applied sciences

MDPI

Review

A Way towards Reliable Predictive Methods for the Prediction of Physicochemical Properties of Chemicals Using the Group Contribution and other Methods

Robert J. Meier

DSM Materials Science Center, 6160 BB Geleen, The Netherlands; meier014@planet.nl

Received: 8 March 2019; Accepted: 18 April 2019; Published: 24 April 2019

Abstract: Physicochemical properties of chemicals as referred to in this review include, for example, thermodynamic properties such as heat of formation, boiling point, toxicity of molecules and the fate of molecules whenever undergoing or accelerating (catalytic) a chemical reaction and therewith about chemical equilibrium, that is, the equilibrium in chemical reactions. All such properties have been predicted in literature by a variety of methods. However, for the experimental scientist for whom such predictions are of relevance, the accuracies are often far from sufficient for reliable application We discuss current practices and suggest how one could arrive at better, that is sufficiently accurate and reliable, predictive methods. Some recently published examples have shown this to be possible in practical cases. In summary, this review focuses on methodologies to obtain the required accuracies for the chemical practitioner and process technologist designing chemical processes. Finally, something almost never explicitly mentioned is the fact that whereas for some practical cases very accurate predictions are required, for other cases a qualitatively correct picture with relatively low correlation coefficients can be sufficient as a valuable predictive tool. Requirements for acceptable predictive methods can therefore be significantly different depending on the actual application, which are illustrated using real-life examples, primarily with industrial relevance. Furthermore, for specific properties such as the octanol-water partition coefficient more close collaboration between research groups using different methods would greatly facilitate progress in the field of predictive modelling.

Keywords: thermodynamic properties; molecular properties; physical chemical properties; chemicals; group contribution method; Ab initio method; neural network; QSPR; predictive method

1. Predictive Methods for Physicochemical Properties of Molecules

1.1. Introduction

The title of this review suggests a very broad domain and indeed it is and very many techniques can be applied to predict a larger variety of physicochemical properties or a single specific property. In this tutorial review we want to focus on selected properties that are of eminent importance in industry but not only in industry. These properties include heat of formation, boiling point, toxicity of molecules and the fate of molecules whenever undergoing or accelerating (catalytic) a chemical reaction and therewith about chemical equilibrium, that is, the equilibrium in chemical reactions. For larger molecules like polymers also properties including the glass transition temperature can be evaluated as a chemical groups based property. In addition, we focus on the required accuracy and reliability of the prediction, which are key in applications such as chemical process design, as well as the speed with which the properties can be predicted. *For speed-up of innovation one needs sufficiently accurate and reliable prediction and preferably 'on the fly' so that different scenarios can be compared while*

designing a process or the next series of experiments in the chemistry lab. It is for these reasons that we focus on certain properties, for example those mentioned and at the same time do not discuss simulation methods that need specific expertise and /or long simulation times. All the methods we do not discuss have been extensively discussed including discussions on performance in previous reviews we will refer to below.

There are more than 1 billion organic molecules with 13 heavy atoms (the GDB-13 database [1], whereas GDB itself is just a data base file format), where we only refer to a subset of the organic molecules only. When we need the properties of any molecule, it will be evident that determining all properties experimentally is undoable. Even more so the other way around: you are looking for a molecule with certain properties, which one has these properties? Moreover, in practice we need to deal with organometallics and mixtures too. In this review, we will primarily focus on the very generically used and important class of organic molecules. To ensure that such a method would bring true added value, the reliability and accuracy must fulfil certain requirements. However, is this realistic when we have strict requirements on accuracy and the reliability of prediction?

For physicochemical property prediction, it is to be noticed that for many properties, if not all, the predictions are not yet of the quality they should be to allow them to be used with full confidence. With quality, we not only refer to the accuracy of the predicted value but also to the reliability of the prediction. Predictive methods are often qualified by mentioning root-mean-square deviations from experimental values or a similar quantity. Even if such average deviations are small, often cases with very high deviations are found without any clear reason why exactly this molecule is an outlier. This implies the method is not sufficiently reliable. Good methods should next to having a small root-mean-squared deviation also have a maximum deviation. As an example, a report has been published [2] comprising the results of an investigation on industrial requirements for thermodynamic and transport properties carried out by the Working Party on Thermodynamic and Transport properties (http://www.wp-ttp.dk/) of the European Federation of Chemical Engineering (EFCE) (www.efce.info). An overview of desired accuracy (at least for industrial use) is shown in Table 1 where it is emphasized that the numbers in the column achieved refer to average accuracies and does not exclude large outliers that should be avoided for any appropriate predictive method.

Table 1. Indications of achieved and realistically needed accuracies for various properties.

Physical Property	Achieved	Needed
Heat of formation	2.5–4	4 kJ/mol
Liquid heat capacity	>10%	10%
Liquid density	>2%	2%
Vapour pressure	>10%	10%
Normal boiling point	6 K	3 K
Melting point	20 K	<10 K
Toxicity	typically 70% correct	>95% correct

Thermodynamic data are key in the understanding and design of chemical processes. Next to the experimental evaluation of such data, computational methods are valuable and sometimes indispensable tools in obtaining heats of formation and Gibbs free energies. The major toolboxes to obtain such quantities by computation are quantum mechanical methods and group contribution methods. Although a lot of progress was made over the last decade, for the majority of chemical species we are still quite a bit away from what is often referred to as chemical accuracy, that is, '1 kcal/mole.' Currently, for larger molecules the combination of group contribution methods with group additive values that are determined with the best available computational *ab initio* methods seems to be

a viable alternative to obtain thermodynamic properties near chemical accuracy. New developments and full use of existing tools may lead to further improvements.

Thus, whereas predictive methods for thermodynamic properties and phase equilibria are not new, the current accuracy of predictive methods is far from what is needed for reliable designs, equipment and process optimization. The accuracy should be pushed by a factor of 2–5 to get into the desired regime. In many scientific publications the % deviation is quoted. However, as we see from the required accuracies in the above, some properties require an absolute accuracy, for example heat of formation with chemical accuracy '1 kcal/mole' and boiling point 3 K and not a percentage of the boiling point. This is because the chemical equilibrium depends on the absolute difference, needed is often the '1 kcal/mole' and separation by distillation requires a certain absolute difference in boiling points and not a percentage. An excellent and extensive review entitled "Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction" has been reported by Katritzky et al. [3] and it provides, next to the state of the art in 2010, many examples that illustrate the arguments we put forward in this paper. Another, more recent review addresses the prediction of physico-chemical properties for regulatory processes [4].

In the present review, we discuss the current situation at a general level and we address the question: why are current methodologies often not leading to the desired accuracies in prediction and is there a consistent way to improve quality? The focus is on how to arrive at predictive methodologies that fulfil the requirements for the practitioner to whom these methods are potentially useful. These include chemists, process technologists and engineers, toxicologists and environmental chemists. The cases we present for illustration generally do not contain details for reasons of confidentiality of the systems discussed but they originate from real-life industrial experience.

1.2. Current Methods

As mentioned in the introduction here we focus on the required accuracy and reliability of a predictive method, whereas a significant additional benefit would be the on the fly (instantaneous) prediction of a property, the latter being of particular interest for technologists carrying out (dynamical) process design.

Ab−initio quantum mechanical calculations are based on solving the Schrödinger equation in one way or the other and thus start from basic physical principles. They allow the calculation of a range of properties including heat of formation, free energy of formation, dipole moment and many more, without the necessity to have available any experimental data for parametrisation. Many other properties are not within reach, for example the melting point of a substance or the prediction of toxicity (unless, in the latter case, we would know the exact mechanism of action of what makes a molecule toxic at the molecular level). Furthermore, ab−initio methods and most certainly when the required accuracy needs to be approached, is computationally expensive and therefore inappropriate for rapid screening in for instance process design. A single quantum mechanical ab initio or Density Functional Theory (DFT) calculations take in the order of hours per molecule (on a single CPU). Calculating 'on the fly' would is not an option and, moreover, it generally requires expert knowledge to perform these calculations properly.

Unfortunately and despite some reports in literature that appear optimistic, these methods have, generally speaking, not reached the accuracy and reliability level required [5]. The optimism in literature is often based on the fact that a specific class of related molecules is being treated with a specific method (e.g., alkanes or alcohols only) or so-called hybrid calculational schemes (technically speaking we refer to methods like G2, G4, etc.) which are in fact semi-empirical approaches as parameters were optimized in these schemes to obtain the best possible result for a certain, rather finite, set of test molecules. On the other hand, *ab-initio* methods can be used to allow for the calculation of molecular descriptors that can be used in Quantitative Structure Activity Relationship (QSAR) approaches involving those molecular descriptors (see further below in this review). To obtain reasonable bond lengths and other geometrical information as well as for instance atomic charges

ab initio methods can be used which are not very expensive. In this way ab initio calculations can make a most valuable contribution in conjunction with statistical QSAR methods involving molecular descriptors, as they can be applied to compound classes including organometallics. Many of the most prominent papers up till 2010 are quoted in the Reference [5]. A paper involving molecular simulations starting from molecular structures computed by quantum mechanics revealed predicted molecular liquid densities that are getting close to what is needed (< 2%) with the root-mean-squared deviations for halogenated compounds being the largest outlier (3.4%) [6].

As an alternative to the more rigorous but time-demanding ab initio calculations, for which one should also have the proper expertise, semi-empirical quantum mechanical methods such as the AM1 or PM6 methods can be very useful but have become somewhat forgotten despite good performance in areas like molecular entropy of organics [7], whereas we also refer to a more recent study involving semi-empirical quantum calculations review focusing on heat capacities and energies of organics [8]. There current use seems limited, despite some very useful specific applications in combination with the calculations being very fast, as the performance area is limited (the classes of molecules and properties for which proper parametrization was accomplished in the past) or simply insufficiently accurate (non-organics),

In the ***QSPR (Quantitative Structure Property Relationships)*** approach, in which so-called molecular descriptors are involved that characterise the molecule, descriptors are linked with the physicochemical property of interest by establishing a relation

$$\text{Physicochemical property of interest} = f\ (\text{molecular descriptors}) \quad (1)$$

Molecular descriptors are parameters that describe characteristics of the molecule and vary from geometrical parameters such as bond lengths and bond angles to atomic charges, atom-atom connectivity, molecular volume, cone angles, HOMO and LUMO energies and so on. Formally, a definition of such descriptors was given by Todeschini and Consonni [9]. Descriptors can be evaluated by well-established procedures and available programs. One methodology largely uses quickly calculable parameters including those based on geometry and connectivity, (partial) molar volumes, components of the dipole and quadruple moment vectors and many more. In fact, the data presented in Figure 4 were based on such an approach (homogeneous catalysts). The descriptors can be computed using software programs such as DRAGON (http://www.moleculardescriptors.eu/resources/resources.htm) or CODESSA (http://codesa.weebly.com/software.html), the MS.QSAR module of Accelrys Inc. and many more, all easy to connect to on the Internet, several of these are freeware. For a useful further reference on molecular descriptors see Reference [10]: "The great descriptor melting pot: mixing descriptors for the common good of QSAR methods." An excellent introduction describing various approaches and their advantages and disadvantages was presented by Le et al. [11]. The other way to obtain (additional) molecular descriptors is to use ab initio quantum mechanical calculations to compute the geometry and electronic structured of a molecule and from that derive a large set of molecular descriptors [12]. Both approaches can be used individually or combined. Next to using ab-initio methods, the much faster semi-empirical quantum mechanics methods can be used to generate and evaluate molecular descriptors.

Next, a mathematical relation (viz. Equation (1)) is established between the descriptors and each property of interest. This can be accomplished by statistical methods optimizing the parameters to be fitted in the model. In many cases a PCA (principal component analysis) is performed first so only the most relevant (statistical correlation) descriptors are retained. As the total number of descriptors easily runs into many hundreds, this PCA reduces the risk of overfitting (many fitting parameters many lead to very different models for the same property, a risk that can only be overcome by having a much larger set of property data compared to the number of molecular descriptors). This class of methods is generally applicable, for any type of molecule. When a property is required for a new molecule, the descriptor values for that new molecule are calculated and substituted into the model, which then provides the predicted value for the property of relevance. This methodology has been

applied to a very wide range of properties, from which we can only name a few. Surfactant cloud points [13], pKa [14], normal boiling points [15] and another one in which the impressive number of 17768 pure chemical compounds were considered and which includes an overview of previous models [16], octanol-water partition coefficients [17] and homogeneous catalysis [18]. The observation though is that accuracy and reliability are often still too limited, something we will further comment on after the next paragraph on the Group Contribution methodology.

The third well-known method is the ***Group Contribution (GC) method***. The GC methods (see [19–22]) belong to the class of empirical property models, also known as additive methods, where the molecular structure of a compound is broken down into *building blocks* and the property of a compound is estimated by the summation of the contributions of each building block as schematically illustrated in Figure 1. In GC-methods, the building blocks are functional groups and these methods are based on the assumption that a property value of any group has the same contribution in all the compounds where it appears and that the property value of the compound is a function of the contributions of all the groups needed for a unique representation of the molecular structure of the compound. In this way, the GC-methods are similar to methods based on topological indices [23], where the building blocks are topological indices [24,25], bond contributions [25,26], conjugates [27,28] or a combination of them. The GC-methods could also be considered a special class of QSAR methods [29,30] where the molecular descriptors are the building blocks, that is, the functional groups. Group or Fragment based QSAR is also known as GQSAR.

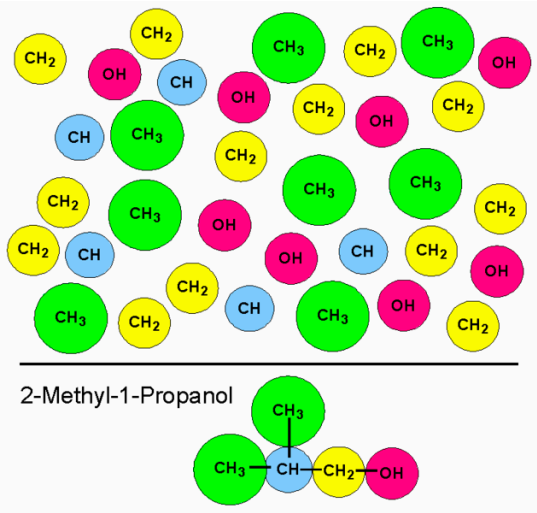

Figure 1. Schematic representation of the Group Contribution concept where the molecule is broken into building blocks.

The simplest form of a GC approach has a large set of simple groups that allows description of the molecular structures of a wide variety of organic compounds, for example

$$\text{Property P} = \text{constant} + \sum_i C_i \tag{2}$$

in which the C_i are values characteristic of a chemical group. So, for instance for alkanes one can well imagine that some properties depend on the CH_3 end-groups and the number of CH_2 groups only. In such a case, very few parameters can make a very good method. Obviously, the entire idea originates from the empirical knowledge that individual chemical groups have an individual and additive contribution to many properties. However, these groups capture only partially the proximity effects and are unable to distinguish among isomers. For this reason, the first level of estimation is intended to deal with simple and mono-functional compounds. To accommodate this deficiency, Marrero and Gani [23] and, before them, Constantinou and Gani [21] and Benson et al. [20] (who also used non-linear

group contributions) introduced higher-order groups, the purpose of which was to add molecular structural information (interactions between groups) of the compound as a higher-level contribution or correction to the value from the first level. The ultimate objective of the proposed multilevel scheme is to enhance the accuracy, reliability and the range of application for a number of important pure component properties. Also for the Group Contribution approach there exist many references for many properties. The Gani group has probably the currently most extensive set of properties combined with, at least on average, the best accuracy (http://www.capec.kt.dtu.dk/Software/ICAS-and-its-Tools/). Among the properties we find vapour pressure and heat of vaporization [31], enthalpy of formation with the required chemical accuracy [32] and environment related properties [33].

2. Accuracy of Predictive Methods and What Is Required

2.1. On the Assessment of the Quality of the Methods Evaluating Physicochemical Properties

When we develop or validate predictive methods, we need a way to judge about the quality of a method. Often the accuracy of methods is quantified by quantities such as the quality of fit, for example R^2 as the proportion of variability in a data set that is accounted for by a statistical method, $0 \leq R^2 \leq 1$, where for R^2 =1 we have perfect correlation. There are other quantities that may be preferred or that provide additional information on the quality of the method but the discussion below actually applies to most of these. As often used in statistical analysis, quantities like R^2 can be a good means to assess the quality of one method but depending on the application this may be a rather meaningless quantity as we will see further below. In many references, as is common in statistics, R^2 values in the range 0.6–0.8 are considered a good correlation. A typical interpretation for different fields of application is given in the table (Table 2) below.

Table 2. Typical variability values for acceptable or good methods for various disciplines (these should be interpreted as typical values; the numbers in the table were taken from http://condor.depaul.edu/sjost/it223/documents/correlation.htm).

Discipline	r Meaningful if	R^2 Meaningful if
Physics	$r < -0.95$ or $0.95 < r$	$0.9 < R^2$
Chemistry	$r < -0.9$ or $0.9 < r$	$0.8 < R^2$
Biology	$r < -0.7$ or $0.7 < r$	$0.5 < R^2$
Social Sciences	$r < -0.6$ or $0.6 < r$	$0.35 < R^2$

A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. A study in chemistry reporting R^2=0.85 could therefore be considered as state-of-the-art.

Despite what on average is considered 'good' in the different disciplines, the absolute quality of prediction is of course worse with decreasing R^2 values. Is not a method with R^2 = 0.6 far from sufficient when we look at what this means in a practical sense? And would a true human expert not make an equally good prediction without the use of methods but based on experience? (see also Table 3 further below for an explicit example). Another important factor is whether (i) the predictions should be quantitatively correct over the entire numerical data range or (ii) it is only of interest to see whether values fall within in a certain domain. We will elaborate on this in the following two sub-sections.

2.2. Cases Requiring Quantitative Predictions over the Data Range

For a polymer, it can be important to predict the glass transition temperature Tg within a few degrees. Depending on the application, one is seeking polymer leading to a certain Tg. In Figure 2, we have displayed three correlations between experimental and modelled Tg values, with different methods providing different quality. Here we see that even R^2=0.99 still does not yield Tg prediction

that are all within a few degrees accuracy. In fact, one would need $R^2 = 0.999$ level to have sufficiently good predictions, something that might be difficult to achieve.

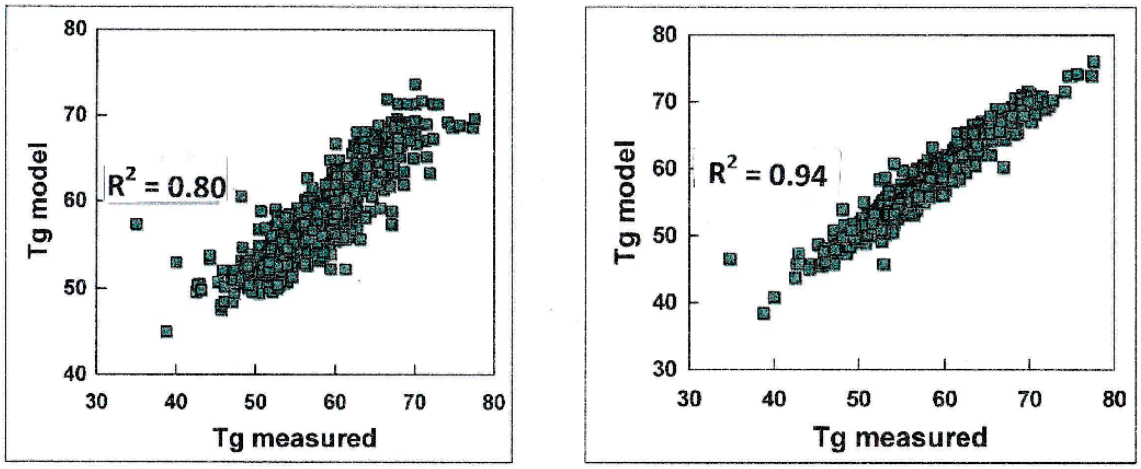

Figure 2. Neural network based methods for the glass transition temperature Tg for a series of polymers. Different models were built leading to different quality with R^2 ranging from 0.80 to 0.99. It can be seen clearly that for $R^2 < 0.99$ the model does not describe all Tg values within a few degrees.

A similar case is the prediction of a boiling point with the application of separation using distillation in mind. To predict whether two compounds can be separated, not only the relative order of the boiling points should be correct but also the reliability of the predicted difference in boiling points is crucial. The last requirement must be satisfied by any predictive method to be really useful.

A further class of examples comprises the cases of chemical equilibria. These are the cases where it is important to know whether in the reaction

$$A + B \rightarrow C + D \tag{3}$$

the equilibrium lies clearly to the right (all reactants transform in products) or to the left or the equilibrium is in between, all species being present all the time. When there is 0% or 100% yield, we have a large positive or negative ΔG and an accurate prediction is less important. For cases where ΔG is close to zero we actually need a very accurate, at the level of 4 kJ/mole accuracy, prediction in order to assess whether we have 20% or 80% yield for instance as that may well determine the economic feasibility of a process. A specific example is the transaminase reaction

$$X\text{-}NH_2 + YC{=}O \rightarrow XC{=}O + Y\text{-}NH_2 \tag{4}$$

with the final amine as given target molecule, the initial ketone is known and the question is what is the right choice for the initial amine to have a successful transformation (equilibrium largely on the right-hand-side). Relatively simple quantum mechanical calculations have been shown to be a successful toll to predict whether the reaction is thermodynamically favourable [34]. The oval in Figure 3 shows the effect of a 4 kJ/mole difference on the conversion and the data displayed reveal a good correlation between experimental data and model to allow for adequate predictions in an industrial context. The data that are really outside the oval on the right-hand side are those for which from experiments it is known that enzyme inhibition is known likely to occur.

Similar arguments and cases can be put forward for many other properties of interest to chemists and chemical engineers.

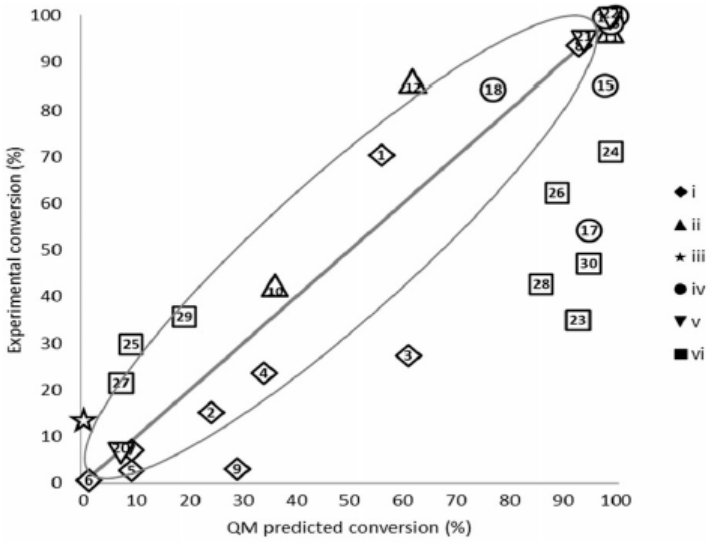

Figure 3. Comparison of predicted and measured transaminase catalysed conversions. For further details see the text and Reference [34].

2.3. Cases Requiring only Qualitative Predictive Power

For a catalyst, it can be important to know whether its selectivity is high, preferably close to 100%. This if different from, for example, the Tg case in the previous section, as now it only matters if a predicted high selectivity is indeed high in experiment. Whether low selectivity is predicted well or not is of no practical relevance. We illustrate this in the graphs in Figure 4. Although the model with R^2 = 0.999 shows an almost perfect correlation and the one with R^2 = 0.87 not, for the purpose of predicting high selectivities they do equally well—both predict the same 3 data points at the right-hand top of the graph as being good catalysts with high selectivity (both prediction and experiment show high selectivity).

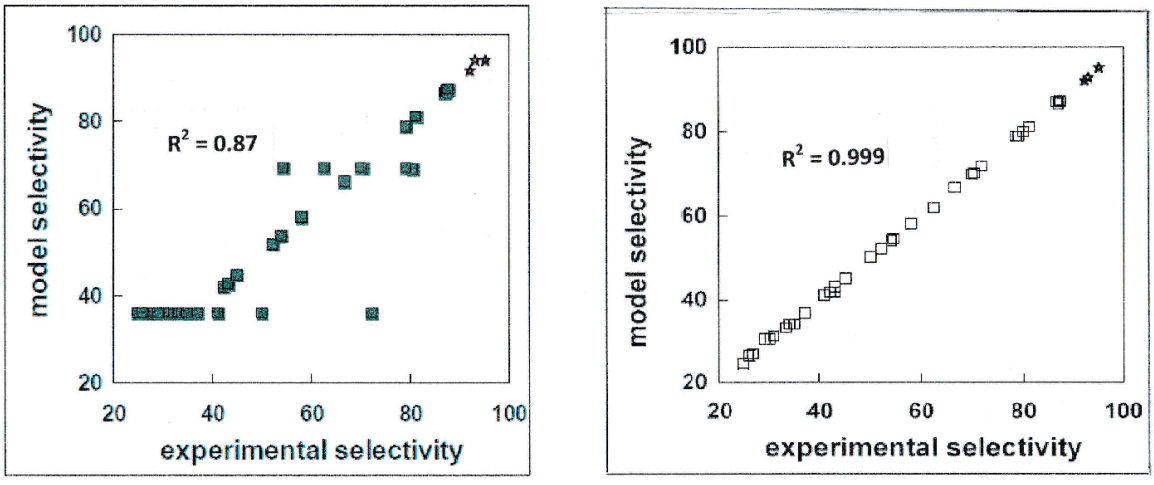

Figure 4. Experimental versus model selectivity for a realistic catalyst system. These plots serve to illustrate that in this case, where the main interest is to have a high selectivity, it does not matter that much if the method is not that good: in both cases, the same 3 data points represent the best catalysts (right hand top corner, star symbols).

A similar argument applies to a completely different domain of application, the prediction of toxicity properties such as LD_{50} type quantities (there are various such toxicity properties assessed by LD_{50}). LD stands for "Lethal Dose." LD_{50} is the amount of a material that, given all at once, causes the

death of 50% of a group of test animals. The LD_{50} is one way to measure the short-term poisoning potential (acute toxicity) of a material. The lower this value the more toxic the substance is, as it means a low dose is already poisonous and deathly. The quantity is normally expressed it is –log LD_{50} form and therefore the higher the value is the more toxic the compound. For the assessment of toxicity, this value is the threshold above which the compound is considered safe. Thus, as long as the predictions as well as the corresponding experimental data points are *both* above this threshold, the predictions are useful and valid. Quite similar are other toxicological endpoints such as mutagenicity as evaluated by the AMES test (named after Bruce Ames): it is qualified in a binary way too: mutagenic or not mutagenic.

In the cases discussed in this section R^2 values or similar statistical quantities do not seem really meaningful here, as it will not be high for the data shown but the method gives reliable prediction. There are different parameters to assess the quality of these methods: sensitivity, specificity and concordance. These quantities are statistical measures of the performance of a binary classification test, for example a substance is toxic or non-toxic. *Concordance* is the fraction of correct predictions, so if predicted safe it is indeed safe and in case predicted unsafe (toxic) indeed experimentally unsafe. Concordance levels below 70% are, however, no exception in the field, of toxicology. This implies that 30% of the predictions is wrong! Would not it be fair to ask one selves whether an experienced expert who also uses read-across can be equally good predictions? There is data around that support this, viz. Table 3. Although the predictive methods might have improved since this study, it is seen that human experts arrived at 75% good answers, which is higher (and this more reliable predictions) at the high end of values for predictive methods.

Table 3. The performance of some well-known in silico tools for the prediction. Table taken from Reference [35].

Tool	Concordance (%) = Agreement with Experiment
TOPKAT	58
DEREK	59
CASE	49
COMPACT	54
Human Experts	75

Although it might seem very odd at first instance, it seems realistic to state that if it was to be known that for a certain toxicological endpoint only 20% of all compounds that have been experimentally screened is to be considered toxic, a default answer of 'non-toxic' for any new molecule would eventually (for a larger number of compounds) have a higher concordance than some of the current methods. Finally, a very interesting study was reported by Thomas et al. [36]. The authors used an impressive number of 84 different statistical models on 60 in vivo endpoints. It was concluded that the used in vitro assays had limited applicability in predicting chemical hazards using statistical classification methods. It seems a valid conclusion to state that several methods need to be seriously improved to be of concrete, practical value.

2.4. Why do Methods often not Perform up to the Required Level of Accuracy and Reliability?

The question that is now pertinent is why property prediction methods, sufficiently often, do not yet lead to the required accuracy and reliability of prediction. We recall that the required accuracy and reliability is what bench chemists, process engineers and toxicologists need to make these predictions preferably first time right. We believe that, by experience, the following factors are relevant

(i) the preferred or exclusive use of a specific class of methods, for example quantum mechanics or statistical analysis using molecular descriptors

(ii) no optimal parametrization of statistical models or GC methods
(iii) quality of data
(iv) inappropriate account of the physics involved
(v) insufficient transfer of knowledge between different fields of science

Ad *(i) preferred or exclusive use of particular methods.* Scientist are experts in their field of expertise, and, generally speaking, we tend to apply this expertise to new upcoming problems. Theoretical chemists mostly apply ab initio calculations and molecular simulations to evaluate molecular properties but even after many years and high-level calculations the desired chemical accuracy of "1 kcal/mole" for heats of formation and limited number of outliers has not been achieved for a broader class of molecules using one methodology [5]. Here the recent results obtained by applying the Group Contribution method by Gani et al. [32] has delivered an appropriate solution that serves the need of both accuracy and reliability for heat of formation of several classes of organic molecules. Moreover, there is a straightforward and natural way to make that GC method better or extend it to new classes of molecules and results are produced 'on the fly' (this will be further elaborated in one of the next sections), a promise *ab initio* approaches are far away from.

This all applies to organic molecules. However, as Group Contribution methods rely on the summation of individual group contributions with some higher order contribution corrections, generally speaking, organometallic compounds with their much more complicated electronic structure and diffuse molecular orbitals cannot be realistically treated with such method as the overall property cannot be split into molecular group contributions in a simple and uniform way which is transferrable over the entire class of similar species. Here we thus still rely on, for example, quantum mechanical calculations even though the required accuracy will be difficult to achieve and the calculations are still expensive and not 'on the fly.' The difficulty with ab initio calculations remains that to arrive at the required accuracy and reliability extremely demanding and computationally intense calculations are required that also require the involvement of a true specialist. This will not facilitate the use of predictive methods by practitioners. An alternative for, for example, organometallics is the use of QSPR based on molecular descriptors rather than molecular group. See Section 3.3 for further reference.

Finally, thus far GC methods were not really developed for charged or radical type organic molecules but the available methodologies can be applied without any restriction to such classes of molecules. Also here ab initio calculations can be very useful to provide input, that is not available experimentally, to construct GC methods.

When looking at the molecular descriptors based and the Group Contribution methods these are, when appropriately developed, methods that can be applied 'on the fly' and, except for constructing the methods, can be used by practitioners without having detailed knowledge of the methodologies behind.

Ad *(ii) No optimal parametrization of statistical methods and GC methods* relates to the way one performs the fitting of the adjustable parameters in both methods. Once a set of properly screened (see previous paragraph) experimental data has been obtained, the fitting of the adjustable parameters can start. Although there is no need from a mathematical point of view, many of the statistical methods that are being applied involve linear models. There is, however, no physical reason why linear models would be sufficient to describe and properly predict physicochemical properties of chemicals of various kind. Many might argue and not without reason, that practice has shown that such methods work and indeed they do. A problem arises when we want to go for the, what we have called, required accuracy and reliability. Moreover, often a pre-filtering is applied which reduces the number of molecular descriptors taken into account, most of the time accomplished by a Principal Component Analysis. On the one hand, this procedure prevents overfitting (many parameters to be fitted without having an overwhelmingly larger number of experimental input data for the fitting) but on the other hand one retains typically 90–95% of the variability which implies that the required accuracy (see also the discussion in the Introduction) which often requires R^2 =99% or even more, see Section 2, will not be achieved.

Another reason that might lead to sub-optimal results relates to the way how the entire fitting is accomplished. When the total number of available experimental data is taken at once and the selected method parametrized using this data set, it is by no means guaranteed that this leads to a unique and best possible solution. This is what the authors explicitly saw when developing the GC method that finally led to the long-desired 1'kcal/mole accuracy' with no larger outliers [32]. It was found particularly helpful to look for trends among the molecules with larger outliers or incidental outliers which turned out strange as similar molecules were described very well. For instance, after one of the models was constructed and the root-mean-square deviation already better than that of at the time existing models, it was observed that molecules with a $-NO_2$ group has a severe error in the heat of formation and those with two NO_2 groups about twice that error. It turned out necessary to fit NO_2 containing molecular data first and then more or less fix (some flexibility should be left) the group parameters related to the NO_2 group and subsequently optimizing all other parameters. Also, third order groups, accounting for distant effects such as conjugation, turned out to be crucial. This is an iterative process, but, in the end, we obtained a result that was very satisfactory and a standard deviation of less than 3 kJ/mole and a maximum deviation of 10 kJ/mole for a widely varying set of organic molecules. When dealing with a different property, the type of third order groups might need to be chosen differently, so in fact different physical properties require a slightly different selection of terms which are, though not in a direct way, related to the different physical nature of the different properties.

Ad (iii) Quality of data. This requires serious attention when we use methods that require data fitting such as QSPR and Group Contribution methods. The source of data must reliable but even in good data sets errors of different kind occur. Depending on the magnitude of the error and the characteristics of the molecule concerned, for example does it involve special combination of chemical group, not removing erroneous data can either have a small or a detrimental effect on the quality of the methods. As we focus on accurate and reliable methods, with numerical requirements set, the effects of erroneous data are expected a problem. Actually, data curation should *always* be considered necessary [37] and a procedure needs to be in place to ensure that the data set does contain any erroneous data.

The larger issue here will be that the experimental data must have at least the accuracy and reliability we require from the method to be developed. For log K_{ow}, related to the octanol-water partition coefficient and a parameter of importance in various domains including toxicity, it has been discussed that experimental values can differ by an order of magnitude [38]. For the AMES test (mutagenicity), it is known that this test has about 85% agreement with true real-life mutagenicity tests. As the AMES test is easy and cheap to apply, the results are often taken to establish a method to predict mutagenicity. But also for parameters like heat of formation and boiling point one finds inaccurate or even wrong data in literature, even in some well-known databases.

Our experience, on a larger variety of data sets, is that erroneous data can often be easily detected by the following approach. Establish a method based on neural networks using a large number of nodes and several layers. The consequence will normally be that one arrives at a method showing overfitting and therefore not good as a predictive tool. However, as the neural network framework can fit any proper data set to any level of accuracy, it turned out in practice that all data points that were off-line an almost perfect correlation model value – experimental value turned out, after careful check, as erroneous input data.

Ad (iv). Regarding *neglecting of the physics of involved*, what is sometimes missing in building models is the discussion on the physics and /or specific interactions involved. Without doing this one builds models based purely on mathematical tools which is some cases do but in other cases do not represent the real-life behaviour and therefore erroneous predictions start popping up when making predictions for systems that were not involved whilst establishing the method. This can be illustrated with the example of free energies of formation, more specifically the entropy part.

Group contribution methods but similarly molecular descriptor like methods, can be applied to molecular enthalpy prediction. For large molecules, the number of atomic connectivity can become very large. However, as in organic chemistry many (but not all) properties have a high degree of additivity of individual molecular group contributions, one still needs to fit relatively few parameters only. Higher order terms are introduced if necessary, as mentioned in the above. If a property depends on the *integral structure of the molecule*, this picture breaks down. This is the case for the vibrational contribution to the molecular entropy, which can be obtained from

$$S_{vibration} = R\sum_i [\frac{(hvi/kT)\exp(-hvi/kT)}{[1-\exp(-hvi/kT)} - \ln[1-\exp(-hvi/kT)]]. \quad (5)$$

with R the gas constant, T absolute temperature, h Planck's constant and ν_i the vibrational frequency associated with the *i*th normal mode and σ the symmetry number [39]. This expression indicates that the low-lying molecular vibrations count most and in practice these are by far dominating. In a not very small organic molecule these vibrations extend over a large part, if not all, of the molecule. This in turn make a group contribution method, which tries to work with a minimum number of neighbours (third order groups as currently the maximum), unsuitable for describing the entropy term and therewith the Gibbs free energy. There are, however, semi-empirical quantum mechanics methods that very quickly provide the low-lying frequencies with high accuracy thereby leading to accuracy and reliable molecular entropies [7,8]. Consequently, a hybrid method involving, for example, a Group Contribution method and a fast semi-empirical quantum mechanics method can yield accurate and reliable predictions of the Gibbs free energy. To arrive at the highest accuracy that seems currently realizable, anharmonic contributions need to be taken into account, where those related to the relatively low rotational barrier of C-C bonds are most often the most important tones and a practical procedure can be put in place [40].

This is just a specific example, but more examples exist. Referring to toxicity prediction, when there is an understanding of the molecular interactions responsible, methods may be developed that are more realistic than statistical methods only, an example being "Thiol Reactivity and Its Impact on the Ciliate Toxicity of α,β-Unsaturated Aldehydes, Ketones and Esters" [41]. These are a few, though real-life relevant examples.

Ad (v) refers to the large class of neutral organic molecules, several physico-chemical properties are of wide interest in different science communities and different methods have been developed for the same purpose. This particularly applies to properties including log K_{ow} where technologists, environmental scientist and chemists all have an interest, the pKa and boiling point. There is no a priori and natural communication between the different communities, although there are obviously exceptions. Looking in the different areas might provide you with a more suitable predictive method. Eventually, close collaboration between such research groups would greatly facilitate progress in the field of predictive modelling.

2.5. Validation

Validation is a key item in any modelling exercise and should be associated to any modelling method, molecular or macroscopic modelling. We have been discussing different methods in this paper and these generally involve different methods of validation.

For ab initio methods, about three approaches are practiced. There is quite a few papers where ab initio data are published without a validation. Still some researchers arguing on the basis that ab initio is from first principles are therewith good. The second approach is to use some reference data on similar systems which serve as calibration for the ab initio results of new but similar systems. When it is about general property prediction, as we primarily discuss in the present paper, reference sets of molecules have been proposed and used for the purpose of validation. One example is the G2 set proposed by Curtiss et al. [42]. This is a good and valid procedure, however these sets consist of really

small molecules only and this is really small, so no real-life relevant molecules for industry or academic research into larger organic or organometallic species. This means we lack proper validation for most molecules. As the error in ab initio calculations is largely at the total energy of the atom level, the error increases with the size of the molecule and we have no proper referencing. Of course one could alleviate this problem but that needs a much larger data base band massive quantum computations.

For the statistical and neural network based models the validation is commonly done using a training set and a test set. We have a set of molecules and for the property of interest also experimental data exist for comparison. A possibly arbitrary subset is taken out of the data set. With arbitrary it is meant that those taken out contain typical groups, double bonds and so forth, resembling the molecules in the larger training set. This is to guarantee that when a good predictive model is constructed for the training set and that model is subsequently applied to the test set, also for the test set we will obtain good results. In order to further improve and guarantee good predictivity one may repeat this process with other test sets, which is a method known as cross-validation. These approaches are generally applied in statistical an neural network based modelling as there is no a priori physical meaning of the correlations obtained. It is in essence all purely mathematical, though these models can have very good predictive properties.

Finally, for the Group Contribution methods a similar validation procedure involving a test set has been practiced [22]. In practice the model is commonly developed on a larger data set, for example thousands of compounds, primarily organic species. We know that many properties are additive to some or a large extent. Higher order groups (neighbour effects) are introduced to correct for deviations due to specific neighbouring.

3. Is There a Way forward towards Reliable Predictive Methods for Physicochemical Property Prediction of Chemicals? How to Arrive at Good Methods for the Prediction of Physicochemical Properties

3.1. Introduction

The state-of-the-art thus suggests it to be difficult to build models that have a high accuracy and high reliability (no significant outliers) and currently there are few methods satisfying this requirement. As we have argued not all cases require accurate predictions, whereas reliability (few outliers, toxic is toxic, order of boiling points is correct, etc.) is obviously always a relevant item. So the question is can we have more accurate and reliable methods that produce results 'on the fly' so the chemist designing the next experiment or the engineer designing a process has quick access to high quality while screening options. As elucidated in the previous sections there will be no single generic solution. From the discussion in the above, it will be evident that for certain classes of problems the quality of data is more problematic and consequently predictions might be less certain (AMES test, log K_{ow}). For organometallics, we will remain largely dependent on either *ab initio* methods or, when its properties can be linked to activity, with descriptor based QSAR methods. The limitations of the former have been referred to above. *Ab initio* methods are applicable to a limited range of properties but remain indispensable for molecules for which we cannot develop appropriate QSPR or GC methods. For organometallics a key property is reactivity. A good tutorial review with many appropriate references was provided by Maldonado and Rothenberg [18].

3.2. Organic Molecules: the Virtue of Group Contribution Methods

For the chemist and the chemical engineer, as well as the corresponding environmental scientists, a significant and very important part of the total number of possible molecules for which we need property data is the class of organic molecules. For these one can argue, as we believe, there is a preferred approach.

There is a unique notation for each organic molecule, known as the Simplified Molecular Input Line Entry Specification (SMILES) [43]. It is a kind of linguistic construct, for example acetic acid reads CC(=O)) and it is *a unique notation for each unique molecule*. It is the most concise way in writing down

the structure of a molecule whilst still retaining all necessary information. A SMILES actually only contains the atom types and the entire connectivity within the molecule. This suggests that it should be possible to establish a QSPR relation, linking the unique property value of a specific molecule to the unique structure of that molecule, on the basis of the SMILES notation only. At least for small molecules this implies only relatively few elements and thus few parameters to be fitted. The SMILES notation does not contain the hydrogen atoms, as these are implicit when the molecular structure is drawn.

The next logical step is the recognition that we are now close to the group contribution approach: a C in the SMILES notation within is a CH_3 or CH_2 are the groups in a GC approach. Furthermore, a combination of atoms in a SMILES might form a group, for example C(=O) or C(=O)O. As we only have atom types and atom-atom connectivity in a SMILES notation, to describe a property of a molecule could still imply we would need the entire connectivity table of a molecule to arrive at proper description of the physicochemical property. Although this still means the SMILES and therefore the molecular structure can be uniquely connected to properties, this would possibly be at the cost of many adjustable parameters and possibly non-linear relations. We now come back to what we know as chemists, namely the fact that it is well-known that in organic chemistry many physicochemical properties can be based on additivity based on molecular groups for instance a CH_3, an acid group COOH and so forth. Thus, we need only limited contribution of groups that are connected at further distance. This is precisely what the first order Group Contribution (GC) methods entail. It is based upon the recognition that a molecular group sufficiently far away from another does not influence the contribution of that other group to the physicochemical property of interest. This will not always be the case but we have seen before that such effects can be incorporated by introducing higher order effect [22]. Thus, the Group Contribution method is the most logical and efficient approach to start developing predictive methods for physicochemical properties, requiring the minimum on input (only the SMILES in effect) and the lowest number of adjustable parameters. Using the GC approach also circumvents the potential problem of overfitting which may readily occur when using a larger number of molecular descriptors in a QSPR approach (see next section). One could see the GC approach as a QSPR approach with only relatively few specific descriptors, namely atom type and connectivity. The downside is of course that the approach will work for organics but not for systems like many of the organometallics. As mentioned earlier, an alternative for organometallics is the use of QSPR based on molecular descriptors rather than molecular group. See Section 3.3 for further reference.

Another advantage in using GC-based methods is illustrated by the following example. The boiling points of the lower cycloalkanes show a slight deviation for cyclohexane as shown in Figure 5. These are typically features known and understood by the chemist as particular structures, like six-membered rings, sometimes slightly deviating properties compared to extrapolation between the two neighbouring members of the family. Straightforward method development might require the need for highly non-linear models to cope with this. The more appropriate way of dealing with this is to introduce such a species as a group by itself with its individual property. Upon requesting the property value from the model, the software tool will simply return the experimental value that was provided along with the definition of the group. By introducing one specific experimental value associated with one specific chemical group (molecule in this case), we arrive at a very good method without further complexity of the method.

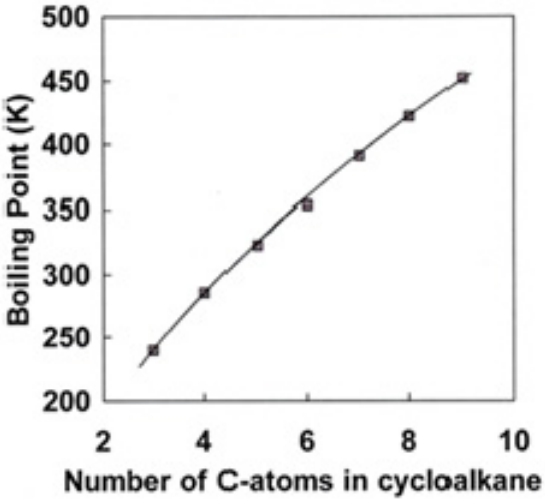

Figure 5. The boiling points of alkanes as a function of chain length.

Such an approach may also very well apply to the prediction of toxicological end-points. If the mode of action of a toxic molecule is through very specific interactions of that particular molecule, something that is not an a priori exceptional as it is similar to the action of drug molecules that need a combination of very specific function for being active and molecules with a slightly different composition or structure having no or hardly any activity, the introduction of such species as groups by themselves is a very appropriate and effective action. In fact, in this way one effectively combines a method that was based on parameter optimization with the knowledge of the expert, an example is the so-called Read-Across knowledge of the toxicologist which is further elucidated in http://www.ecetoc.org/wp-content/uploads/2014/08/ECETOC-TR-116-Category-approaches-Read-across-QSAR.pdf.

Whereas various GC methods have been developed and applied over time, the Gani group has probably made the more important *recent* advances for a wide variety of properties [32,33,44]. These works show and in particular the recent one on heats of formation in which the methodology was adapted to obtain the most accurate results and few outliers, that indeed the Group Contribution methods can do the job in a very appropriate and reliable manner with a standard deviation of less than 3 kJ/mole and a maximum deviation of 10 kJ/mole for a widely varying set of almost 900 organic molecules. Other properties can be modelled in a similar way, leading to better predictive methods.

In summary, the GC method seems, for organic molecules, to be a method requiring a minimum of parameters to uniquely describe a molecule and its physicochemical properties by a QSPR. There is no risk of overfitting (see next section) and there are clear ways within the methodology to increase accuracy and scope of predictions. Earlier in this review we have indicated how one can achieve the desired accuracy.

3.3. QSPR Methods using Molecular Descriptors

After molecular descriptors (see Section 1.2 for the introduction of these) have been evaluated for a series of molecules for which experimental property data are available, for example boiling, points, heats of formation, toxicity data or any other property, statistical or artificial intelligence methods are applied to build a method between descriptors and physicochemical property. Subsequently, for a new molecule for which the property value is unknown, the molecular descriptors are calculated and inserted in the model, leading to a prediction for the physicochemical property. The methods applied are generically of two different types: classical statistical methods and artificial intelligence. The former include the regularly used linear regression and, if not sufficient, polynomial regression and other ways to fit a given mathematical function of relatively simple form. We have seen, however, that current methods are mostly not sufficiently accurate and predictive for the experimental chemist and engineer, the environmental scientist or the toxicologist. The classical methods allow for limited flexibility and

there is no reason to assume that properties exhibit a linear or any other fixed mathematical form. Practice (literature) has shown the accuracy and reliability remain limited.

Neural networks are more generic and can model any complex relationship to any desired accuracy (see, for example, Reference [11]). Surprisingly, although Gasteiger and Zupan pointed out the capabilities of neural networks as early as 1993 [45], still many use classical statistical methods. Still, meanwhile neural networks and other artificial intelligence methods are gaining exposure, not only for physicochemical properties but in a much wider area including describing the relation between (chemical plant) process parameters like catalyst type, catalyst concentration, flow, feed characteristics and so forth. Boiling points of organics were modelled using neural networks using 17,768 pure chemical compounds as available input data for constructing the model [16]. The average absolute relative deviations of the predicted properties from existing literature values: 3.2% and squared correlation coefficient R^2 was 0.94. Another study compared various publicly available methods using over 2000 experimental data points [15]. Also the more difficult to predict melting points were reported [46], as well as pKa neural network methods [47]. Very recently a neural network based approach was reported for the calculation of molecular energies, claiming less than 3 kJ/mole (0.6 kcal/mole) difference between DFT results and the neural network method [48].

The molecular descriptor based approach can equally well be applied to organometallics. Reports date more than a decade ago, for example on the catalytic efficiency in the Heck reaction [49], More than a decade ago. The method was applied to other classes of homogeneous catalysts [18] and to the heat of formation of organometallics [50]. There are two excellent reviews available from Fey [51,52].

A critical issue in modelling using molecular descriptors is how many should be used. As we have indicated tools are available to readily evaluate thousands of parameters. Obviously, when we have only hundreds or even a few thousands of experimental data as input during the construction of the method, we easily have more parameters that can be fitted compared to the number of experimental data. This leads to a phenomenon known as overfitting. Overfitting and ways to deal with this has been discussed by, for example, Chang et al. [53]. Having too many fitting parameters results in multiple methods with different parameter values describing the experimental data set equally well but which is the correct method? As these methods are often produced as black boxes, albeit for a good reason, they will be less predictive. Thus, we need to reduce the number of descriptors to a value much less than the number of available experimental parameters. The question arising is how to do this and still having a good predictive method. In practice, often Principal Component Analysis (PCA) is applied, which is a statistical procedure to reduce the number of parameters to a level that the remaining parameters still describe a large part of the variability of the data set. However, as argued before, we need very high-quality correlations to arrive at the required accuracy of prediction and then a retention of variability of 90–98%, the normal range attempted, is often not sufficient. Therefore, we do need sufficient experimental data to allow for more descriptors and therefore more parameters to be fitted. That this works in practice is illustrated by the work of Espinosa et al. on the boiling points of organics [54]. They applied neural network modelling involving a few atomic connection descriptors (maximum was seven) which resulted in very good results. For saturated hydrocarbons average errors below 2K were obtained. The low number of descriptors ensures that no overfitting is involved. This, however, was for saturated hydrocarbons only, for the alkenes the results were worse (average absolute errors up to 6K) but it illustrates what is possible with a judicious choice of descriptors.

4. Conclusions

In this review we have focused on the accuracy and reliability of predictive methods for a series of physicochemical properties of chemicals that are of significant relevance for industrial as well as academic research. In addition, in order for these to have significant impact, results should preferably be obtained 'on the fly.'

While in first instance one may think these requirements are not mutually compatible, we think we have shown that both Group Contribution (GC) methods as well as QSPR approaches with methods

constructed using statistical or artificial intelligence methods can provide physicochemical properties 'on the fly,' and, as soon as a method is built by the expert, it can be easily applied by the experimental chemist, the process engineer or the environmental scientist. Such methods can, when developed appropriately, as demonstrated with a few examples, make reliable and sufficiently accurate predictions for the end-user. From its basic principles, the Group Contribution method is the method of natural choice for property prediction of organic molecules. Further work is needed to capture many more properties than the few for which there are good methods at present. The same applies for molecular descriptor based QSPR methods for all cases in which the GC approach is not the method of choice. Ab initio calculations will serve in those cases where we need very specific data or where we have no experimental data set to build methods and otherwise for the calculation of molecular descriptors. Close collaboration between such research groups would greatly facilitate progress in the field of predictive modelling.

Importantly, we have emphasized that good correlations are not always a requirement, for example for the prediction whether a catalyst is very good we need to have good qualitative results, viz. Figure 5, whereas in other cases we need quite accurate and reliable energy differences, viz. Figure 4.

Finally, collaboration between different communities could more efficiently lead to the best methods for certain properties, for example log Kow.

Conflicts of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Fink, T.; Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353. [CrossRef]
2. Hendriks, E.; Kontogeorgis, G.M.; Dohrn, R.; de Hemptinne, J.-C.; Economou, I.G.; Fele Žilnik, L.; Vesovic, V. Industrial Requirements for Thermodynamics and Transport Properties. *Ind. Eng. Chem. Res.* **2010**, *49*, 11131–11141. [CrossRef]
3. Katritkzy, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D.; Karelson, M.; Kahn, I.; Dobchev, D.A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110*, 5714–5789. [CrossRef]
4. Nieto-Draghi, C.; Fayet, G.; Creton, B.; Rozanska, X.; Rotureau, P.; De Hemptinne, J.-C.; Ungerer, P.; Rousseau, B.; Adamo, C. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chem. Rev.* **2015**, *115*, 13093–13164. [CrossRef] [PubMed]
5. Van Speybroeck, V.; Gani, R.; Meier, R.J. The calculation of thermodynamic properties of molecules. *Chem. Soc. Rev.* **2010**, *39*, 1764–1779. [CrossRef] [PubMed]
6. Rozanska, X.; Ungerer, P.; Leblanc, B.; Saxe, P.; Wimmer, E. Automatic and systematic Atomistic Simulations in the MedeA® Software Environment: Application to EU-REACH. *Oil Gas Sci. Technol. Rev. IFP Energies Nouv.* **2015**, *70*, 395–403. [CrossRef]
7. Barrett, R.A.; Meier, R.J. The calculation of molecular entropy using the semiempirical AM1 method. *J. Mol. Struct.* **1996**, *363*, 203–209. [CrossRef]
8. Rozanska, X.; Stewart, J.J.P.; Ungerer, P.; Leblanc, B.; Freeman, C.; Saxe, P.; Wimmer, E. High-Throughput Calculations of Molecular Properties in the MedeA Environment: Accuracy of PM7 in Predicting Vibrational Frequencies, Ideal Gas Entropies, Heat Capacities and Gibbs Free Energies of Organic Molecules. *J. Chem. Eng. Data* **2014**, *59*, 3136–3143. [CrossRef]
9. Todeschini, R.; Consonni, V. (Eds.) *Handbook of Molecular Descriptors*; Wiley-VCH: Hoboken, NJ, USA, 2000.
10. Tseng, Y.; Hopfinger, A.J.; Esposito, E.X. The great descriptor melting pot: Mixing descriptors for the common good of QSAR models. *J. Comput. Aided Mol. Des.* **2012**, *26*, 39–43. [CrossRef]
11. Le, T.; Epa, V.C.; Burden, F.R.; Winkler, D.A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889–2919. [CrossRef]

12. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027–1044. [CrossRef]
13. Ren, Y.; Zhao, B.; Chang, Q.; Yao, X. QSPR modeling of nonionic surfactant cloud points: An update. *J. Coll. Interf. Sci.* **2011**, *358*, 202–207. [CrossRef] [PubMed]
14. Lee, A.C.; Crippen, G.M. Predicting pKa. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033. [CrossRef]
15. Oprisiu, I.; Marcou, G.; Horvath, D.; Brunel, D.B.; Rivollet, F. Publicly available models to predict normal boiling point of organic compounds. *Thermochim. Acta* **2013**, *553*, 60–67. [CrossRef]
16. Gharagheizi, F.; Mirkhani, S.A.; Ilani-Kashkouli, P.; Mohammad, A.H.; Ramjugernath, D.; Richon, D. Determination of the normal boiling point of chemical compounds using a quantitative structure-property relationship strategy: Application to a very large dataset. *Fluid Phase Equilib.* **2013**, *354*, 250–258. [CrossRef]
17. Piliszek, S.; Wilczyńska-Piliszek, A.J.; Falandysz, J. N-octanol-water partition coefficients (log K(OW)) of 399 congeners of polychlorinated azoxybenzenes (PCAOBs) determined by QSPR- and ANN-based approach. *J. Environ. Sci. Health A Tox. Hazard Subst. Environ. Eng.* **2011**, *46*, 1748–1762. [CrossRef] [PubMed]
18. Maldonado, A.G.; Rothenberg, G. Predictive modeling in homogeneous catalysis: A tutorial. *Chem. Soc. Rev.* **2010**, *39*, 1891–1902. [CrossRef]
19. Joback, K.G.; Reid, R.C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Commun.* **1987**, *57*, 233–243. [CrossRef]
20. Benson, S.W.; Cruickshank, F.R.; Golden, D.M.; Haugen, G.R.; O'Neal, E.H.; Rodgers, A.S.; Shaw, R.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279–324. [CrossRef]
21. Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE J.* **1994**, *40*, 1697–1710. [CrossRef]
22. Marrero, J.; Gani, R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilib.* **2001**, *183–184*, 183–208. [CrossRef]
23. Bicerano, J. (Ed.) *Prediction of Polymer Properties*; Marcel Dekker Inc.: New York, NY, USA, 1993.
24. Kier, L.B.; Hall, H.L. (Eds.) *Molecular Connectivity in Structure Activity Analysis*; John Wiley & Sons: New York, NY, USA, 1986.
25. Randic, M. The connectivity index 25 years after. *J. Mol. Graph. Model.* **2001**, *20*, 19–35. [CrossRef]
26. Brown, R.D.J. A quantum-mechanical treatment of aliphatic compounds. Part I. Paraffins. *J. Chem Soc.* **1953**, 2615–2621. [CrossRef]
27. Constantinou, L.; Prickett, S.E.; Mavrovouniotis, M.L. Estimation of thermodynamic and physical properties of acyclic hydrocarbons using the ABC approach and conjugation operators. *Ind. Eng. Chem. Res.* **1993**, *32*, 1734–1746. [CrossRef]
28. Constantinou, L.; Prickett, S.E.; Mavrovouniotis, M.L. Estimation of Properties of Acyclic Organic Compounds Using Conjugation Operators. *Ind. Eng. Chem. Res.* **1994**, *32*, 395–402. [CrossRef]
29. Katrizky, A.R.; Slavov, S.H.; Dobchev, D.A.; Karelson, M. Rapid QSPR model development technique for prediction of vapor pressure of organic compounds. *Comput. Chem. Eng.* **2007**, *31*, 1123–1130. [CrossRef]
30. Kahrs, O.; Brauner, N.; Cholakov, G.S.; Stateva, R.P.; Marquardt, W.; Shacham, M. Analysis and refinement of the targeted QSPR method. *Comput. Chem. Eng.* **2008**, *32*, 1397–1410. [CrossRef]
31. Ceriani, R.; Gani, R.; Meirelles, A.J.A. Prediction of heat capacities and heats of vaporization of organic liquids by group contribution methods. *Fluid Phase Equilib.* **2009**, *283*, 49–55. [CrossRef]
32. Hukkerikar, A.S.; Meier, R.J.; Sin, G.; Gani, R. A method to estimate the enthalpy of formation of organic compounds with chemical accuracy. *Fluid Phase Equilib.* **2013**, *348*, 23–32. [CrossRef]
33. Hukkerikar, A.S.; Kalakul, S.; Sarup, B.; Young, D.M.; Sin, G.; Gani, R. Estimation of Environment-Related Properties of Chemicals for Design of Sustainable Processes: Development of Group-Contribution+ (GC+) Property Models and Uncertainty Analysis. *J. Chem. Inf. Model.* **2012**, *52*, 2823–2839. [CrossRef] [PubMed]
34. Meier, R.J.; Gundersen, M.T.; Woodley, J.M.; Schürmann, M. A Practical and Fast Method to Predict the Thermodynamic Preference of ω-Transaminase-Based Transformations. *Chem. Cat. Chem.* **2015**, *7*, 2594–2597. [CrossRef]
35. Dearden, J. Expert Systems for Toxicity Prediction. In *Situ Toxicology, Chapter 19*; Cronin, M., Madden, J., Eds.; The Royal Society of Chemistry: London, UK, 2010.
36. Thomas, R.S.; Black, M.B.; Li, L.; Healy, E.; Chu, T.M.; Bao, W.; Andersen, M.E.; Wolfinger, R.D. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicol. Sci.* **2012**, *128*, 398–417. [CrossRef]

37. Fourches, D.; Muratov, E.; Tropsha, A. Trust but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204. [CrossRef]
38. Renner, R. The KOW Controversy. Doubts about the quality of basic physicochemical data for hydrophobic organic compounds could be undermining many environmental models and assessments. *Environ. Sci. Technol.* **2002**, *36*, 411A–413A. [CrossRef]
39. Atkins, P.W. (Ed.) *Physical Chemistry*, 2nd ed.; Oxford University Press: Oxford, UK, 1982.
40. Vansteenkiste, P.; Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. Ab initio calculation of entropy and heat capacity of gas-phase n-alkanes with hetero-elements O and S: Ethers/alcohols and sulfides/thiols. *Chem. Phys.* **2006**, *328*, 251–258. [CrossRef]
41. Böhme, A.; Thaens, D.; Schramm, F.; Paschke, A.; Schüürmann, G. Thiol Reactivity and Its Impact on the Ciliate Toxicity of α,β-Unsaturated Aldehydes, Ketones and Esters. *Chem. Res. Toxicol.* **2010**, *23*, 1905–1912. [CrossRef] [PubMed]
42. Curtiss, L.A.; Raghavachari, K.; Trucks, G.W.; Pople, J.A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *J. Chem. Phys.* **1991**, *94*, 7221–7230. [CrossRef]
43. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
44. Ceriani, R.; Gani, R.; Liu, Y.A. Prediction of vapor pressure and heats of vaporization of edible oil/fat compounds by group contribution. *Fluid Phase Equilib.* **2013**, *337*, 53–59. [CrossRef]
45. Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem. Int. Ed.* **1993**, *32*, 503–527. [CrossRef]
46. Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. Prediction of Melting Point for Drug-like Compounds Using Principal Component-Genetic Algorithm-Artificial Neural Network. *Bull. Korean Chem. Soc.* **2008**, *29*, 833–841. [CrossRef]
47. Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. Prediction of basicity constants of various pyridines in aqueous solution using a principal component-genetic algorithm-artificial neural networks. *Monatsh. Chem.* **2008**, *139*, 1423–1431. [CrossRef]
48. Smith, J.S.; Isayev, O.; Roitberg, A.E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203. [CrossRef] [PubMed]
49. Tabares-Mendoza, C.; Guadaramma, P. Predicting the catalytic efficiency by quantum-chemical descriptors: Theoretical study of pincer metallic complexes involved in the catalytic Heck reaction. *J. Organometal. Chem.* **2006**, *691*, 2978–2986. [CrossRef]
50. Jover, J.; Bosque, R.; Martinho Simões, J.A.; Sales, J. CORAL: QSPRs of enthalpies of formation of organometallic compounds. *J. Organometal. Chem.* **2008**, *693*, 1261–1268. [CrossRef]
51. Fey, N. The contribution of computational studies to organometallic catalysis: Descriptors, mechanisms and models. *Dalton Trans.* **2010**, *39*, 296–310. [CrossRef] [PubMed]
52. Fey, N. Lost in chemical space? Maps to support organometallic catalysis. *Chem. Cent. J.* **2015**, *9*, 38. [CrossRef] [PubMed]
53. Chang, C.-Y.; Hsu, M.-T.; Esposito, E.X.; Tseng, Y.J. Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition. Molecular Descriptors and Predictive Modeling Methods. *J. Chem. Inf. Model.* **2013**, *53*, 958–971. [CrossRef]
54. Espinosa, G.; Yaffe, D.; Cohen, Y.; Arenas, A.; Giralt, F. Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 859–870. [CrossRef]