Journal of Molecular Graphics and Modelling 30 (2011) 120–128

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Estimation of boiling points using density functional theory with polarized continuum model solvent corrections

Poh Yin Chan, Chi Ming Tong, Marcus C. Durrant*

School of Life Sciences, Northumbria University, Ellison Building, Newcastle-upon-Tyne NE1 8ST, United Kingdom

ARTICLE INFO

Article history:
Received 2 February 2011
Received in revised form 21 June 2011
Accepted 27 June 2011
Available online 5 July 2011

Keywords:
Boiling points
DFT
Implicit solvent corrections
QSPR
Quantum calculations

ABSTRACT

An empirical method for estimation of the boiling points of organic molecules based on density functional theory (DFT) calculations with polarized continuum model (PCM) solvent corrections has been developed. The boiling points are calculated as the sum of three contributions. The first term is calculated directly from the structural formula of the molecule, and is related to its effective surface area. The second is a measure of the electronic interactions between molecules, based on the DFT-PCM solvation energy, and the third is employed only for planar aromatic molecules. The method is applicable to a very diverse range of organic molecules, with normal boiling points in the range of −50 to 500 °C, and includes ten different elements (C, H, Br, Cl, F, N, O, P, S and Si). Plots of observed *versus* calculated boiling points gave $R^2 = 0.980$ for a training set of 317 molecules, and $R^2 = 0.979$ for a test set of 74 molecules. The role of intramolecular hydrogen bonding in lowering the boiling points of certain molecules is quantitatively discussed.

1. Introduction

The prediction of physicochemical properties such as boiling points is a basic goal of computational chemistry. The normal boiling point (BP) can be defined as the temperature at which the vapour pressure of a pure liquid reaches 760 mm Hg. It is obvious that the BP of a compound is related in general terms to its molecular structure, but the nature of the relationship is subtle and often difficult to predict. For example, one can make the qualitative prediction that an alcohol will have a higher BP than an isomeric ether; but it is much more difficult to make the quantitative prediction that *n*-butanol and diethyl ether boil at 118 and 34 °C, respectively. An early success in the development of such quantitative structure–property relationships (QSPR) was provided by Wiener [1], who developed a method for the prediction of boiling points of alkanes, based on two parameters derived from the structural formula of the alkane. Both parameters quantify the topology of the molecule; one of them came to be known as the Wiener index *w* and is a measure of the extent of branching within the molecular structure. For isomeric alkanes, increased chain branching generally correlates with lower boiling points. This can be rationalised on the basis that the van der Waals surface area is reduced for branched molecules, thereby reducing the strength of the intermolecular interactions which must be overcome to progress from the liquid to gas phase.

In progressing beyond alkanes, the prediction of boiling points becomes a more exacting problem. A variety of QSPR approaches have been investigated. Cramer [2,3] has shown by principal component analysis that a range of physical properties, including the BP, can be correlated with just five derived parameters, of which the first two are most important. These two parameters, *B* and *C*, were associated with the molecular bulk and cohesiveness, respectively. Thus, small molecules such as H_2 have the smallest *B* values, whilst highly polar molecules such as water have the highest *C* values. The parameters *B* and *C* can be calculated from the structural formula of a molecule, using a fragment-based approach. Using this method, a plot of predicted *versus* experimental boiling points for a test set of 139 diverse molecules, including nine different elements (C, H, Br, Cl, F, I, N, O, and S), gave an overall correlation coefficient R^2 of 0.932 as a benchmark for future studies.

The advent of inexpensive quantum calculations offered a fresh line of attack for this problem. The value of quantum calculations in the general context of QSPR has been demonstrated by Popelier and co-workers, who have developed a method called quantum topological molecular similarity (QTMS) to address the prediction of both physical properties such as pK_a values and Hammett constants, and biological properties such as the activities of drug molecules [4–6]. QTMS has been used with a range of quantum methods including Hartree–Fock and DFT. We have found that DFT calculations with PCM solvent corrections can be applied to the prediction of pK_a values of both organic and inorganic molecules

* Corresponding author. Tel.: +44 191 2437239.
E-mail address: marcus.durrant@northumbria.ac.uk (M.C. Durrant).

1093-3263/$ – see front matter.
doi:10.1016/j.jmgm.2011.06.010

[7]. Meanwhile, a number of groups have investigated the use of quantum calculations at various levels of sophistication for the prediction of boiling points. Thus, Katritzky et al. [8] combined a QSPR approach with molecular descriptors extracted from AM1 semi-empirical calculations to fit the boiling points of a training set of 298 organic molecules, containing six different elements. This returned an overall R^2 value of 0.973 and an average prediction error of 2.3% with the use of four parameters. For comparison, the error associated with the experimental values was previously estimated as 2.1% [9]. As with Cramer's analysis, the two most significant parameters were associated with molecular bulk (the so-called gravitation index) and electrostatic effects (defined as hydrogen bonding ability). This work was subsequently extended to a set of 612 organic compounds, containing nine elements (C, H, Br, Cl, F, I, N, O, and S) [10]. This resulted in a versatile eight parameter model, with $R^2 = 0.965$ and a standard prediction error of 15.5 °C which was comparable to the estimated experimental RMS error for the data set (11.4 °C) [11].

In a series of papers, Jurs and co-workers developed a suite of methods for the prediction of boiling points based on molecular descriptors such as the Wiener index and surface charge areas [9,11,12]. PM3 semi-empirical calculations were used to provide the descriptors and the analyses were carried out by regression, computational neural network and genetic algorithm methods. Different models were developed for different types of molecule; for example, a model using 10 variables was developed for a set of 277 compounds containing eight different elements but excluding nitrogen (i.e. C, H, Cl, Br, F, I, O and S). After exclusion of two outliers, the method gave RMS errors of 11.6 and 10.5 °C for a training subset of 248 compounds and a test subset of 27 compounds, respectively. Further refinement using a computational neural network reduced the RMS error for the prediction set to 9.0 °C [11]. Although these methods gave accurate predictions using relatively few descriptors (typically up to 10), an important consideration is that the set of descriptors used for each individual model was chosen from a much larger set of available descriptors, in order to give the best fit between observed and calculated boiling points. Thus, individual models used very different descriptor sets, which were very strongly dependent on the set of compounds chosen as the data set [12]. This is also true of subsequent work by Sola et al. [13], who followed a similar QSPR approach to calculate both boiling points and critical properties, using AM1 semi-empirical calculations to generate about 500 descriptors. Their final model gave $R^2 = 0.985$ with an RMS error of 9.1 °C for a training set of 135 compounds and an RMS error of 7.3 °C for a test set of 20 compounds, including five elements (C, H, Cl, N and O). There was however little overlap between the eight descriptors used by Sola et al. and those used by Jurs et al. Hence, QSPR approaches to boiling point prediction often feature a degree of arbitrariness in their selection of molecular descriptors, and may not provide a very clear conceptual model for the factors that determine the BP.

A similar QSAR approach was taken by Stanton [14], but using molecular mechanics with electrostatic terms rather than quantum calculations to provide the molecular descriptors. A diverse training set of 268 molecules (after removal of 26 outliers) was used to develop a model from 12 of the available descriptors. The mean error for the training set was 12.3 °C, whilst the prediction error for a test set of 78 additional molecules was 16.7 °C. An interesting observation from this work was that intramolecular hydrogen bonding can significantly reduce the boiling point. Thus, the BP for 2-hydroxybenzaldehyde was predicted as 241 °C, much higher than the observed value of 196 °C. This discrepancy was attributed to an intramolecular hydrogen bond between the two functional groups in the molecule. Consistent with this suggestion, the experimental BP of 3-hydroxybenzaldehyde which cannot form an equivalent intramolecular hydrogen bond is 240 °C.

Clark and co-workers combined a neural network approach with descriptors calculated by AM1 and PM3 semi-empirical methods to analyse a very large and diverse set of molecules, including 17 different elements [15]. AM1 was found to give more accurate results than PM3. Use of 18 descriptors with a training set of 6000 molecules gave an overall R^2 of 0.959, whilst standard deviations for the training set and a validation set of a further 629 molecules were 16.5 and 19.0 °C, respectively. The structural descriptors included surface area and globularity, which varies from 1 for a perfect sphere to close to 0 for long unbranched chains such as normal hydrocarbons. The importance of correct selection of tautomers was considered, along with the effects of varying the conformation of flexible molecules.

The problem of selection of a small number of molecular descriptors from a vast number of possible candidates has been discussed by Duchowicz and co-workers [16,17], who advocated the use of flexible as well as rigid molecular descriptors for regression analysis. Thus, a set of 200 diverse molecules, containing 10 elements (C, H, Br, Cl, F, N, O, P, S and Si) was subjected to regression analysis to give a linear equation for the boiling point. This equation included one flexible descriptor DCW^1, derived from a graphical description of atomic orbitals, plus five rigid descriptors selected from 1199 candidates, to give an overall $R^2 = 0.942$.

One would hope that more sophisticated quantum calculations should give better molecular descriptors, and this proved to be true of recent work from Kumar [18], who correlated the boiling points of a set of 75 alkanes with descriptors from different quantum calculations. The AM1, PM3 and DFT calculations returned R^2 values of 0.891, 0.910 and 0.941, respectively (in each case, the worst three data points were rejected as outliers). Interestingly, the R^2 value for the DFT calculations could be improved to 0.959 by the inclusion of Klopman atomic softness values, calculated in water as solvent. Recently, Chen et al. used DFT to execute a QSPR study of the boiling points of some organic compounds [19]. They found that linear equations using the molecular average polarizability, most negative atomic net charge, and dipole moment as descriptors gave R^2 values of 0.933 and 0.945 for data sets of oxygen- and sulfur-containing compounds, respectively.

To summarise, QSPR calculations using descriptors based on quantum chemical calculations can provide reasonably accurate models for the prediction of boiling points, with accuracies approaching typical experimental errors in the best examples. Nevertheless, a number of shortcomings in current methods are evident. First, different models generally use different descriptors which are chosen from very large sets of calculated parameters. Although this allows individual models to maximise their R^2 values and still be valid for unknown but chemically related molecules, this approach inevitably raises questions about the robustness and generality of the models. Second, the arbitrary nature of the descriptors, when chosen for their statistical performance, means that the resulting models often have rather limited physical meaning. We know that in general terms, boiling point increases with molecular size, strength of intermolecular interactions, and deviation of the molecule from sphericality. Most molecular descriptors capture aspects of these observations, but often the conceptual links are tenuous. It would be useful to have a model that more explicitly reflects these general observations, but still gives accurate predictions.

Whilst considering this problem, we surmised that implicit solvation energies might provide a useful molecular descriptor. In recent years, the development of implicit solvent models such as the polarized continuum model (PCM) has allowed for reasonably accurate calculation of solvation energies, both in organic solvents and in water [20]. The non-specific solute–solvent interactions