020)	auto-ignition temperature	binary	organic compounds	[39]
Shen (2020)	upper flammability limit	binary and ternary	organic compounds	[57]
Toropova (2020)a	flash point	binary	organic compounds	[45]
Toropova (2020)b	flash point	ternary	organic compounds	[35]
Yao (2020)	flash point	binary	organic compounds	[43]
Zhou (2020)	lower flammability limit upper flammability limit	binary and ternary	organic compounds	[58]
He (2021)	decomposition temperature	binary	Ionic liquids	[31]
Ni (2021)	lower flammability limit upper flammability limit	binary and ternary	organic compounds	[59]
Aljaman (2022)	flash point	up to 5 compounds*	organic compounds	[37]

<sup>#</sup> with test on ternary mixtures

\* including large number of pure compounds not involved in mixtures

**Table 2** - Different mixture descriptor formulas tested for the development of a QSPR model for flash point prediction of binary mixtures [25]

Formulas based directly on the mole fraction of each component	$D_m = x_1d_1 + x_2d_2$
	$D_m =  x_1d_1 - x_2d_2 $
	$D_m = x_1^2d_1 + x_2^2d_2$
	$D_m = \sqrt{x_1}d_1 + \sqrt{x_2}d_2$
	$D_m = (x_1d_1 + x_2d_2)^2$
Formulas based on the difference in mole fractions	$D_m = \sqrt{(x_1d_1)^2 + (x_2d_2)^2}$
	$D_m = (1 - \Delta x)\Delta d$
	$D_m = (1 - \Delta x^2)\Delta d$
Other formulas	$D_m = (1 - \Delta x)^2\Delta d$
	$D_m = (d_1 + d_2)/2$
	$D_m = (d_1 - d_2)^2$
	$D_m =  d_1 - d_2 $

**Table 3** - Performance of the different models developed by Jiao et al. (2020) [47] for the LFL of mixtures

Method	R <sup>2</sup>	
	Training (n=43)	Validation (n=11)
MLR	0.9486	0.9186
k-NN	0.8735	0.8583
RF	0.9973	0.9831
BT	0.9926	0.9523
SVM	0.9900	0.9704

## **Figure captions**

**Figure 1** - Flash point profile of ethanol-octane mixtures [6]

**Figure 2** - Properties and complexity of the mixtures studied in the reviewed publications

**Figure 3** - Main constituents represented in the flash point data used by Jiao et al. (2020) [34]

**Figure 4** - Algorithms used in the reviewed publications

**Figure 5** - Validation methods (left) and types of external validation (right) used in the identified publications (CV: Cross-Validation ; Boot.: Bootstrapping ; Y-Rand.: Y-randomization ; Ext.: External Validation)

**Figure 6** - Predictive power observed in external validation for the models dedicated to the flash point compared to the performances observed by Rowley et al. [53] for pure compounds models on 1096 organic compounds













