



**HAL**  
open science

## How to use QSPR-type approaches to predict properties in the context of Green Chemistry

Guillaume Fayet, Patricia Rotureau

► **To cite this version:**

Guillaume Fayet, Patricia Rotureau. How to use QSPR-type approaches to predict properties in the context of Green Chemistry. *Biofuels, Bioproducts & Biorefining*, 2016, 10 (6), pp.738-752. 10.1002/bbb.1723 . ineris-01863106

**HAL Id: ineris-01863106**

**<https://ineris.hal.science/ineris-01863106>**

Submitted on 28 Aug 2018

**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.

# How to use QSPR type approaches to predict properties in the context of Green Chemistry

Guillaume Fayet\*, Patricia Rotureau

INERIS, Accidental Risk Division, Parc Technologique Alatta, 60550 Verneuil-en-Halatte, France

\*Corresponding author: guillaume.fayet@ineris.fr; tel: +33(0)344618126; fax: +33(0)344556565

## Abstract

Green Chemistry is an active field of chemical research in which ultimate targets are to promote high value, safe and clean products from renewable resources using inherently safer and clean processes. Until recently, main efforts focused on the production of chemicals based on renewable resources by cleaner processes notably by eliminating or substituting solvents.

Quantitative Structure-Property Relationships (QSPR) offer the opportunity to also take into account safety issues (in particular fire and explosion risks) in the early steps of development of chemicals and processes in the context of Green Chemistry. Based on robust methods for their development and validation, these predictive approaches allow accessing the properties of substances based on the only knowledge of their molecular structures, even before their synthesis. QSPR models have been developed for various kinds of properties, including physico-chemical hazards, for diverse families of compounds.

They can be used as virtual screening tools to identify the best candidate among a series of possible compounds (within databases of products) for a target application and even in computer aided molecular design to propose alternative molecular structures, for instance for substitution purposes. So they represent very relevant tools to take into account the physico-chemical hazards of substances together with targeted functional properties from the early stages of R&D projects towards safe-by-design “green” products and processes.

**Keywords:** Quantitative Structure-Property Relationships, virtual screening, *in silico* design, Green Chemistry, safety-by-design.

## Introduction

If the concept of Green Chemistry is not new<sup>1</sup>, it represents currently an important source of motivation for research notably in terms of innovative substances and processes. The main idea that supports this multidisciplinary approach consists in “changing the intrinsic nature of a chemical product or process so that it is inherently of less risk to human health and the environment”<sup>1</sup>. Among the twelve principles of Green Chemistry, the resort to alternative raw materials, products, reagents, solvents and catalysts with satisfactory properties towards target applications and lower concern for environment and human health are promoted.

Thus, the products issued from Green Chemistry researches target various purposes. First, they aim to fit with target functions in applications, like conductivity for electrolytes<sup>2</sup> or critical micelle concentration for surfactants<sup>3</sup>. They must also fit with process specifications, e.g. with acceptable viscosity in the process conditions or with a desired boiling point to stay in liquid phase at a specific temperature. Green Chemistry also promote durability criteria including processes economics and environmental costs which may favor the use of renewable resources as raw materials, including biomass residues, etc. Another axis of progress addresses the use of more energetically efficient production routes<sup>4</sup>. Finally, all proposed solutions may stay economically viable to reach industrial applications<sup>5</sup>. Moreover, green chemistry may keep concerned on regulatory compliance and on hazards of substances produced or involved in the production processes. Up to now, many researches addressed the substitution or the elimination of toxic products and in particular solvents<sup>6</sup>.<sup>7</sup>. But, physico-chemical hazards (flammability, explosibility, oxidizing properties) and regulatory compliance are in general considered in the last steps of development, to qualify the final solutions that have been selected to fit with the functional and economic requirements.

Such a strategy can reveal risky and can lead to important waste of R&D time and money ending to the development of a solution with remaining critical safety issues, that could have been identified and solved in the earliest steps of development with a better approach. In pharmaceutical industry for example, the introduction of computational screening approach allowed to take into account the maximum of decision making criteria at the earliest stage of drug design to reduce the proportion of drugs that failed in the (costly) clinical testing<sup>8,9</sup>. Another example was exposed by Wolf et al.<sup>10</sup> in the early 1990s for the substitution of chlorinated solvents, due to their contribution to the reduction of stratospheric ozone or their suspected carcinogenic potential. Different alternative solvents were proposed and adopted but revealed posing other health and environmental safety problems like flammability, chronic toxicity or polluting processes. Therefore, it is advisable that safety issues may be taken into account as early as possible in the R&D decision process to identify high potential candidates regarding desired functional properties while eliminating among them potentially hazardous molecules in a single stage.

Examples of coupled experimental/theoretical approaches involved in the development of green solvents already exist. For instance, Aparicio et al.<sup>11</sup> used different theoretical approaches to complement the experimental characterization of the thermo-physical properties (used in process engineering) of ethyl lactate, which has recently been proposed as a green solvent. The Density Functional Theory (DFT) was used to elucidate the 3D structures and organization of the ethyl lactate molecules (notably in the liquid phase), to compute infrared spectra and to investigate further intra- and inter-molecular hydrogen bonding based on Atoms in Molecules (AIM) and Natural Bond Orbitals

(NBO) studies. Equations of States, Molecular Dynamics and Monte Carlo simulations were also used to calculate thermophysical properties and phase equilibria behaviors. More details on these methods are available in dedicated references<sup>12,13,14</sup>. If they provide reliable predictions for some properties, they are not applicable for the whole range of properties like hazardous properties<sup>14</sup>.

Among existing computational approaches to estimate the properties of chemical products, Quantitative Structure-Activity/Property Relationships (QSAR/QSPR) represent a very active and productive field of research nowadays. Indeed, the number of research articles in this field increased from 164 to 914 per years between 1993 and 2012<sup>15</sup> and extended in application fields from the drug design<sup>16</sup> and toxicology<sup>17</sup> to environmental sciences<sup>18</sup>, physico-chemical properties<sup>19</sup> and safety<sup>14, 20, 21</sup>. In drug design<sup>16</sup>, QSAR/QSPR models are considered as low-cost tools to estimate physico-chemical properties and biological activities for the selection and for the optimization of promising chemical structures.

In that context, this contribution highlights how the use of QSPR models can help to take into account the physico-chemical hazardous properties (e.g. flammability, explosibility) or process safety important parameters (like thermal stability) of chemicals at early steps of development of “green” chemicals or processes. At first, the principles of QSPR models are presented with some guidelines towards a relevant use of QSPR models for the prediction of properties. An overview of existing models to predict physico-chemical hazardous properties is then proposed. Finally, different strategies are described to use QSPR models for the development of “green” chemicals and processes.

### **The Quantitative Structure-Property Relationships Approach**

QSPR models are predictive molecular based methods that allow the prediction of measurable macroscopic properties  $P$  of chemicals from the only knowledge of their molecular structure. So they can be used even before synthesis of chemicals. Such models are developed according to a similarity principle, by considering that compounds with similar molecular structures will have similar properties. It consists in looking for correlation between the target property and a series of descriptors  $d_i$  of the molecular structures of a dataset of compounds similar to those targeted by the final model (as illustrated in Figure 1Figure 1). QSPR models can be summarized by the general form in Eq. 1.

$$P = f(d_i) \tag{1}$$

Molecular descriptors characterize the molecular structure of the chemical compounds and their molecular scale properties. Among years, thousands of descriptors have been developed to encode the whole diversity of chemicals and their molecular properties<sup>22, 23</sup>. They can be of different types either obtained from simple elemental formula, from the 2-dimension structure or from 3-dimension structure, requiring preliminary structure determination, in general by quantum chemical calculations. These molecular descriptors can be classified into four classes.

- Constitutional descriptors identify and count particular features in the molecule (atoms, bonds, fragments, functional groups).

- Topological descriptors encode the molecular structure of compounds from molecular graphs that characterize the connection of atoms in the molecule. They relate to the size, shape and ramification rate of molecular structures.
- Geometric descriptors characterize the 3D structure of molecules. They consist in interatomic distances, angles, dihedral angles, molecular volume and surface areas.
- Quantum chemical descriptors gather the information obtained from Quantum chemical calculations about electrostatic properties (e.g. atomic charges), reactivity (e.g. bond dissociation energies) or molecular orbitals (e.g. molecular orbital energies).

The final QSPR model is established using various data mining tools to fit the model, select the best set of descriptors, estimate its performances and define its applicability domain. Models can be issued from simple multi linear regression (MLR), as in Eq. 2, or from more complex methods like artificial neural networks<sup>24</sup> or support vector machine<sup>25</sup> that sometimes improve the performances of the models but, in general, with a loss of chemical interpretation.

$$P = a_0 + \sum a_i d_i \quad (2)$$

where  $a_i$  are the regression constants.

Qualitative predictive approaches can also be derived through decision trees or based on principal component analyses that are useful in different situations. At first, they are used in case of intrinsically qualitative properties like the determination of the order of elution (first or second) of chiral compounds<sup>26</sup>. They can also be used for quantitative properties notably when large uncertainties are expected and for applications for which a first “High/low” or “Yes/No” estimation could be sufficient.

The selection of the descriptors included in the model is an important task to avoid a risk of any over-parameterization of the model that would lead to lower its predictive power. Several data mining strategies can be used<sup>27</sup>. For instance, MLR models are often derived based on stepwise methods, by gradually adding or cancelling descriptors; genetic algorithm, that mimics the natural evolutionary phenomena on applying randomized modifications of the set of descriptors (additions, cancellations, substitutions), is also currently used to evidence the best set of descriptors<sup>28</sup>. The final model is chosen to represent the best compromise between the quality of fit of the model and the chemical meaning of the including descriptors.

At last, one critical parameter to access an accurate QSPR model is the experimental data set used for its development and validation<sup>29</sup>. At first, the database must be as large as possible to allow a robust fitting of the model and to keep aside an external validation set to evaluate the predictive power of the final model. Moreover, these data must be as reliable and homogeneous as possible. Indeed, uncertainties and errors in the experimental training data propagate in the model and affect its accuracy. Furthermore, experimental protocol must be at best clarified to avoid contradictory data that can be due to the use of data issued from different protocols.

To evaluate the performances of models, several internal and external validation methods are used<sup>30, 31</sup>. The cross validation techniques are internal validation methods that consist in excluding part of the training set, refitting the model and check that the refitted model keeps reliable for the excluded compound(s). They characterize the robustness of the model, i.e. that the model is not too

dependent on particular molecules in the training set. Another internal validation is Y-randomization<sup>32</sup> that aims to ensure that the model was not issued from a chance correlation. It consists in randomizing the property data among molecules in the training set and to check that models refitted on these erroneous data give erroneous predictions regarding actual experimental property values. At last, the predictive capabilities of the model are checked on an external validation set of molecules not used to develop it, by comparing the predicted data to the experimental ones<sup>33-35</sup>.

At last, the applicability domain (AD) of the model has to be defined, i.e. the domain in which predictions can be expected to be accurate. As QSPR is an interpolation approach, this AD is limited by the training set with which the model has been developed, in terms of property domain and chemical space. The definition of AD can be done using different kinds of tools as reviewed by Jaworska et al.<sup>36</sup> or Eriksson et al.<sup>37</sup>. One of the simplest ways to proceed is to define the ranges of values of property and descriptors (for those included into the model) in the training set.

Once a model is validated, it can be used for prediction. But to ensure a correct and relevant application of the model, some precautions must be taken. As explained before, a model is only applicable in a defined applicability domain in terms of chemical diversity (within identified families and within a defined domain in the chemical space described by the values of descriptors in the training set) and in terms of property domain (by the range of property values in the training set). To ensure it, the following questions may be answered:

- Question 1: Is the endpoint of the model relevant for the use of the expected prediction (e.g. protocol in agreement with regulation, property predicted at the same temperature than the one expected in a process)?
- Question 2: Is the model dedicated to the family of molecules to which the target compound belongs?
- Question 3: Is the compound in the applicability domain of the model from the used molecular descriptors view point?
- Question 4: Is the final calculated property in the applicability domain of the model?

Moreover, the model must be applied following the exact procedure defined by the developers of the model. In particular, in the case of quantum chemical descriptors, computational details, like the basis set and the method (e.g. functional), must be strictly followed.

### **Existing Models for Hazardous Physico-chemical Properties**

QSPR models have been developed and reviewed among years for very diverse properties of chemicals<sup>19</sup> (notably related to the REACH regulation<sup>14, 20</sup>), of materials<sup>38</sup> and also for hazardous substances<sup>14, 20, 21</sup>. Table 1 summarizes the capabilities of QSPR models available to predict physico-chemical properties used to classify substances according to the European regulation related to the Classification, Labelling and Packaging of substances and mixtures<sup>39</sup>. This table indicates that QSPR models nowadays concern only one part of the properties required to classify chemicals according to physical hazards. So, new models would deserve to be developed. In particular, no QSPR model exists for oxidizing properties, probably due to the lack of available experimental data to develop models.

As a matter of fact, a number of different models dedicated to physico-chemical hazards have indeed been proposed in the last decades, for different properties and different families of compounds. For example, extended works were devoted to nitro compounds. Concerning their heat of decomposition ( $\Delta H$ ), an accurate model<sup>40</sup> has been obtained for nitrobenzene derivatives presenting no substituent in ortho position to the nitro group (as illustrated in Figure 2), taking into account the fact that such last compounds could present specific decomposition mechanisms as demonstrated on ortho-nitrotoluenes by a Density Functional Theory study<sup>41</sup>.

This model consists in a four-parameter equation (Eq. 3) with a correlation coefficient  $R^2$  of 0.90 and an average deviation of 12% (for the 31 molecules of the training set).

$$-\Delta H \text{ (kcal/mol)} = 0.8 G - 3.8 WPSA1 - 4255.1 Q_{max} + 26.8 RPCS - 251.2 \quad (3)$$

with  $G$  the gravitational index,  $WPSA1$  the weighted positive surface area,  $Q_{max}$  the maximal partial charge and  $RPCS$  the relative positively charged surface area.

The predictive power of the model was estimated on a validation set of 11 molecules not already used for the development of the model with a  $R^2$  of 0.84 and an average error of 18%. No such performance was reached by including ortho substituted nitro derivatives, demonstrating the importance of understanding the molecular mechanisms involved into the target properties.

If the model presented in Eq. 3 requires quantum chemical calculations, other models were developed with simpler descriptors. For instance, QSPR models dedicated to the impact sensitivity of nitramines<sup>42</sup> were developed using only constitutional descriptors that can be calculated from 2D molecular structures (Eq. 4).

$$\log h_{50\%} = 0.94 + 86.3 n_{C=O}/Mw - 0.017 OB + 0.14 n_{C-O-C} - 0.21 n_{C=O} \quad (4)$$

where  $n_{C=O}/Mw$  is the ratio of the number of C=O fragments on the molecular weight,  $OB$  is the oxygen balance<sup>43</sup>,  $n_{C-O-C}$  and  $n_{C=O}$  are the numbers of C-O-C and C=O fragments, respectively.

This model presented even similar results than a quantum chemical based model obtained in a previous work<sup>44</sup> with a  $R^2_{ext}$  of 0.90 (vs. 0.88 for the quantum chemical model) and a  $RMSE_{ext}$  of 0.14 (vs. 0.16) estimated on an external validation set (see Figure 3) in which only one molecule revealed out of the applicability domain. Such simple model can reveal less accurate and meaningful than a quantum chemical one in some cases but it presents the advantage to be faster and easier to implement and to apply for non-expert users.

Regarding the hazards of nitro compounds, we also proposed a decision tree model to predict if the heat of decomposition of a nitroaromatic compound is higher or lower than 500 kJ/mol<sup>45</sup>. The final algorithm, presented in Figure 4 is very simple and presented good performances with 82 % of good classifications as evaluated on an external validation set.

QSPR models for hazardous properties were also developed for other families of compounds like the heat and temperature of decomposition of organic peroxides<sup>46</sup>, as shown in Eqs. 5 and 6.

$$-\Delta H/C \text{ (kJ/g)} = 54^1 K - 990 n_{OO} + 12934 d_{OO} + 2631 Q_{OO} - 19371 \quad (5)$$

$$T_{onset} = 144 F_{OO} + 29 n_{OO} - 20 gap + 194 \quad (6)$$

Where  $C$  is the concentration of organic peroxide,  ${}^1K$  is the order 1 Kier shape index,  $n_{OO}$  is the number of peroxide bonds,  $d_{OO}$  is the distance between the oxygen atoms of the peroxide bond,  $Q_{OO}$  and  $F_{OO}$  are the average Mulliken charges and the average local Fukui function on these two O atoms and  $gap$  is the energy difference between the LUMO and HOMO orbitals.

These models presented good predictive powers as estimated on external validation sets with  $R^2_{in}$  of 0.81 and 0.83, respectively in their applicability domains. Moreover, they include descriptors directly related to the presence and strength of the peroxide bond which is recognized to be initiated by its homolytic cleavage<sup>47, 48</sup>.

Recent models were also developed for safety related parameters like complete heats of combustion of ionic liquids (IL)<sup>49</sup>. Their gross heat of combustion (HHV) can be obtained from the knowledge of their stoichiometric composition in C, H, O, N and Cl atoms with an error of 3.9 % ( $R^2_{ext}=0.98$ ).

$$HHV = 34.95 C + 135.82 H - 3.37 O + 6.45 N + 2.79 Cl - 1.86 \quad (7)$$

Eventually, the most recent developments allow predictions for the properties of mixtures, in particular for the flash points of organic liquid mixtures. A first approach consisted in including QSPR models to provide predictions on pure compounds in existing mixing rules. Such approach was successfully proposed by Saldana<sup>50</sup> and Gaudin<sup>51</sup> with good predictive performances (with errors about 5 K on binary mixtures for both works). For instance, the full-predictive approach of Gaudin et al. predicts the profile of flash point of the octane/isopropanol mixture with a mean absolute error of only 2.5°C (Figure 5). It is interesting to note that this mixture is more hazardous than the pure compounds with lower value of flash points. In such a case, the predicted profiles allow overcoming the sometimes misleading practice consisting in taking into consideration the flash point of the most flammable constituent of a mixture in absence of data on the mixture.

The full-predictive approach of Gaudin reveals also reliable when applied to ternary mixtures. For example, the predicted flash points for methanol/toluene/2,2,4-trimethylpentane mixtures (used to build the ternary diagram presented in Figure 6) are close to experimental ones obtained by Liaw et al.<sup>52</sup>, with a mean absolute error of only 0.5°C (on 63 data). Here again, a mixture of 50 % of methanol and 50 % of 2,2,4-trimethylpentane reveals having a lower flash point (calculated at -12.5°C) than the pure compounds with experimental FP<sup>52</sup> of 10.0°C, 7.2°C and -8.1°C for methanol, toluene and 2,2,4-trimethylpentane, respectively. So, these two examples encourage the use of such predictive approaches to anticipate the possible increase of hazards when mixing chemicals.

Another strategy was also recently developed by Gaudin et al.<sup>53</sup> consisting in defining mixture descriptors by applying mixing formula on molecular descriptors to achieve mixture QSPR models. Such approach can be particularly useful when no mixing rule is (easily) available. First promising results were obtained for 284 flash points of binary liquid mixtures with a model (Eq. 8) presenting a mean absolute error of 10.3°C (evaluated on an external validation set of 151 data).

$$FP = 50.3 + 16.3 (x_1 {}^3\chi_1 + x_2 {}^3\chi_2)^2 + 5.5 \times 10^{-3} (x_1 HDCA_1 + x_2 HDCA_2)^2 \\ - 2.4 \times 10^{-6} (x_1 \Delta\alpha_1 + x_2 \Delta\alpha_2)^2 - 88.0 (x_1 V_{min,H,1} + x_2 V_{min,H,2})^2 \quad (8)$$

where  $x$  is the molar fraction,  ${}^3\chi$  is the Randic index (order 3),  $HDCA$  is the HDCA H-donors charged surface area (order 2),  $\Delta\alpha$  is the anisotropic polarizability and  $V_{min,H}$  is the minimum valency of a H atom.

To complement this overview of existing QSPR models for physico-chemical hazards, it is worth noting that QSAR models have been developed for other properties of high concern in Green Chemistry like biodegradability<sup>18, 54</sup> and (eco)-toxicity<sup>17, 55, 56</sup> that must also be considered in view of developing/selecting safer and cleaner products, for instance for substitution purpose.

## **Proposed uses of QSPR models in the context of Green Chemistry**

### ***QSPR models as screening tools***

QSPR models can be used early in the development of chemicals or processes as virtual screening tools to estimate the properties of substances when not already available, purchased or even before their synthesis. These predictions can help to select a series of candidates that can be in a following step further studied by performing synthesis and experimental characterizations.

Using QSPR models in such a way is very useful to reduce the possible candidates among large sets of potential chemicals, as illustrated in Figure 7 and is a common computer aided molecular design procedure used in pharmaceutical research to discover new drugs<sup>57</sup> according to desired activity, for instance for anti-cancer agents<sup>58</sup>. The dataset of possible chemical structures can be gathered from supplier's chemicals portfolio datasheets or from internal databases of compounds. For instance, pharmaceutical companies developed their own compound libraries that, from years, compile the structures and/or data for millions of compounds<sup>59</sup>. Commercial and open-source databases have been also widely developed in this field<sup>60</sup>. Similar strategies of data collection could be considered for biobased compounds in the future.

In the context of Green Chemistry, virtual screening can also be used to select safe candidates holding the target functionalities among databases of compounds and to select safer solvents, catalysts or reagents in processes. It can also assist the formulation of mixtures not only for the choice of constituents in mixtures but also by computing property profiles versus concentrations (even if only few examples of QSPR models for mixtures nowadays exist<sup>61</sup>) allowing to determine the concentrations that offer the best functional properties with the lowest hazards.

The first step of such approach is to define the properties required to satisfy the expected specifications of the final product. These specifications can be related to:

- application properties, like critical micelle concentration for surfactants or the ionic conductivity for electrolytes;
- process properties, like boiling point, melting point or maximum critical temperature in agreement with specific conditions of process;
- hazardous properties, like toxicity, flammability and explosivity;
- economic and environmental costs;
- regulatory compliance, based on properties required to classify and register substances according to regulations like the Transport of Dangerous Goods (TDG)<sup>43</sup> or the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)<sup>62</sup>.

Of course, some specifications are not related to the chemical structures of products, in particular economic costs, and QSPR models cannot help to evaluate these criteria. For other specifications (in particular application, process and hazardous properties), existing QSPR models can be used to at least obtain an estimation of the properties to screen the chemical structures of the studied database and identify the best candidates for the target application.

When used for screening purpose, QSPR models can have different levels of predictive powers depending on the property considered and the expected level of screening. Indeed, first screening can be done with models that are very quick and easy to use in order to eliminate the largest part of non relevant alternatives. Then, additional screening steps can be performed on the most promising candidates using more time-consuming and/or less automated models. Of course, other computational complementary approaches can also be introduced in such process, for instance by using UNIFAC or equations of states that are commonly used to predict the fluid properties<sup>14</sup> needed for process engineering.

Some examples of virtual screening for drug design are notably provided by Tropsha et al.<sup>63</sup> for the discovery of anticonvulsant compounds and anticancer agents. For anticancer agents, Zhang et al.<sup>64</sup> started by developing QSAR models based on an experimental dataset of half maximal effective concentrations ( $EC_{50}$ ) for 52 phenanthrine-based tylophorine derivatives (PBT). Using topological descriptors and a  $k$  nearest neighbours ( $kNN$ ) approach for the descriptor selection, 10 models were selected from multiple chemically different partitions of the dataset into training and validation sets. These models (constituted from 10 to 20 descriptors) presented performances from 0.71 to 0.81 in  $R^2$ . They were then applied on the molecular structures available in the ChemDiv Database<sup>65</sup>, a commercial database of about 500 000 compounds. 34 compounds were identified as relevant candidates from a consensus prediction of moderate or high biological activity based on the ten developed models. 10 out of those 34 compounds were experimentally investigated and eight of them confirmed their activity against the target cancer cell. The best activity was found for the structure presented in Figure 8, with an activity in line with the predicted one (with  $(-\log EC_{50})_{exp} = 5.74$  vs.  $(-\log EC_{50})_{pred} = 5.21$ ).

These examples demonstrate the potential of QSPR models to evidence the best candidates for a target application and could also be used to take into account physical hazards as relevant models already exist. By considering both hazardous and functional properties in early stages of selection of chemical candidates, both (toxic, physico-chemical) hazard and functional property unnecessary testings are avoided for compounds that are already identified as not relevant for the target application or exhibiting too important hazards for workers or end-users.

### ***QSPR models as in silico design tools***

Based on QSPR models, full *in silico* design of substances is possible (notably for substitution purposes) by not only selecting existing molecules but also proposing new ones with particular specifications. In such approaches, QSPR models can not only give predictions but much more they can also give structural trends towards (better) target properties and lower hazards.

For such an inverse exercise, a current strategy (originated also from drug design<sup>16, 58</sup>) is to generate a virtual combinatorial library of structures in a defined chemical space and to screen it using computational methods to predict the target specifications. Then, the most relevant compounds are

further investigated for the final application, as shown in Figure 9. Such an approach is notably used in drug design to optimize the molecular structure of a compound that has been identified in the virtual screening of databases of existing chemical products. In such cases, analog structures are generated by applying structural modifications on the molecular scaffold of the selected structure<sup>16</sup>.

As detailed in the previous section, the target specifications can be defined not only in terms of functional properties but also in terms of hazards. In the case of substitution studies, the properties of the compound to be substituted give in general the main expectations both for the required properties and the undesired hazards that represent the motivation for searching an alternative substance.

Then, a virtual library of chemical structures is generated within a defined chemical space using different possible methods. A first one is based on the definition of a series of building blocks and structural constraints. Such approach is notably proposed by Gani et al.<sup>66</sup> who considered as building blocks the contribution groups of the predictive methods used for the prediction of some functional properties (e.g. octanol/water partition coefficient, boiling point, surface tension). This approach has been used for various applications like solvents for crystallization<sup>67</sup>, polymers<sup>68</sup> or surfactants<sup>69</sup>. Within the same type of approach, Weis and Visco<sup>70</sup> used Signature molecular descriptors both as descriptors in the QSPR models and as building blocks in the structure generation in a solvent selection application. In the framework of Green Chemistry, the use of bio-based raw materials favors this approach by the definition of a series of bio-based building blocks for the generation of virtual combinatorial libraries. Moity et al.<sup>71</sup> notably applied a similar approach in a tool called GRASS that associates bio-building blocks with co-reactant to generate new molecules with an application on itaconic acid based solvents. Similarly, Heintz et al.<sup>72</sup> proposed computer aided tools (IBSS, for Integrated Bio Sourced Search) for the development of sustainable products, including mixtures, by the combination of groups (chemical functional groups or bio-sourced synthons) into molecules and, then, into mixtures, using a genetic algorithm.

A second interesting way to build virtual libraries consists in generating chemical structures by virtual reactions from selected initial molecules of interest. Indeed, an important field of work in Green Chemistry relates to so-called platform molecules, e.g. succinic acid<sup>73</sup> (Figure 10) or lactic acid<sup>74</sup> (Figure 11), which can be derived from renewable resources and from which a huge variety of chemicals can be produced. So, virtual chemical libraries can be defined from such particular platform molecules.

Once the virtual library is generated, it is screened according to the target properties using theoretical approaches, like group contribution, as proposed by Gani et al.<sup>66</sup>, or COSMO-RS, used by Moity et al.<sup>71</sup>, to select best eligible candidates which can be then confirmed by experimental means. The same virtual screening can also be supported by QSPR models. For instance, Rucker et al.<sup>75</sup> proposed such inverse-QSPR study for the identification of C<sub>1</sub>-C<sub>4</sub> acyclic haloalkanes containing C, H, F, Cl and Br presenting boiling points (T<sub>b</sub>) between 130°C and 140°C. The boiling points of 507 C<sub>1</sub>-C<sub>4</sub> haloalkanes were gathered from literature and used to develop a 7-descriptor MLR model with high quality of fit (R<sup>2</sup>=0.99) and good robustness (R<sup>2</sup><sub>cv</sub>=0.98) using the MOLGEN-QSPR program<sup>76</sup>.

$$T_b = -153.251 n_{F,rel} + 73.1663 n_{Br,rel} + 53.3144 {}^1\chi^s + 100.227 SCA1 - 16.7507 slogP - 0.828538 {}^2TC^v + 1.12749 {}^4TC_c - 223.678 \quad (9)$$

where  $n_{F,rel}$  and  $n_{Br,rel}$  are the relative numbers of F and Br atoms,  ${}^1\chi^s$ ,  $SCA1$ ,  ${}^2TC'$  and  ${}^4TC_c$  are topological descriptors (Kier and Hall valence chi indices of first order, sum of coefficients of principal eigenvector of the adjacency matrix and two Bonchev's overall topological indices) and  $slogP$  is a calculated partition coefficient.

Then, 28 600 compounds were generated by exhaustive and redundancy-free construction in the target chemical space using the MOLGEN software<sup>76</sup>. The model in Eq. 9 was applied on these compounds and 655 of them were found to fall between 130°C and 140°C in boiling point.

Another interesting example concerns ionic liquids (IL) which are particularly relevant systems for such approaches since extended virtual combinatorial libraries can be built by combining diverse anions and cations. In their study, Matsuda et al.<sup>77</sup> address the design of IL with target ionic conductivity. At first, a non-linear QSPR model was developed based on group contributions for the different cations, chain lengths ( $R_1$ ), other side chains ( $R_{2-4}$ ) and anions (as illustrated in Figure 12) in the training set. This model developed from 206 ionic conductivities presented a correlation coefficient  $R^2$  of 0.91.

Then, a reverse design of IL was performed to evidence IL presenting an ionic conductivity of  $15 \pm 1 \text{ mS.cm}^{-1}$  at 40°C. All structures that can be built by combining several cations, chain lengths, other side chains and anions were exhaustively generated. The ionic conductivities of the built structures were calculated using the developed QSPR model and the compounds highlighting values of properties falling into the target range of ionic conductivity (i.e. between 14 and 16  $\text{mS.cm}^{-1}$ ) were identified. Finally, 13 generated IL were exhibited as possible relevant candidates.

To go further in the inverse QSPR problem, one can take advantage of the trends involved into the QSPR models to evidence the best chemical candidates for target applications and then guide the generation of virtual chemical structures. Indeed, once a model has been developed, it is sometimes possible to identify how descriptor values influence the final properties. Then, one can change the molecular structure in a way that descriptors reach acceptable values to reach the expected property. To do that, some types of models are more relevant than others. Indeed, it is easier to investigate the influence of a descriptor in a multi linear model than in an artificial neural network.

Such type of approach has been automated by Miyao et al.<sup>78</sup> for the exhaustive generation of all chemical structures satisfying a target property. The efficiency of this approach was demonstrated on the search of chemicals with desired boiling points. At first, a QSPR model was developed on a dataset of 600 acyclic hydrocarbons (with  $R^2=0.95$ ) that was validated on an external set of 282 compounds (with  $R^2_{ext}=0.95$ ).

$$T_b = -217.953 + 37.622 \text{ nSK} - 0.943 \text{ nDB} + 30.870 \text{ MWC02} + 249.128 \text{ MWC03} \\ - 185.337 \text{ MWC04} - 34.971 \text{ X1} \quad (10)$$

where nSK is the number of main atoms, nDB is the number of double bonds, MWC02, MWC03 and MWC04 are the molecular walk counts with order 2, 3 and 4 and X1 is the Randic connectivity index.

Then, probability distributions of descriptors were defined as a function of the target property and mixed to define a chemical space in which all chemical structures presenting the target property will be found. Then, all chemical structures in this chemical space were systematically generated.

If this kind of strategies nowadays are used focusing on the identification of high performance compounds, the availability of reliable QSPR models for physico-chemical hazards open new perspectives towards the search for safer compounds. In particular, in the field of Green Chemistry, all requirements towards more “green” chemicals could be taken into account from the beginning of the development process, including solubility in water, biodegradability for instance.

### **Conclusions and perspectives**

Computational approaches have increased their implications in R&D since now many decades in different fields. Nowadays, molecular scale computations have proven their relevance and performances to support innovation. Notably, computer aided molecular design has been already extensively used in drug design to reduce costs by using virtual screening to identify the products of greatest potential.

Green Chemistry is an active field of chemical research in which these computational methods could allow reducing experimental times and costs by anticipating in the earliest steps of development not only the functional properties but also the hazards of final products, reactants or solvents used in processes. In particular, Green Chemistry can nowadays take advantage of the increasing developments of QSPR models to predict physico-chemical hazards of chemicals from the only knowledge of their molecular structures. In addition to the prediction of a single property value for instance for the classification of hazardous chemicals or in process safety analyzes, these predictive methods are powerful alternative approaches for virtual screening or even for the *in silico* design of new products. This could also be complemented with modeling tools like process engineering software to go further into an even more global *in silico* R&D strategy.

To exploit the full potential of the QSPR approach in Green Chemistry, further research actions may still be encouraged:

- The fields of applications of existing QSPR models can be enlarged to fulfill all specifications of chemicals in the context of Green Chemistry. Considering safety issues, QSPR models for all chemical hazards can be further developed and used, notably for properties for which no model already exists like oxidizing properties.
- Further development and evaluation of models for specific compounds with promising applications in Green Chemistry like ionic liquids, electrolytes or surfactants should be encouraged. Indeed, some of these systems, like ionic liquids, are particularly relevant for computer aided molecular design by targeting the combination of anions and cations towards specific properties using QSPR models.
- The prediction of mixture properties remains also a challenge of great importance when looking for bio-based products since their compositions can be complex and variable in time, dependent on bio-resources. Nevertheless, first models revealed already promising for the flash point of binary liquid mixtures and encourage further development to improve the performances of models and to use them for complex multi-component mixtures.
- To support the development, evaluation and improvement of QSPR models, robust and large databases on bio-based compounds or class of compounds, e.g. solvents, surfactants or ionic liquids, could be collected and organized. Once consolidated, QSPR models can be used to screen these databases to identify high potential candidates for industrial applications.

## Acknowledgment

The authors thank Guy Marlair for fruitful discussions.

## References

1. Anastas PT and Warner JC, *Green Chemistry: Theory and Practice*. Oxford University Press, (1998).
2. Chang Z, Yang Y, Li M, Wang X and Wu Y, Green energy storage chemistries based on neutral aqueous electrolytes. *J Mater Chem A* **2**:10739-10755 (2014).
3. Foley PM, Phimpachanh A, Beach ES, Zimmerman JB and Anastas PT, Linear and cyclic C-glycosides as surfactants. *Green Chem* **13**:321-325 (2011).
4. Anastas PT and Kirchhoff MM, Origins, Current Status, and Future Challenges of Green Chemistry *Acc Chem Res* **35**:686-694 (2002).
5. Hjeresen DL, Kirchhoff MM and Lankey RL, Green Chemistry: Environment, Economics, and Competitiveness. *Corpor Env Strategy* **9**:259-266 (2002).
6. Clark JH and Tavener SJ, Alternative Solvents: Shades of Green. *Org Process Res Dev* **11**:149-155 (2007).
7. Capello C, Fischer U and Hungerbuhler K, What is a green solvent? A comprehensive framework for the environmental assessment of solvents. *Green Chem* **9**:927-934 (2007).
8. Hay M, Thomas DW, Craighead JL, Economides C and Rosenthal J, Clinical development success rates for investigational drugs. *Nat Biotech* **32**:40-51 (2014).
9. DiMasi JA, Success rates for new drugs entering clinical testing in the United States. *Clinic Pharm Therap* **58**:1-14 (1995).
10. Wolf K, Yazdani A and Yates P, Chlorinated Solvents: Will the Alternatives be Safer? *J Air Waste Manag Assoc* **41**:1055-1061 (1991).
11. Aparicio S and Alcalde R, The green solvent ethyl lactate: an experimental and theoretical characterization. *Green Chem* **11**:65-78 (2009).
12. Cramer CJ, *Essentials of Computational Chemistry - Theories and Models*. Wiley, Chichester, U.K., (2004).
13. Young DC, *Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems*. John Wiley & Sons Inc., New York, (2001).
14. Nieto-Draghi C, Fayet G, Creton B, Rozanska X, Rotureau P, De Hemptinne J-C, Ungerer P, Rousseau B and Adamo C, A General Guidebook for the Theoretical Prediction of Physico-Chemical Properties of Chemicals for Regulatory Purposes. *Chem Rev* **115**:13093-13164 (2015).
15. Li L, Hu J and Ho Y-S, Global Performance and Trend of QSAR/QSPR Research: A Bibliometric Analysis. *Mol Inform* **33**:655-668 (2014).
16. Bajot F, The Use of Qsar and Computational Methods in Drug Design. In *Recent Advances in QSAR Studies*, Puzyn, T.; Leszczynski, J.; Cronin, M. T., Eds. Springer Netherlands, Vol. 8, pp 261-282 (2010).
17. Schultz TW, Cronin MTD, Walker JD and Aptula AO, Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. *J Mol Struct (Theochem)* **622**:1-22 (2003).
18. Pavan M and Worth AP, Review of Estimation Models for Biodegradation. *QSAR Comb Sci* **27**:32-40 (2008).
19. Katritzky AR, Kuanar M, Slavov S, Hall CD, Karelson M, Kahn I and Dobchev DA, Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem Rev* **110**:5714-5789 (2010).
20. Dearden JC, Rotureau P and Fayet G, QSPR prediction of physico-chemical properties for REACH. *SAR QSAR Environ Res* **24**:545-584 (2013).

21. Quintero FA, Patel SJ, Munoz F and Mannan MS, Review of Existing QSAR/QSPR Models Developed for Properties Used in Hazardous Chemicals Classification System. *Ind Eng Chem Res* **51**:16101-16115 (2012).
22. Karelson M, *Molecular Descriptors in QSAR/QSPR*. Wiley, New York, (2000).
23. Todeschini R and Consonni V, *Handbook of Molecular Descriptors*. Wiley, Weinheim, (2000).
24. Gasteiger J and Zupan J, Neural Networks in Chemistry. *Angew Chemie Int Ed* **32**:503-527 (1993).
25. Ivanciuc O, Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*, John Wiley & Sons, Inc., pp 291-400 (2007).
26. Del Rio A and Gasteiger J, Encoding Absolute Configurations with Chiral Enantiophore Descriptors. Application to the Order of Elution of Enantiomers in Liquid Chromatography. *QSAR Comb Sci* **27**:1326-1336 (2008).
27. Shahlaei M, Descriptor Selection Methods in Quantitative Structure-Activity Relationship Studies: A Review Study. *Chem Rev* **113**:8093-8103 (2013).
28. Valadi J, Siarry P, Sukumar N, Prabhu G and Saha P, Applications of Genetic Algorithms in QSAR/QSPR Modeling. In *Applications of Metaheuristics in Process Engineering*, Springer International Publishing, pp 315-324 (2014).
29. Dearden JC, Cronin MTD and Kaiser KLE, How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ Res* **20**:241-266 (2009).
30. Tropsha A, Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform* **29**:476-488 (2010).
31. Gramatica P, Principles of QSAR models validation: internal and external. *QSAR Comb Sci* **26**:694-701 (2007).
32. Rücker C, Rücker G and Meringer M, Y-Randomization and Its Variants in QSPR/QSAR. *J Chem Inf Model* **47**:2345-2357 (2007).
33. Chirico N and Gramatica P, 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**:2320-2335 (2011).
34. Chirico N and Gramatica P, 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**:2044-2058 (2012).
35. Consonni V, Ballabio D and Todeschini R, Evaluation of model predictive ability by external validation techniques. *J Chemometr* **24**:194-201 (2010).
36. Jaworska J, Nikolova-Jeliazkova N and Aldenberg T, QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Altern Lab Anim* **33**:445-459 (2005).
37. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM and Gramatica P, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Persp* **111**:1361-1375 (2003).
38. Le T, Epa VC, Burden FR and Winkler DA, Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem Rev* **112**:2889-2919 (2012).
39. Regulation (EC) N°1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) N° 1907/2006.
40. Fayet G, Rotureau P, Joubert L and Adamo C, Development of a QSPR model for predicting thermal stabilities of nitroaromatic compounds taking into account their decomposition mechanisms. *J Mol Model* **17**:2443-2453 (2011).
41. Fayet G, Joubert L, Rotureau P and Adamo C, A theoretical study of the decomposition mechanisms on substituted ortho-nitrotoluenes. *J Phys Chem A* **113**:13621-13627 (2009).
42. Fayet G and Rotureau P, Development of simple QSPR models for the impact sensitivity of nitramines. *J Loss Prevent Proc Ind* **30**:1-8 (2014).
43. UN, Recommendations on the Transport of Dangerous Goods: Manual of Tests and Criteria, 5th revised Edition. United Nations (2011).

44. Fayet G, Rotureau P, Prana V and Adamo C, Global and local quantitative structure–property relationship models to predict the impact sensitivity of nitro compounds. *Process Saf Prog* **31**:291-303 (2012).
45. Fayet G, Del Rio A, Rotureau P, Joubert L and Adamo C, Predicting the thermal stability of nitroaromatic Compounds using Chemoinformatic Tools. *Mol Inform* **30**:623-634 (2011).
46. Prana V, Rotureau P, Fayet G, André D, Hub S, Vicot P, Rao L and Adamo C, Prediction of the thermal decomposition of organic peroxides by validated QSPR models. *J Hazard Mater* **276**:216-224 (2014).
47. Benassi R, Folli U, Sbardellati S and Taddei F, Conformational properties and homolytic bond cleavage of organic peroxides. I. An empirical approach based upon molecular mechanics and ab initio calculations. *J Comput Chem* **14**:379-391 (1993).
48. Benassi R and Taddei F, Homolytic bond-dissociation in peroxides, peroxyacids, peroxyesters and related radicals: ab-initio MO calculations. *Tetrahedron* **50**:4795-4810 (1994).
49. Diallo AO, Fayet G, Len C and Marlair G, Evaluation of Heats of Combustion of Ionic Liquids through Use of Existing and Purpose-Built Models. *Ind Eng Chem Res* **51**:3149-3156 (2012).
50. Saldana DA, Starck L, Mougin P, Rousseau B and Creton B, Prediction of Flash Points for Fuel Mixtures Using Machine Learning and a Novel Equation. *Energy Fuels* **27**:3811-3820 (2013).
51. Gaudin T, Rotureau P and Fayet G, Combining mixing rules with QSPR models for pure chemicals to predict the flash points of binary organic liquid mixtures. *Fire Saf J* **74**:61-70 (2014).
52. Liaw H-J, Gerbaud V and Chiu C-Y, Flash Point for Ternary Partially Miscible Mixtures of Flammable Solvents. *J Chem Eng Data* **55**:134-146 (2009).
53. Gaudin T, Rotureau P and Fayet G, Mixture Descriptors toward the Development of Quantitative Structure-Property Relationship Models for the Flash Points of Organic Mixtures. *Ind Eng Chem Res* **54**:6596-6604 (2015).
54. Rücker C and Kümmerer K, Modeling and predicting aquatic aerobic biodegradation - a review from a user's perspective. *Green Chem* **14**:875-887 (2015).
55. Cronin MTD and Worth AP, (Q)SARs for Predicting Effects Relating to Reproductive Toxicity. *QSAR Comb Sci* **27**:91-100 (2008).
56. Netzeva TI, Pavan M and Worth AP, Review of (Quantitative) Structure–Activity Relationships for Acute Aquatic Toxicity. *QSAR Comb Sci* **27**:77-90 (2008).
57. Gasteiger J and Engel T, *Chemoinformatics - A textbook*. Wiley, Weinheim, (2003).
58. Kumar V, Krishna S and Siddiqi MI, Virtual screening strategies: Recent advances in the identification and design of anti-cancer agents. *Methods* **71**:64-70 (2015).
59. Comley J, Tools and Technologies that Facilitate Automated Screening. In *High-Throughput Screening in Drug Discovery*, Wiley-VCH Verlag pp 37-73 (2006).
60. Oprea TI and Tropsha A, Target, chemical and bioactivity databases - integration is key. *Drug Discov Today Tech* **3**:357-365 (2006).
61. Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG and Kuz'min VE, Existing and Developing Approaches for QSAR Analysis of Mixtures. *Mol Inform* **31**:202-221 (2012).
62. 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).
63. Tropsha A, Integrated Chemo- and Bioinformatics Approaches to Virtual Screening. In *Chemoinformatics Approaches to Virtual Screening*, Varnek, A.; Tropsha, A., Eds. The Royal Society of Chemistry, pp 295-325 (2008).
64. Zhang S, Wei L, Bastow K, Zheng W, Brossi A, Lee K-H and Tropsha A, Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J Comput Aid Mol Des* **21**:97-112 (2007).
65. ChemDiv Database. [www.chemdiv.com](http://www.chemdiv.com) [accessed 09/09/2015].
66. Harper PM, Gani R, Kolar P and Ishikawa T, Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equil* **158-160**:337-347 (1999).
67. Karunanithi AT, Achenie LEK and Gani R, A computer-aided molecular design framework for crystallization solvent design. *Chem Eng Sci* **61**:1247-1260 (2006).

68. Satyanarayana KC, Abildskov J and Gani R, Computer-aided polymer design using group contribution plus property models. *Comput Chem Eng* **33**:1004-1013 (2009).
69. Mattei M, Kontogeorgis GM and Gani R, A comprehensive framework for surfactant selection and design for emulsion based chemical product design. *Fluid Phase Equil* **362**:288-299 (2014).
70. Weis DC and Visco DP, Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. *Comput Chem Eng* **34**:1018-1029 (2010).
71. Moity L, Molinier V, Benazzouz A, Barone R, Marion P and Aubry J-M, In silico design of bio-based commodity chemicals: application to itaconic acid based solvents. *Green Chem* **16**:146-160 (2014).
72. Heintz J, Belaud J-P, Pandya N, Teles Dos Santos M and Gerbaud V, Computer aided product design tool for sustainable product development. *Comput Chem Eng* **71**:362-376 (2014).
73. Bechthold I, Bretz K, Kabasci S, Kopitzky R and Springer A, Succinic Acid: A New Platform Chemical for Biobased Polymers from Renewable Resources. *Chem Eng Tech* **31**:647-654 (2008).
74. Dusselier M, Van Wouwe P, Dewaele A, Makshina E and Sels BF, Lactic acid as a platform chemical in the biobased economy: the role of chemocatalysis. *Energy Env Sci* **6**:1415-1442 (2013).
75. Rücker C, Meringer M and Kerber A, QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points. *J Chem Inf Comput Sci* **44**:2070-2076 (2004).
76. Gugisch R, Laue R, Kerber A, Kohnert A, Meringer M, Rücker C and Wassermann A MOLGEN - Molecular Structure Generation. <http://www.molgen.de/> [accessed 09/09/2015].
77. Matsuda H, Yamamoto H, Kurihara K and Tochigi K, Computer-aided reverse design for ionic liquids by QSPR using descriptors of group contribution type for ionic conductivities and viscosities. *Fluid Phase Equil* **261**:434-443 (2007).
78. Miyao T, Arakawa M and Funatsu K, Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Mol Inform* **29**:111-125 (2010).
79. Prana V, Fayet G, Rotureau P and Adamo C, Development of validated QSPR models for impact sensitivity of nitroaliphatic compounds. *J Hazard Mater* **235-236**:169-177 (2012).
80. Xu J, Zhu L, Fang D, Wang L, Xiao S, Liu L and Xu W, QSPR studies of impact sensitivity of nitro energetic compounds using three-dimensional descriptors. *J Mol Graph Model* **36**:10-19 (2012).
81. Wang R, Jiang J, Pan Y, Cao H and Cui Y, Prediction of impact sensitivity of nitro energetic compounds by neural network based on electrotopological-state indices. *J Hazard Mater* **166**:155-186 (2009).
82. Gharagheizi F, Prediction of upper flammability limit percent of pure compounds from their molecular structures. *J Hazard Mater* **167**:507-510 (2009).
83. Gharagheizi F, A new group contribution-based model for estimation of lower flammability limit of pure compounds. *J Hazard Mater* **170**:595-604 (2009).
84. Carroll FA, Lin C-Y and Quina FH, Simple Method to Evaluate and to Predict Flash Points of OrganicCompounds. *Ind Eng Chem Res* **50**:4796-4800 (2011).
85. Rowley JR, Freeman DK, Rowley RL, Oscarson JL, Giles NF and Wilding WV, Flash Point: Evaluation, Experimentation and Estimation. *Int J Thermophys* **31**:875-887 (2010).
86. Hall LH and Story CT, Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks. *J Chem Inf Comput Sci* **36**:1004-1014 (1996).
87. Gharagheizi F, An accurate model for prediction of autoignition temperature of pure compounds. *J Hazard Mater* **189**:211-221 (2011).
88. Liaw H-J, Gerbaud V and Li Y-H, Prediction of miscible mixtures flash-point from UNIFAC group contribution methods. *Fluid Phase Equil* **300**:70-82 (2011).

**Table 1 - QSPR models to classify substances according to the CLP regulation for physical hazards**

<b>Explosive substances</b>		
	Temperature and heat of decomposition	Recent validated models have been developed for nitro compounds with errors lower than 20% on heats of decomposition <sup>40</sup> .
	Impact sensitivity	Few validated models exist for nitro compounds with errors about 0.2-0.25 (log) <sup>42, 44, 79-81</sup> .
<b>Flammable gases</b>		
	Lower and Upper Flammability Limits (LFL/UFL)	Recent models exist with errors about 10% for UFL <sup>82</sup> and 5% for LFL <sup>83</sup> .
<b>Flammable liquids</b>		
	Flash point	Validated models applicable for organic compounds exist with errors lower than 5°C for the best ones <sup>84, 85</sup> . First predictive approaches have been recently proposed for mixtures with errors about 4°C <sup>50, 51, 53</sup> .
	Boiling point	Numerous models have been developed among years for very diverse families of compounds. Validated models reach errors lower than 5°C for the most accurate ones <sup>86</sup> .
	Self ignition temperature	Validated models exist with different performances upon the target compounds. Errors about 15°C are reached for models widely applicable to organic compounds <sup>87</sup> .
<b>Organic peroxides</b>		
	Temperature and heat of decomposition	Validated models have been recently developed with R <sup>2</sup> of 0.82 and 0.90 (in external validation) for the heat and temperature of decomposition, respectively <sup>46</sup> .

**Figure 1 - Principle of the QSPR method**

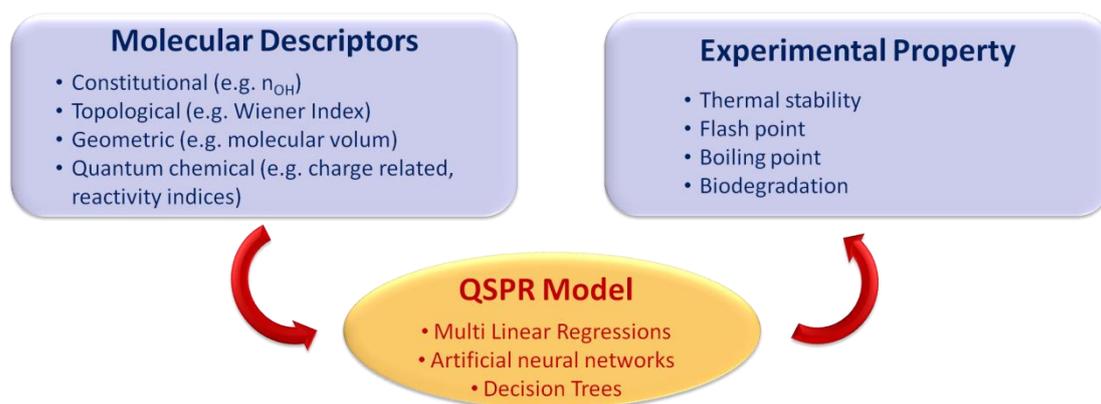


Figure 2 – Nitrobenzene derivatives presenting no substituent in ortho position to the nitro group

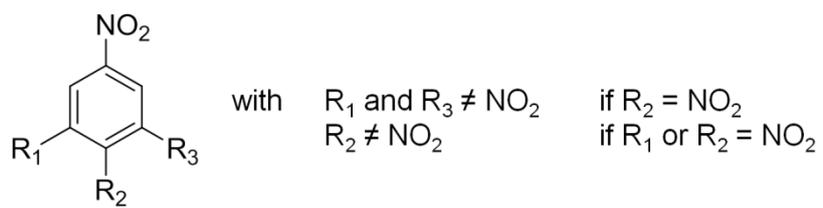


Figure 3 - Experimental vs. predicted impact sensitivity of nitramines from Eq. 4 (adapted from Ref. 42)

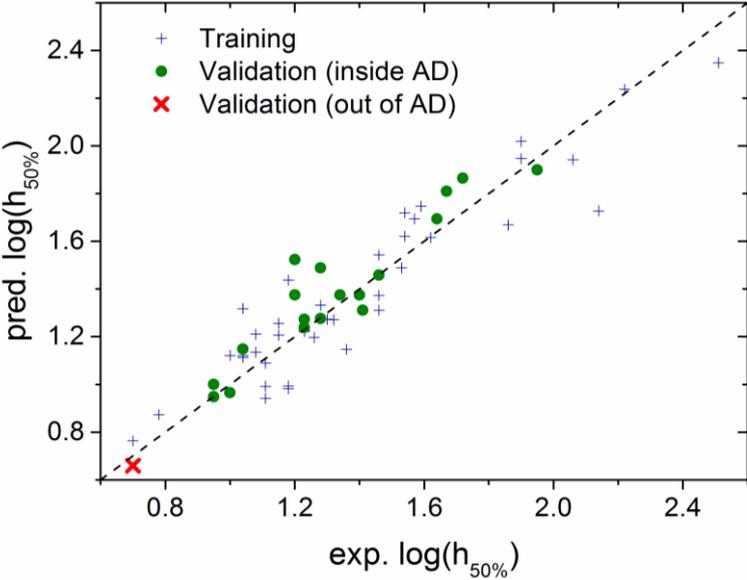


Figure 4 – Decision tree model to predict the heat of decomposition of nitroaromatic compounds (adapted from Ref. <sup>45</sup>).

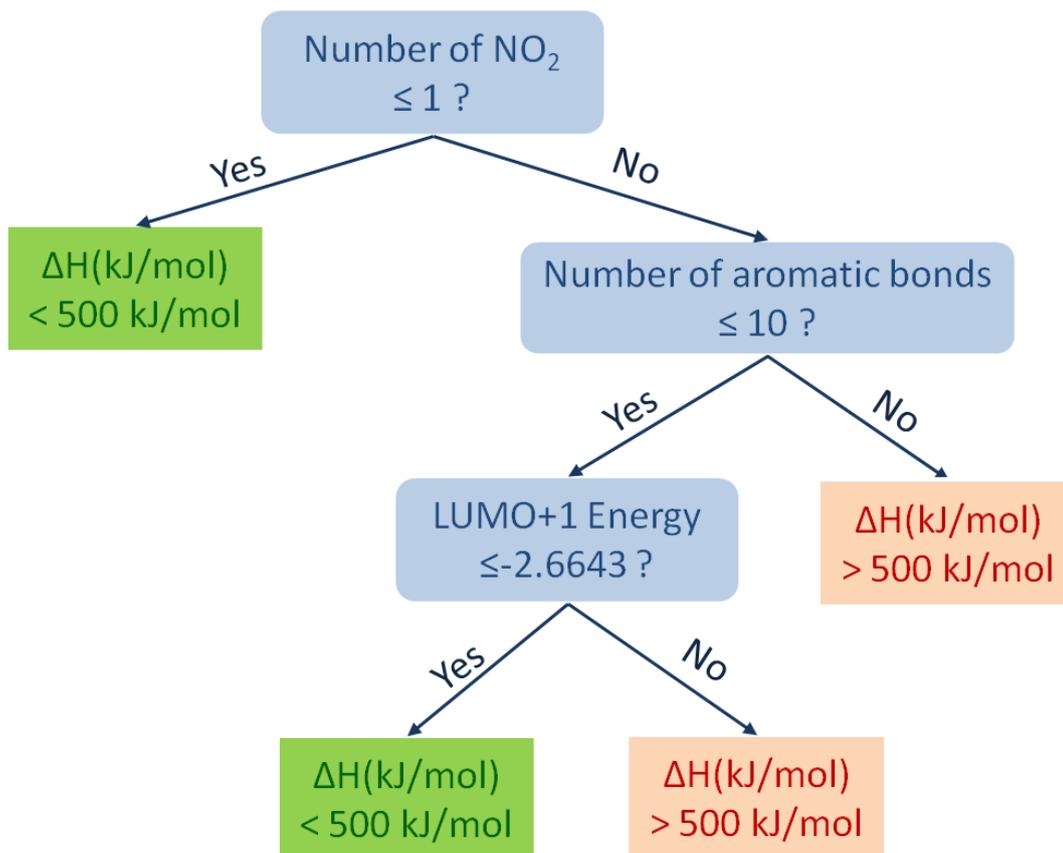


Figure 5 - Flash point of n-octane/isopropanol mixtures from experiments<sup>88</sup> and from the full predictive method of Gaudin et al.<sup>51</sup> (adapted from Ref. <sup>51</sup>).

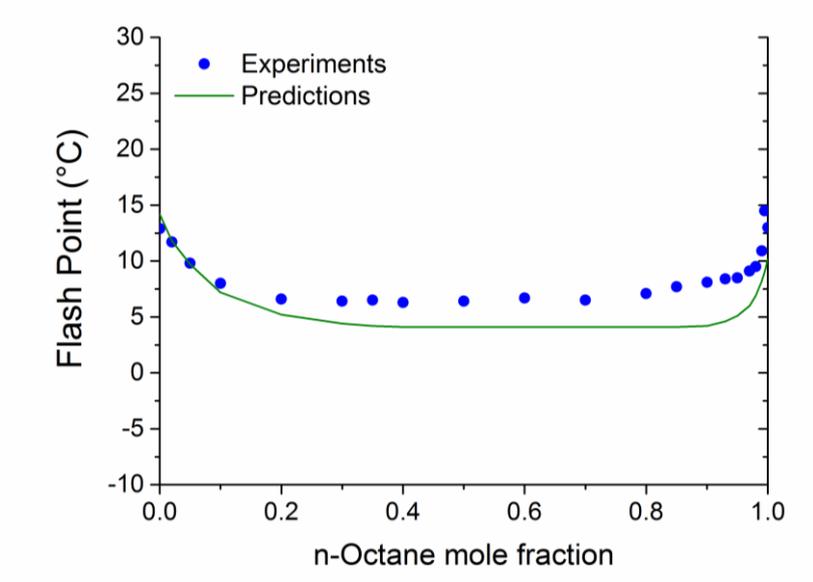


Figure 6 - Ternary diagram of flash points of the methanol/toluene/2,2,4-trimethylpentane mixture calculated by the full-predictive method of Gaudin et al.<sup>51</sup>

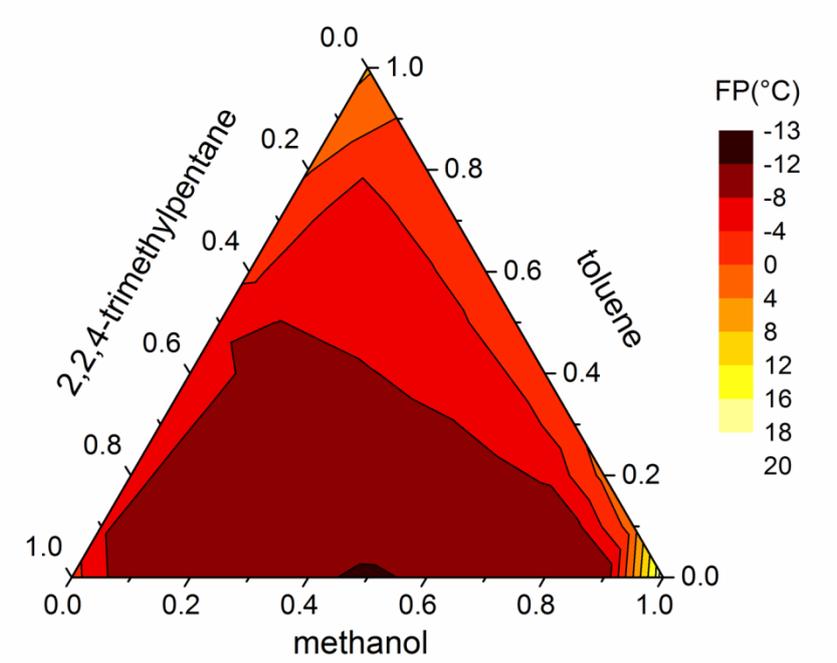


Figure 7 - Screening databases of chemicals using QSPR models

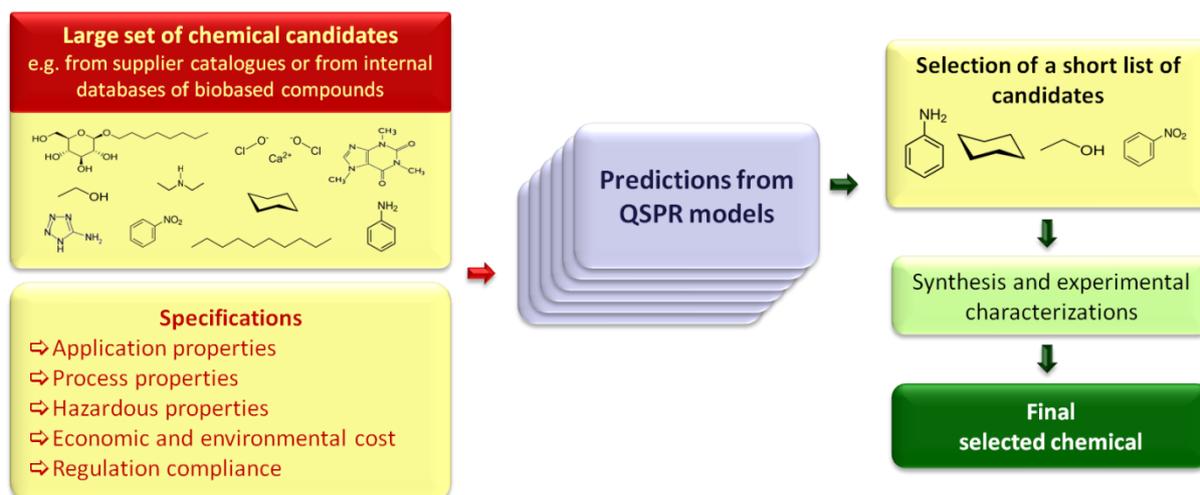


Figure 8 – Structure of the best PBT derivative proposed by Zhang et al.<sup>64</sup>

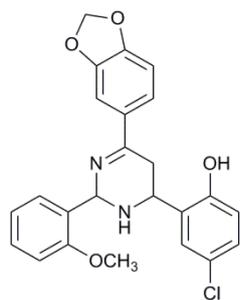


Figure 9 - General structure of an *in silico* study based on the generation and screening of a virtual library

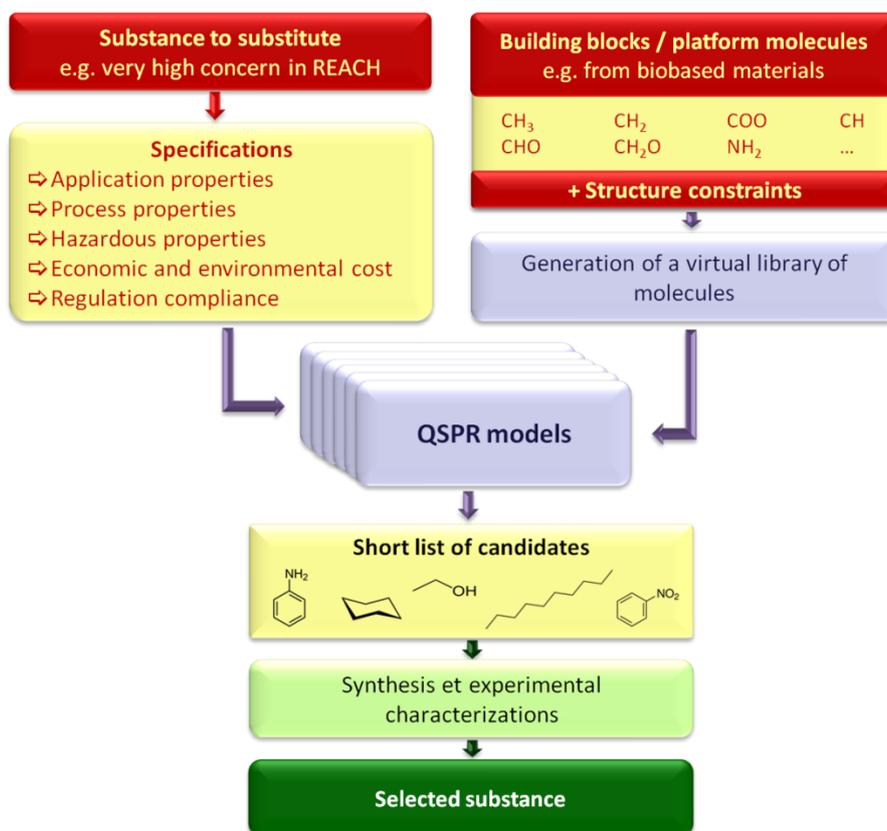


Figure 10 - Possible chemical conversions from succinic acid as platform molecule<sup>73</sup>

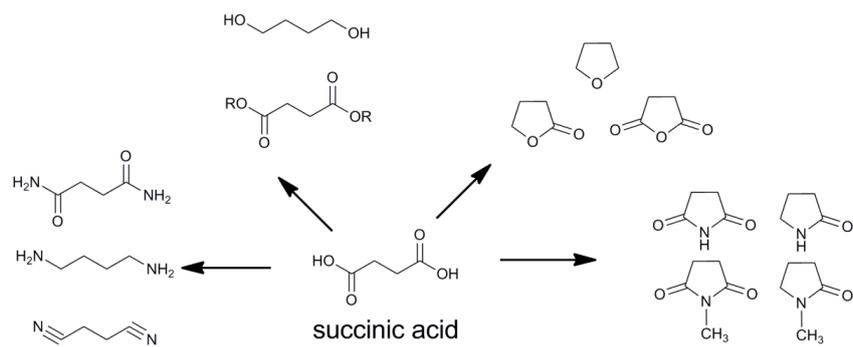


Figure 11 - Possible chemical conversions from lactic acid as platform molecule<sup>74</sup>

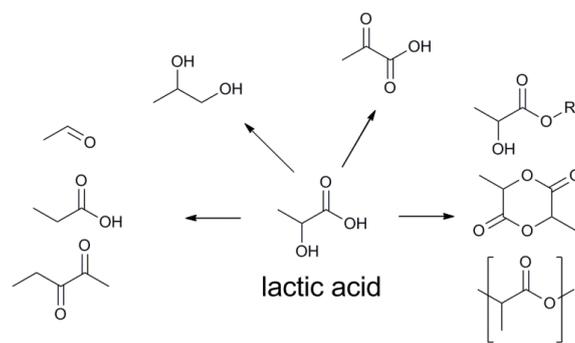


Figure 12 - Structural parameters for the generation of IL by Matsuda et al.<sup>77</sup>

